Uncovering ecological associations by learning latent representations of species effects and responses from their co-distribution

Sara Si-moussi, Esther Galbrun, Mickael Hedde, Wilfried Thuiller

October 14, 2019

1 Background

Understanding the drivers of species distribution and their abundance has always been a long-lasting goal of biogeography [14]. Niche theory explains the spatial dynamics of species by a set of physiological and adaptive properties that lead them to thrive in specific conditions and decline in others ([3], [24]). The ranges of abiotic variables, such as climate and soil characteristics, that match the ecophysiological requirements of a species delimit its potential niche (Grinnellian niche, [9]) or (Fundamental niche, [15]). Habitat suitability models (HSM) or species distribution models (SDMs) [11] aim to infer and model this niche by establishing statistical relationships between observed presences an absence or species' abundances and the environmental characteristics of the corresponding locations.

SDMs have proven useful to predict species ranges in response to climate change, providing operational tools to conservation biologists ([10], [7]). However, they predict multiple species distributions separately, ignoring possible dependencies that can restrict or extend species' ranges from what is expected when considering only abiotic factors. Indeed, species may exclude one another locally (*competitive exclusion*, [12]) or differentiate each other's areas and resources (*niche partitioning*, [27]). 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). Although these interactions take place at a local scale, some of them may restrict the range of the species on a wider, macroscopic scale. Thus, the inability to take into account the presence or absence of other species is an important source of underfitting for SDMs ([31]).

Over the last decade, several approaches have been proposed to infer interspecific associations within the latter models by jointly analyzing the co-distribution of multiple species. This led to the development of Joint Species Distribution Models, JSDM ([23],[21]). The gist is that once we have accounted for abiotic factors, the unexplained variance, typically captured by the residuals' correlation matrix, is attributed to the effect of species on one another. As they rely purely on correlation, JSDMs are limited to inferring symmetric associations between species, such as mutualism and competition. Asymmetric interactions, such as commensalism, amensalism and predation, are therefore inherently out of reach for these models. [16] proposed a JSDM that allows to capture asymmetric associations but it requires access to longitudinal data.

Another research domain that is also concerned with inferring associations from co-occurrence is natural language processing (NLP). NLP modelers try to understand semantic associations between words on the basis of the contexts they appear in. The context of a word is the list of words surrounding it in a sentence. A state of art method in this area is `word2vec` ([18]). `word2vec` learns distributed representations for words in the form of dense vectors called embeddings. The embedding of a word captures information about other words it co-occurs with. The probability of a word occurring in a particular sentence of a text depends on the semantic compatibility of this word with other words occurring in that sentence. Word embeddings, such as `word2vec` , aim to learn multidimensional representations of words that captures this contextual semantic compatibility. By analogy, in community ecology, the probability of the presence of a species in a given abiotically suitable site depends on its compatibility with other species occuring at that site, i.e. other species in the observed community.

[25, 17] introduced a generalization of word embeddings to any type of data that follow an exponential family distribution, including binary and ordinal data, called exponential family embeddings. Building on existing work, we propose 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. We evaluate the potential of the model to recover correct associations on empirically observed as well as simulated communities.

2 Material and Methods

Three main conditions should be satisfied for a species to inhabit a given location. First, the location must be accessible. This relates to the intrinsic dispersal capacity of the species 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 as *habitat suitability* and is the target of Habitat Suitability Models. Third, intraspecific and interspecific interactions within communities can also alter the range of the species as well as their local abundances ([11]). 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*, also refered to as the *abiotic* and *biotic* factors, respectively.

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 environmental variables measured at these same sites or in their vicinity.The abundance of species i at site k is noted y_{ki}. The abiotic environment in site k is represented by the *abiotic feature* vector x_k computed by applying a feature extraction model on the raw environmental variables.

In what follows, we introduce key concepts used in the inference model. In particular, we explain how we model the interactions 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 represent the biotic factors.

2.1 Biogeographical associations and biotic context

A *biogeographical association* describes the relative influence of a pair of species on one another's abundances. The two directions of this influence can be of different types (positive, negative or neutral) and have different intensities. An association represents a consistent pattern of co-occurrence with potentially multiple mechanistic explanations: a direct biotic interaction, an indirect interaction through the environment or a shared correlation to an unmeasured environmental variable or an unobserved group of organisms.

2.1.1 Representing species associations using embeddings

Formally, we represent the association between species i and j with a pair of numerical values a_{ij} and a_{ji}, representing the strength of the influence of species i on species j and vice-versa, respectively. Specifically, a_{ij} represents the change (excess if positive, deficit if negative, none if null) in *target* species j's abundance induced by the presence of an individual from *source* species i. These values across all pairs of species can be collected into an $m \times m$ non-symetric association matrix A.

The association strength depends on two latent parameters: the *effect* applied by the source and the *response* of the target. We assume these parameters are controlled by intrinsic traits 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 parameter which is typically significantly smaller than half the number of species.

The *effect embedding* of species i is denoted as ρ_i, it captures the type of organisms the species allows when it is present. The *response embedding* of species i is denoted as α_i, it controls 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 $m \times d$ matices, respectively denoted as P and Q.

The association matrix is then written as

$$A = PQ^T$$

2.1.2 Biotic context

The biotic context encodes our assumptions about potential biotic effects a target species is exposed to in a given site. In the simplest case, without any

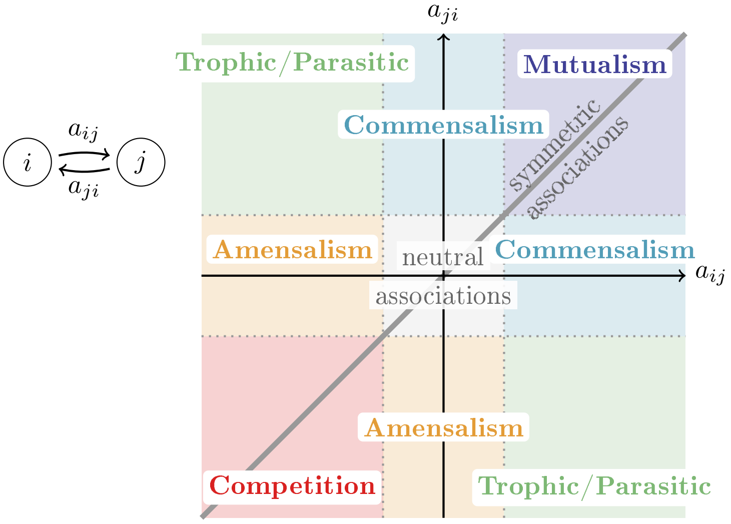


Figure 1: Mapping pairwise association strengths to potential interaction classes. The first bissector represents the association domain covered by correlation-based approaches including Joint Species Distribution Models.

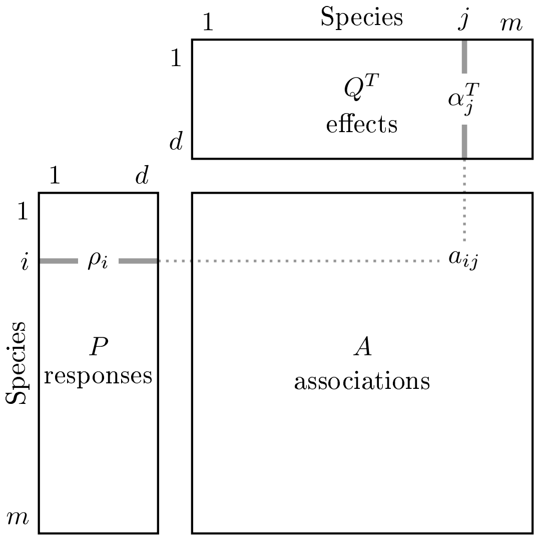


Figure 2: Association matrix factorization.

prior knowledge, it is formed of individuals from other species observed in the same location. Specifically, the biotic context of species i in site k, denoted C_{ki}, is computed as follow:

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

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 compensation between the presence of facilitators and competitors. By weighting with abundance, we consider implicitely that individuals from the same species are similar and contribute equally to the community structure. Conversely, the effect of rare species would only be apparent if their per capita effect is stronger than the aggregated effect of dominant groups.

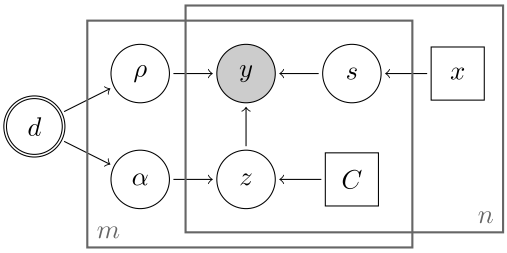
The biotic context carries implicit constraints on the structure of species association networks by restricting the set of potential associations a priori. We present some alternative definitions of the biotic context with the associated data requirements and relevant effect aggregation functions in the appendix (see Section B).

2.2 A conditional generative model of abundance

2.2.1 Formalization

The indicator of abiotic suitability for species i at site k, denoted s_{ki}, follows a Bernoulli distribution, whose parameter (success rate) is given by a habitat suitability model (HSM), h, fitted on the target species's occurrences.

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



d	embedding dimension
x	abiotic feature vector
s	abiotic suitability
C	biotic context
ρ	species response
α	species effect
y	species abundance
z	biotic context effect

Figure 3: Plate diagram of the generative model of abundance.

At sites where the abiotic environment is not suitable (i.e. where $s_{ki} = 0$), the probability mass of the species abundance is concentrated on zero. In other words, η_{ki} follows the Dirac delta function denoted as δ_0. Otherwise (i.e. where $s_{ki} = 1$), the abundance of the species is a function of its biotic context.

Following (Rudolph et al. 2016), we model the abundance using the canonical form of the exponential family $\mathcal{E}$ parameterized by η_{ki}.

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

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

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

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

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

Different choices of probability distributions, depending in particular on the type of data considered (presence/absence vs. abundance) result in different instanciations of this generic model. In Table 1, we provide the mapping from the natural parameter to the expression of the mean for different choices of the link function in the special cases where $\mathcal{E}$ is a binomial (with fixed number of trials), a negative binomial (with fixed number of failures) or a Poisson distribution.

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 and link function choices for common distributions used for presence/absence or count data.

2.2.2 Inference

Having formalized our model, we now outline the inference process, i.e. the procedure for training the model.

In order to prepare the examples for training the model, we gather as positive and negative examples respectively the species that are present and absent at each site k:

$$\mathcal{S}_k^+ = \{i \in \mathcal{S}, y_{ki} > 0\}$$
$$\mathcal{S}_k^- = \{i \in \mathcal{S}, y_{ki} = 0\}$$

Negative examples are typically overrepresented in the dataset, leading to a high inbalance between positive and negative examples. Therefore, we subsample negative examples, selecting $r\%$ of them at random into the training set, alongside all positive examples. By introducing noisiness into the objective function, this sub-sampling procedure also improves the robustness of the estimations and prevents overfitting. Furthermore, we promote the sparsity of the embeddings and of the resulting association matrices by adding lasso penalties on the embedding vectors.

Our model is then trained on this collection of labelled examples. Specifically, we maximize the likelihood of the observed abundances on each site for all positive examples and the selected subset of negative examples.

In summary, given training data (abundances Y and abiotic variables X) together with user-defined hyperparameters (including hyperparameters for the abiotic suitability model ϕ_h, embedding dimension d, vector of species offsets O, negative examples subsample rate r and regularization coefficient on the constraint of embeddings sparsity λ) the model training procedure aims to infer the value of the model parameters (esp. the response and effect embeddings matrices P and Q, and the HSM parameters θ_h) that optimize the objective function.

To do so, we use online stochastic gradient descent. Details of the objective function and its derivative with respect to each parameter are provided in the appendix (see Section A).

The resulting model can then be applied on previously unseen data, e.g. environment features for sites not included in the training data, to predict abundances at these sites for each of the species in the training data given the abundances of other species.

2.3 Unraveling inter-specific interactions

While the primary task addressed by our model is the prediction of abundances, the effect and responses embeddings are also learnt, as by-products of the inference. In fact, these embeddings for the different species, contained in matrices Q and P, as well as the resulting non-symmetric association matrix A are of great interest, as they shed light on the ecological association between the species.

In order to identify the most significative associations, we apply two filtering steps to the matrix A returned by our model. 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 *biogeographical filtering* step aims to further eliminate associations that are predicted to exist from the latent representations of the species, but are not supported by the species occurences because they break some biogeographical constraints.

For instance, facilitation between two species requires co-existence. Thus, we set to zero any inferred positive effect involving two non-co-occurring species ([26]). On the other hand, competition does 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 for us to consider a potential negative association to be valid. Specifically, we compute the ranges of the environmental values corresponding to the occurrences of either 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, which capture the species' respective habitat preferences.

Furthermore, we are often more interested in the polarity of the associations, rather than their precise strength, hence we consider a discretized version of the association matrix, which we call the *interaction 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 interaction matrix can be seen as defining 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. In this context, infering associations can be seen as an edge prediction task, i.e. the task of predicting the existence and polarity of an edge for any ordered pair of species.

Assuming that the ground-truth network of inter-species interactions is available (in the form of the simulation parameters in the case of simulated populations, for instance), we can evaluate the performance of our model on this prediction taks by comparing the network constructed from inferred parameters in matrix A to the ground-truth using standard multi-class classification performance metrics (recall, precision, F1-score).

Note that species with similar response embeddings constitute clusters of rows in the association matrix, called *response groups*. Conversely, species with similar effect embeddings constitute clusters of columns in the association matrix, called *effect groups*. In the associated network, the product of both types of groups, results in the emergence of clusters of exchangeable species occupying the same *structural roles* in the resulting interaction network [8].

2.4 Validation on simulated data

Before applying our proposed model on real-world datasets, we perform a validation experiment in a controlled setting. That is, we evaluate the ability of our model to recover interspecific interactions from simulated datasets of known interactions.

4 We set up an experiment similar to (Pollock et al, in prep) where multiple simulations are run on 500 random points on a gradient between 0 and 100 with different configurations of the prior interaction matrix. Specifically, four configuration modes were tested: absence of interactions, positive interactions only, negative interactions only and both positive and negative interactions. In each mode, we vary the number of species (5, 10 or 20), the proportion of interacting pairs (sparse or dense) and whether the interaction matrix includes asymmetric effects. Positive (resp. negative) effects were all set to +1 (resp. minus −1) as we are interested in the polarity of the interactions rather than their magnitude. The full-factorial design of this experiment produced 29 simulation datasets, summarized in Figure 4).

5

pool size	*density*	*directionality*	
Abiotic filter + positive associations			
5 species	sparse	symmetric	□
	dense	symmetric	■
10 species	sparse	symmetric	□
		asymmetric	◇
	dense	symmetric	■
		asymmetric	◆
20 species	sparse	symmetric	□
		asymmetric	◇
	dense	symmetric	■
		asymmetric	◆
Abiotic only			
5 species	~~dense~~	~~symmetric~~	■
10 species	~~dense~~	~~symmetric~~	■
20 species	~~dense~~	~~symmetric~~	■

pool size	*density*	*directionality*	
Abiotic filter + negative associations			
5 species	sparse	symmetric	□
	dense	symmetric	■
10 species	sparse	symmetric	□
		asymmetric	◇
	dense	symmetric	■
		asymmetric	◆
20 species	sparse	symmetric	□
		asymmetric	◇
	dense	symmetric	■
		asymmetric	◆
Abiotic filter + pos. and neg. assoc.			
5 species	sparse	symmetric	□
	dense	symmetric	■
10 species	sparse	symmetric	□
	dense	symmetric	■
20 species	sparse	symmetric	□
	dense	symmetric	■

Figure 4: Design of the simulation experiment.

☞ **NOTE FROM EG:** I am not sure what it meant when a field was absent in the simulations configurations (except maybe dense/sparse for 5 species, as it is so small)

☞ **NOTE FROM EG:** the connection between the paragraphs above and below are not clear to me. Clarify what are the parameters given to the simulation (matrix of interactions, ...), and highlight them below. Did you use random or hand-crafted initial compositions?

2

2.4.1 Data Generation

3 We use a process-based stochastic model adapted from `Virtualcomm` (Gallien and Münkemüller 2015) 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. For simplicity, the spatial structure of communities and thus dispersal processes are ignored. In other words, there is no exchange of individuals between neighboring communities.

6 The simulation starts with a given or random initial composition for each community independently. Individuals are replaced through time until an equilibrium state is reached or a user-defined number of iterations is completed. In the end, the final composition of the communities is returned as the result of the simulation.

2.4.2 Model training

For each simulated population dataset, we count the number of individuals of each species in each site to produce a site-by-species abundance matrix and binarize these counts to produce a site-by-species occurrence matrix. As our HSM models for the abiotic response of the species, we use independent Generalized Linear Models (GLM) with logistic links and one quadratic term.

We apply the proposed association inference model with a negative binomial distribution to fit the species counts. We set the offsets for each species as the average value of its abundance on its sites of presence (i.e. $o_i = \bar{y}_i$). We also add lasso penalties with $\lambda = 0.01$ on response and effect embeddings to promote the sparsity of their products.

To adjust the embeddings dimension d, we use a 5-fold cross-validation scheme where we monitor the deviance of the predicted abundances. Having set d to the value that minimizes the deviance, we train the model on 1000 bootstrap samples from the training set.

☞ **Note from EG:** how do you measure deviance (is it really deviance or deviation?)

$$D = 2[\mathcal{L}(y_i) - \mathcal{L}(\mu_i)]$$

$$d(y,\mu) = 2\left(y \log \frac{y}{\mu} - y + \mu\right)$$

2.5 Evaluation on a dataset of Alpine plants

In this part, our objective is to evaluate the ability of the model to recover meaningful associations from empirical observations of species abundances.

2.5.1 Data preparation

In this evaluation, we use the plant dataset from ([5]). The data records the counts of 84 plant species, collected in July 2000 on 75 units, of size 5 × 5m each, distributed along a mesotopographical gradient. In addition, a set of environmental and topographic variables is recorded for each unit, namely

slope : the slope inclination in degrees at the site,

snow : the average snowmelt date in Julian days between 1997 and 1999,
physd : the percentage of non vegetated soil due to physical processes,
zoogd : the percentage of non vegetated soil due to marmot activity,
aspect : the relative south aspect, and
form : the microtopographic landform index.

We apply a one-hot encoding scheme to the categorical features Aspect and Form and we scale the numerical features.

2.5.2 Model training

We split the observations into training and test using a multi-label stratification scheme (`scikit-multilearn` Python library[1]) to ensure that all species are covered and their proportions are preserved in both sets.

For each plant species, we pretrain an independent Generalized Linear Model (GLM) with a logit link to predict its occurrences from the environmental features. We use the learnt weights as initial parameter values in the habitat suitability component of our model.

We define the biotic context for a target species as the set of plants observed on the location of interest. We use a negative binomial distribution to fit the plant counts. The embedding vectors are initialized using random samples from a uniform distribution on the $[-0.01, 0.01]$ interval, and subjected to lasso penalties to promote sparsity. Finally, the offset value for each species is set to its average count on occurrence points.

We proceed to training the full model using Stochastic Gradient Descent (with a learning rate of 0.01 and momentum of 0.8) on the counts training dataset with a negative examples subsampling rate of 25%. We monitor the negative log-likelihood of positive examples (presences) on the validation set after each full pass of the training set to assess the convergence of the training. We stop when the loss stops decreasing or when 200 epochs have elapsed.

3 Results

3.1 Recovering simulated associations

In this section, we present and analyse the results of our validation experiment on simulated data.

3.1.1 Association classification performances

We begin by evaluating the performance of our model on an independent test set using two metrics: the area under the ROC curve (AUC) for the presence probabilities predicted by the HSM component and the deviance of predicted abundances on positive examples. Finally, we compute the 95% confidence interval of the inferred association matrix's mean.

[1] `http://scikit.ml/`

☞ **Note from EG:** That does not seem to appear anywhere in the results plots

We compare the ability of our model to recover the simulated associations. To do so, we consider the simulation association classes to constitute the ground-truth, against which we compare the inferred associations using standard multi-class performance metrics (recall, precision, F1-score).

The results are presented in Table 2. On average, recall does not vary significantly between positive and negative associations, whereas precision is higher for negative rather than for postive associations. The prediction of positive and negative has low precision, indicating the detection of spurious associations. Much higher precision is achieved for neutral associations. The prediction performance is better on smaller datasets, with higher recall on the dense configurations but higher precision on sparse ones. The sparse asymmetric positive simulation resulted in the worst predictive preformance.

☞ **Note from EG:** what is the (average) distribution of negative/positive/neutral in the ground truth for the different cases

Association type	Recall (%)	Precision (%)	F1-score (%)
Neutral	60.75	98.64	74.50
Negative	72.00	34.02	41.23
Positive	77.60	17.60	26.72
Macro-averages	62.45	94.71	73.09

Table 2: Association classification performances

On the other hand, biogeographical filtering increases the recall on neutral associations(more correctly identified non interactions) and the precision of positive and negative associations (less spurious interactions).

3.1.2 Simulated patterns vs. inferred associations

☞ **Note from EG:** This section is difficult to understand. As far as I get it, the point of the two measures below is to capture how clearly the association types specified by the simulation parameters actually appear in the simulated data. Where the pattern are a bit obfuscated, it explains why the model might not be able to recover the association as expected. So the results from the two measures are not an evaluation of the prediction but help explain it, when put next to the metric in the paragraph above, right?

In fact, the association types specified by the simulation parameters might not be clearly reflected in the simulated populations. As a proxy for the complexity of the inference problem and a simulation diagnosis tool, we evaluate the ambiguity of the patterns of species dependencies present in the simulated communities for the chosen simulation parameters. To do so, for a given pair of

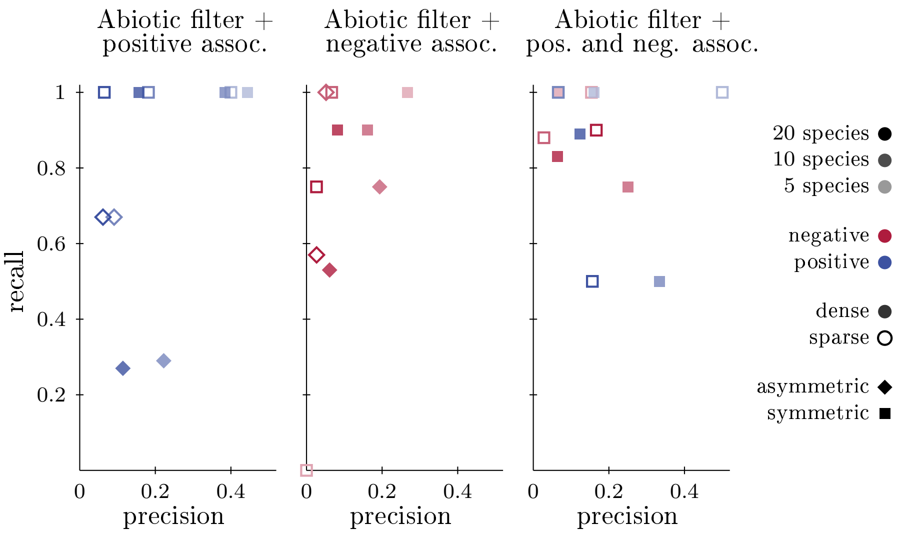


Figure 5: Recall values per simulation groups. `https://chart-studio.plot.ly/~socco/53`

source s and target t species, we define two indices that quantify respectively the overlap in occurrences and the relative variation in abundance of the two species.

The *co-occurrence index* (J) measures the overlap between the occurrences of the two species. It is a symmetric index, computed as the Jaccard similarity between the set of occurrences of either species.

$$J_{st} = J_{ts} = \frac{|\{k \in \mathcal{K}, y_{ks} > 0 \text{ and } y_{kt} > 0\}|}{|\{k \in \mathcal{K}, y_{ks} > 0 \ \text{ or } \ y_{kt} > 0\}|}$$

Values of $J_{s,t}$ close to 1 indicate high overlap whereas values close to 0 indicate near-separation. Values around 0.5 could indicate average overlap, no conclusion can be drawn.

☞ **NOTE FROM EG:** I am not sure what average overlap means. The expected overlap also depends on the total number of sites

The *relative abundance increase* (Δ_y) measures the relative increase in the abundance of the target species when the source species is present. It is an asymmetric measure computed by comparing the abundances of the target species in sites where the source is also present to the average abundance of the target species over all sites where it occurs

$$\bar{y}_t = \text{avg}(\{y_{kt}, k \in \mathcal{K} \text{ s.t. } y_{kt} > 0\})$$
$$\Delta_{st} = \{y_{kt} - \bar{y}_t, k \in \mathcal{K} \text{ s.t. } y_{kt} > 0 \text{ and } y_{ks} > 0\}$$

The larger the standard deviation $\mathrm{std}(\Delta_{st})$, the more ambiguous the strength of the effect of species s on species $\underline{s}$. If the interval $\mathrm{avg}(\Delta_{st}) \pm 1.96\,\mathrm{std}(\Delta_{st})$ does not contain zero, then the simulated dependencies unambigously translate a polarized effect of species s on species $\underline{s}$. Otherwise, the polarity of the effect is ambiguous, often due to confounding effects of other species or a neutral association if the mean is close to zero.

We find that the observed patterns told different stories about the actual associations. The co-occurrence index was higher than 50% for all species with at least one positive effect. However, the species pairs involved in negative associations appeared together more than expected under the independence assumption in some runs and less on others, in particular on bigger pools. Neutral associations induced co-occurrence indices spanning a large spectrum below 70%.

The average relative abundance increase 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 average RAI absolute values as compared to the negative effects. Although more clearly marked, the latter approached neutrality on bigger and more densely connected communities.

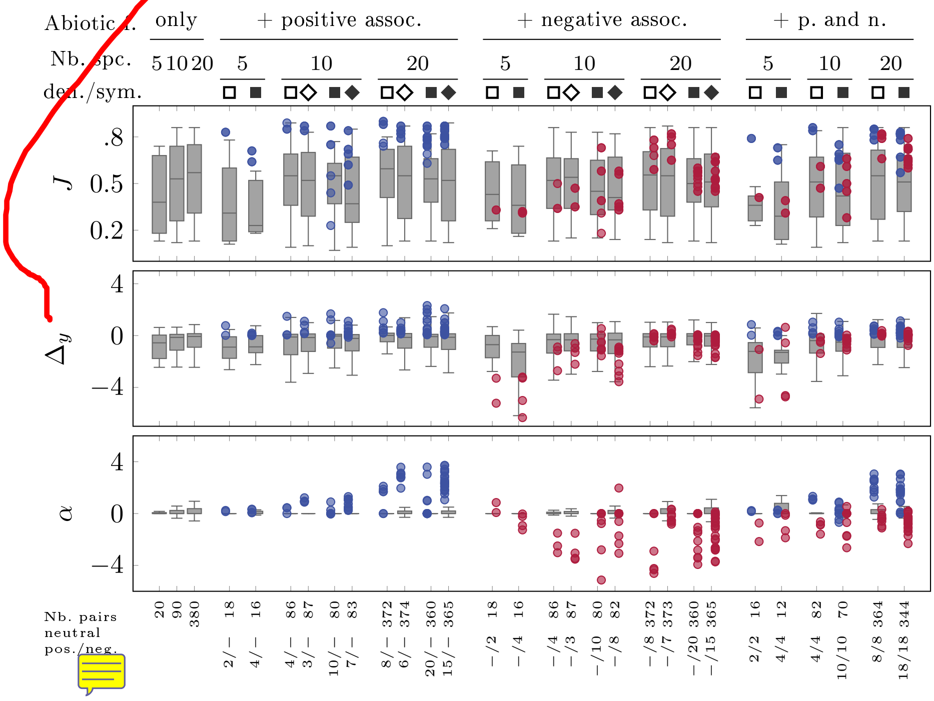


Figure 6: Distribution of co-occurrence, relative abundance effect and inferred association strength per type of association. `https://chart-studio.plot.ly/~socco/30`Interactive mode

The inference model was able to discriminate positive from negative effects while maintaining an average value for non interacting pairs centered on zero. On simulations with a dense mix of positive and negative associations, both observed effects and inferred associations were close to zero, possibly due to opposite effects canceling each other.

3.2 Abiotic and biotic drivers of Alpine plant distributions

In this section, we present and analyse the results of our evaluation experiment on the Alpine plant dataset.

3.2.1 Hyperparameter search and model selection

The first step in this evaluation is to find good values for the hyperparameters of our model. For a species pool of size m, the embedding dimension d is best selected among powers of 2 up to $m/2$, to improve hyperparameter search speed. In our case, with $m = 82$, the embedding dimension is chosen from the set $\{2, 4, 8, 16, 32\}$.

When the value of the lasso penalty parameter λ becomes large, some components of the embedding vectors take extremely small values for all species (below 10^{-5}). These components have no effect on the computed associations. Removing them, shrinks the embeddings to a smaller effective dimension, equal to the number of retained components. In the extreme, very high values of λ lead to effective dimension equal to zero, resulting in a zero association matrix, so that the interaction model is only parameterized by the species offset counts.

For each value of d, we apply the training procedure described previously with increasing values of $\lambda \in \{0.01, 0.015, 0.02, 0.025\}$. We evaluate the resulting models on the test set by computing the effective dimension and the deviance of the predicted counts on positive examples. We summarize the model selection results in Figure 7.

3.2.2 Plant habitat suitability

☞ **Note from EG:** does that mean you evaluate the prediction performance of the entire model, or only the HSM block?

Next, we analyze the parameters and performances of the HSM fitted on plant occurrences. Results are summarized by plant genus groups in Figure 8. The model predicts habitat suitability with at least 70% AUC score for all genuses. The analysis of environmental variable importance shows the dominance of Snow duration followed by zoogenic disturbances, the site form and aspect. Physical disturbance and slope weights were neglectible, probably due to their correlation with Snow.

3.2.3 Analyzing the functional meaning of plant embeddings

Then, we investigate the functional determinants of the associations diversity. To do so, we compute the mutual information between the learnt embeddings and the plant traits (a full list and documentation of the traits are provided in

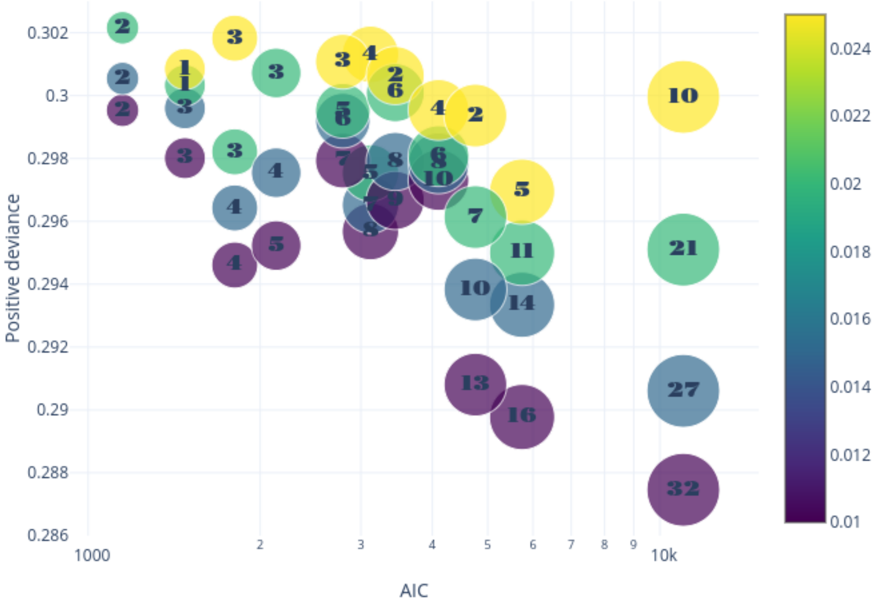


Figure 7: Positive deviance as a function of the Akaike Information Criterion (AIC). Each point represented with a circle indicates a configuration of the hyperparameters. Circle size is proportional to embedding size. Circle color represents the value of the lasso penalty parameter λ ranging from 0.01 (darker blueish colors) to 0.03 (brighter yellowish colors). The labels correspond to the effective dimension (non-zero components). Larger embeddings lead to a higher model complexity which explains the increasing trend in AIC values with the increase in embedding size. Higher values of λ (yellowish points) result in fewer retained components and larger deviance scores. The combination ($\lambda = 0.01, d = 4$) provides the best compromise between model complexity and performance. `https://chart-studio.plot.ly/~socco/45`

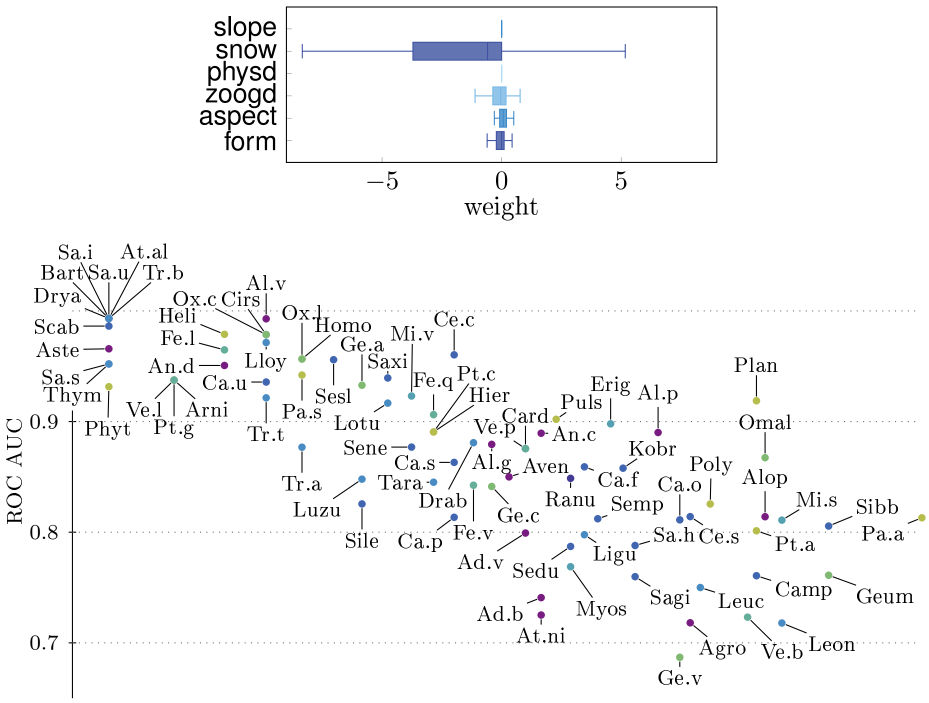


Figure 8: Habitat Suitability Model parameter interpretability and prediction performances.https://chart-studio.plot.ly/~socco/1/

the Supplementary Materials). The Mutual Information [28] is an unbounded symmetric and positive score that measures the amount of information contained in one random variable about another. It quantifies the reduction in uncertainty about one random variable given knowledge of another. Zero mutual information indicates independence.

In general, we expect traits related to dispersal capabilities (seed, spread) to impact the prevalence of the species, consequently increasing or decreasing the opportunity to affect other species (interaction probability). As a result, we expect ~~that~~ such traits to have a higher mutual information with effect embeddings than with response embeddings. Conversely, traits related to nutrient uptake and biomass accumulation capture competitive or cooperative abilities of the plant species. Hence, we would expect a high mutual information between these traits and both responses and effects embeddings.

A histogram of the mutual information between pairs of trait and embeddings is shown in Figure 9. The result show a relatively significant contribution of the Nitrogen mass and Spread to the plants response, whereas the angle was found independent. The Specific Leaf Area contributes significantly to the effect in addition to the Nitrogen mass and on a lesser extent height, spread and seed.

> ☞ **NOTE FROM EG:** I think it is somewhat missleading to present them stacked in this way because the information is not necessarily additive like this, some of the same information about the traits might be encoded in different dimensions of the embedding, if I understand correctly

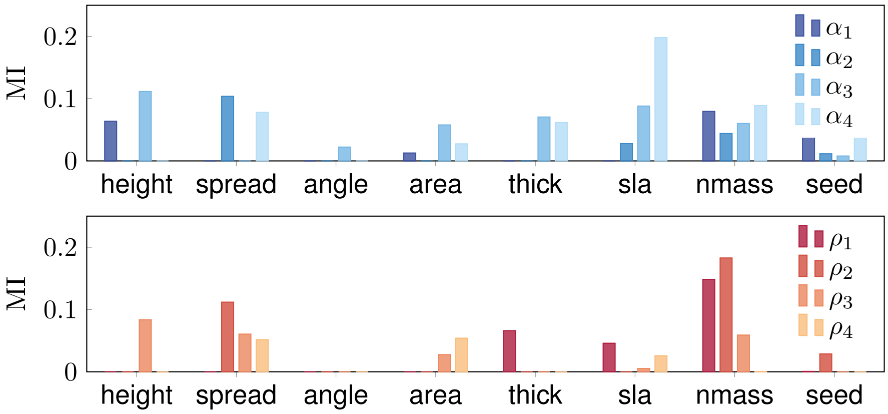


Figure 9: Mutual information between plant traits and their latent representations. Each bar concerns a specific trait, it represents the stack of mutual information scores from the first to the last (fourth) embedding dimension. The lower (resp. upper) figure shows the results for the response (resp. effect) embeddings. `https://chart-studio.plot.ly/~socco/70`

3.2.4 Plant associations on a mesotopographic gradient

Finally, we analyze the inferred associations and the overall network structure in light of existing literature on alpine plants interactions [6].

We perform a hierarchical clustering on both rows and columns of the association matrix to obtain effect and response groups, which are displayed in Figure 10a. In parallel, we apply the modularity maximization algorithm on the association network to identify densely connected modules, referred to as communities [8]. After that, we map the structural roles to the modules to create the group-level network.

☞ **Note from EG:** this needs more explanations

Four densely connected modules stand out. They are structured along the mesotopographical gradient.

1. High-altitude, dry and windy communities densely connected by positive associations: (i) unselective commensalism of forbs by dominant graminoids, (ii) facilitation between graminoids and tall herbs. The latter act as hubs connecting the high elevation sites to the adjacent sites.

2. Mid-gradient communities composed of two groups: (i) Tall herb grasslands occuring in favorables conditions, mostly structured by negative associations (ammensalism and competition); (ii) Short herb meadows, prone to zoogenic disturbance. They present higher abundances when cooccuring with tall herbs. Through their spreadth they also play a central role in connecting extreme and favorable sites' communities.

3. Chinopholous (cold-resistant) vegetation appearing on late-melting sites. They cohabitate positively with some short-herbs but are negatively affected by forbs and tall herbs.

4. North-facing isolated communities dominated by Salix Herbacea negatively associated with high-altitude communities and characterized by high eccentricity.

Modules identified on the basis of link density by the modularity maximization algorithm are spatially structured. Hence, the network of modules reflect the spatial connectivity and as a result the turnover over the gradient. Within each commmunity, we identify various subgroups based on different responses (incoming edges) or effects (outgoing edges) to other groups.

For instance, the high-elevation module (pink nodes) comprises three subgroups with different structural roles. Graminoids dominate these communities, they provide wind protection to forbs and some tall herbs (Festucal Violacea, Trifolium Alpinum). The latter also occur on grasslands where they compete with each other. Hence, this subgroup is separated from other forbs despite their cohabitation and the similar response to graminoids.

As reported in the literature ([1],[5]), the abiotic stressors strongly structure the distribution of the plant species and the dominant interaction types. Indeed, more positive associations are reported in stressful conditions (high-elevation and late-melting sites). Particularly, the average snow duration seems to be the major structuring force.

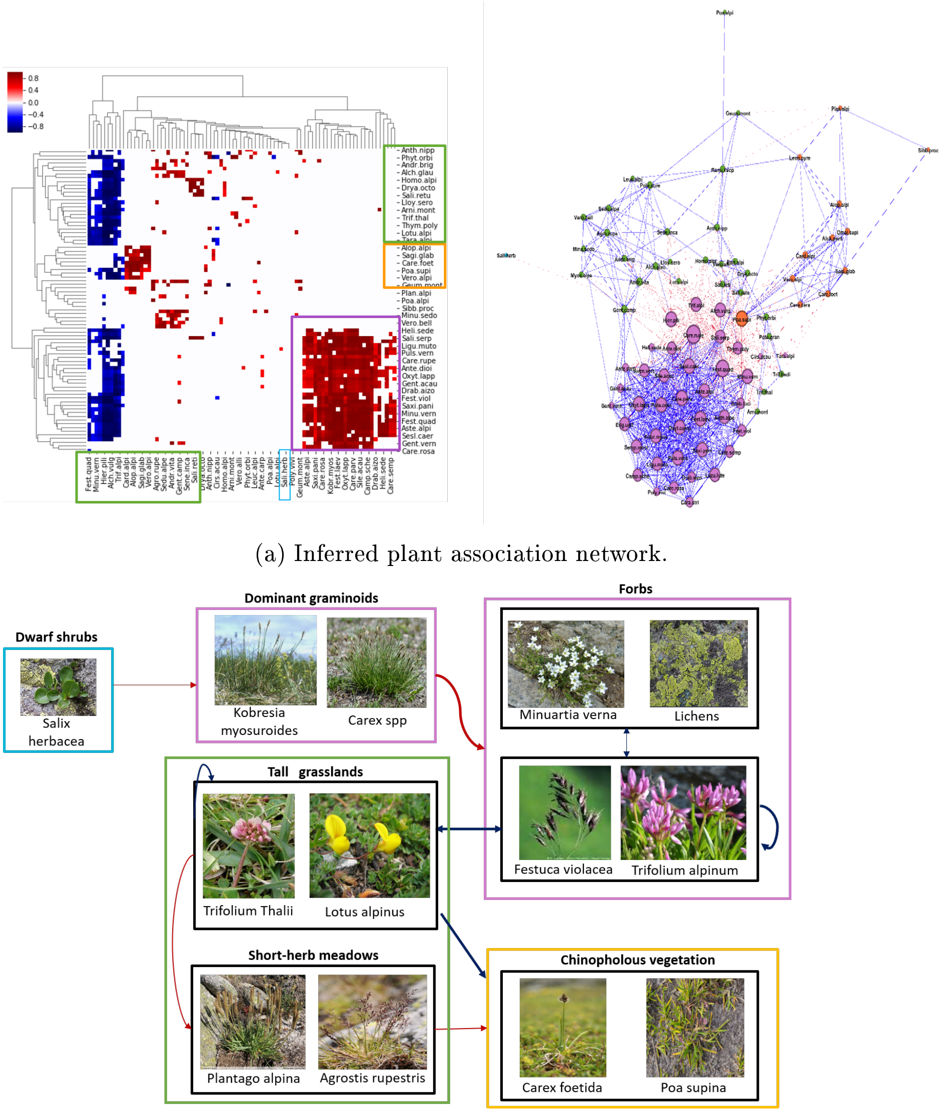
4 Discussion

We first develop a probabilistic model of multispecies abundances that accounts for habitat suitability as well as biotic associations. We parameterize this model with an asymmetric association matrix computed from two sets of low-dimensional embeddings representing: the effect present species have on others' abundance, and the response to other species' effects. We then use the model to infer associations given species abundances.

We use a process-based simulation model to generate synthetic community datasets. By analyzing the observed co-occurrence and pairwise conditional abundances, we note that co-occurrence levels can be high even on known competing pairs if the species pool is small and the communities' carrying capacity is large. On the other hand, the abundance reflects better the nature of associations as it is lower than 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 (Boulangeat et al 2012). Consequently, one should model the pairwise associations conditioned on all other species in order to isolate the different opposing influences. For instance, given a triplet A, B, C such that A (resp. C) influences B positively (resp. negatively). If the same strength is applied by A and B we may observe that all pairs are independent, unless we evaluate the association $A- > B$ conditioned on C and $C- > B$ conditioned on A. In any case, we still need separate observations where we could quantify both associations separately.

Thereafter, we test the ability of the model to recover simulated associations. We find that the model is able to disentange positive, negative and neutral associations. But, it generates many spurious associations due to high levels of co-occurrence induced by the simulation model. ~~As opposed to Joint Species Distribution Models~~, our model is able to recover positive associations between species with similar abiotic niches, for two reasons. First, it does not rely on residuals unexplained by the SDM. Instead, it conditions on the habitat suitability for both species hence the probability of them co-occurring would induce a potential for facilitation. Second, the association is evaluated on the basis of abundance variation in presence of the other species.

Afterwards, we apply the inference model on plants distributed along a mesotopographical gradient in the French Alps. The associations identified by the algorithm conformed to most of the important specific relationships that we expect to find in the upland plant community ([6]), giving us confidence in the novel methods presented here. We observe more positive associations in extreme conditions (dry, windy or frozen) and negative ones in favorable sites, confirming the stress-gradient hypothesis ([1]). We illustrate using the results on Aravo dataset some of the possibilities allowed by the model. For instance, we can use the association matrix to populate a network, study its modular



(a) Inferred plant association network.

(b) Summary network of plant associations.

Figure 10: Plant associations on an alpine mesotopographic gradient. Subfigure (a) shows the normalized pairwise associations on the left and the corresponding network on the right. Blue (resp. red) edges indicate negative (resp. positive) edge weights. Node colors on the graph represent communities identified by the modularity maximization algorithm [20]. We highlight the communities composition using colored squares on the matrix on the right. Species in the association matrix are grouped based on a hierarchical clusterings performed rowise (yielding response groups) and columnwise (yielding effect groups). Subfigure (b) summarizes the association network at the level of previously identified groups. Edge thickness is proportional to the number of links between the interacting group pair.

structure through community detection algorithms and analyze the structural roles, including effect and response groups, occupied by species within communities. Such information is useful to evaluate the functional redundancy within communities. Second, we argue that the learnt representations can be related to functional traits of the species, providing the appropriate traits are measured, to identify the functional drivers of network structure.

It is now agreed upon that inferred associations are not biological interactions. They represent significant spatial colocation or dislocation patterns, that are informative in a predictive rather than causal way ([19]). 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).

The problem of inferring associations from co-occurrence is a major part of ecological modeling publications, as much as biomedical and bioinformatics literature. Existing approaches include JSDMs ([21]) which rely on explaining residual correlations by biotic effects but cannot by design model asymmetric associations. Other major state of art methods are based on probabilistic graphical models including Markov Random Fields ([13],[4]) and Bayesian networks (Mils, Trifonova, Aderhold et al 2012). The former yields undirected networks while the latter impose a directed acyclic structure to the network of associations forbiding feedbacks and the modeling of asymetric effects.

The most challenging part of this task is that it is completely unsupervised, with no prior or guidance on the expected associations or network. Hence, validating the resulting associations is tricky. It is still feasible to evaluate the veracity of the type of associations by using an edge classification scheme and a list of potential interactions as we've shown. However, it is far more difficult to validate the strength of associations especially when working with snapshot data. As many associations would have a strength around zero, it is also valuable to ask whether one should use higher thresholds to decrease the amount of spurious associations and increase model precision and how to select these thresholds ? Moreover, because many processes influence community assembly, multiple scenarios could lead to the same communities making this problem unidentifiable. In this case, we need not one expected list of associations but all the possible ones or a goodness-of-fit measure that accounts for equivalence between different association combinations.

A possible way to prevent the unidentifiability issue is to include known information on ecological interactions in the model [2]. For instance, [29] use a Bayesian network with a structure defined a priori, and trains its parameters using an SDM 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 regional metaweb.

There is now growing evidence that ecological interactions are context-dependent ([22], [30]), we show in appendix B how 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 (Tikhonov et al 2017, Clark et al in prep). Despite the new possibilities offered by these developments, the question of how to validate their results still araises itself.

Finally, although the main purpose of the model is to infer species associations, it is also useful to make conditional predictions of abundances. Another possible usage is to perform link prediction on an incomplete network. The idea is to complete the associations involving a target species by leveraging information from similar node species.

5 Conclusion

Biological interactions and other processes induce spatial patterns of co-occurrence. Our objective is to disentangle these processes and isolate species dependencies. We present a model of species co-abundances as a function of the habitat and biotic associations. We propose an asymetric 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.

A Derivatives of the optimized objective

Canonic derivatives (refer to Liping liu et al 2017).

B Extensions of the biotic context definition

B.1 Adding conditioning covariates

In the base model, the estimation of any pairwise interaction is oblivious to the abiotic or biotic conditions surrounding it. To account for these neighborhood conditions, we extend the base model by allowing the embeddings used to represent the biotic context to depend on some chosen variables.

Each site is associated to p conditioning covariates, These covariates are stored alongside an offset in a $n \times (p+1)$ matrix V, such that each of the first p columns of V contains the values of the corresponding covariate for the different sites while the last column is filled with ones. Then, given an embedding dimension d, the covariates are mapped to d dimensions by applying a regression with a weight matrix $W \in \mathbb{R}^{d\times(p+1)}$. The resulting conditioning vectors are such that $\beta_k = W v_k^T$.

The extended biotic context is then written as follows, where $\odot$ is the element-wise vector product:

$$z_{ki} = \beta_k \odot \Big(\frac{1}{|C_{ki}|} \sum_{j \in C_{ki}} y_{kj}\rho_j\Big) = \frac{1}{|C_{ki}|} \sum_{j \in C_{ki}} y_{kj} \cdot (\beta_k \odot \rho_j)$$

The biotic associations can be recovered as in the base model, by isolating the pairwise interactions in the response variable. However, in this case, the associations we obtain are represented by a three-dimensional tensor instead of a two-dimensional matrix. Each slice along the first dimension of this tensor represents a local association network.

$$a_{kij} = \sum_{l=1}^{d} (\beta_k \odot \alpha_i \odot \rho_j)_l$$
$$\eta_{ki} = f\big(\sum_{j \in C_{ki}} y_{ki} a_{kij} + o_j \big)$$

By incorporating the environmental covariates on the latent space, we gain two desirable properties. First, we get a fixed number of parameters that is a factor of the embedding dimension, which is significantly smaller than the number of modeled species. Second, we ensure species with similar latent traits, as captured by the response and effect embeddings, share associations regardless of the surrounding conditions. As a result, response or effect groups of species computed from the learnt embeddings remain consistent in the environmental space.

B.2 Temporal extension

When longitudinal data are available, we denote the abundance of species i at site k at time-point t as $y_{ki}^{(t)}$. Accordingly, the definition of the biotic context for a target species at a given time-point is extended to contain the species, including the target, that were observed in the previous time-point:

$$C_{ki}^{(t)} = \{ j \in \mathcal{S}, y_{kj}^{(t-1)} > 0 \}$$
$$z_{ki}^{(t)} = \frac{1}{\left| C_{ki}^{(t)} \right|} \sum_{j \in C_{ki}^{(t)}} y_{kj}^{(t-1)} \rho_j$$

B.3 Spatial extension

Given a function d that measures the distance between any pair of sites and a radius r, we consider a spatial extension of the base model where the biotic context is defined to contain species that were observed at locations within distance r of the considered site.

$$C_{ki} = \{ (j,l) \in \mathcal{S} \times \mathcal{K}, y_{lj} > 0 \text{ and } \mathrm{d}(k,l) \leq r \}$$

One can use multiple radius values customized to the dispersal abilities of each target group or species for instance. The effect of each contextual element is weighted in inverse proportion to its distance to the target location. The hyperparameter τ controls the decrease in weight per unit of distance. Similarly, τ can be customized for each group of species based on expert knowledge.

$$z_{ki} = \sum_{(j,l) \in C_{ki}} y_{lj} \cdot \exp(-\tau \, \mathrm{d}(k,l))$$

B.4 Graph extension

So far, we defined the biotic context using the community composition in terms of species, possibly involving their abundances. At this point, we are able to capture pairwise additive effects. However, we miss the impact of interactions between context species or the whole network structure around the target location on the abundance distribution of the target group.

Fortunately, graph embedding algorithms permit the incorporation of structured data such as knowledge graphs into predictive models. For instance, we can redefine the biotic context as the interaction network at site of interest k minus the target species i, noted $G_{k/i}$. The context embedding is then obtained by applying a graph kernel function k with parameter θ on the contextual network

$$z_{ki} = \mathrm{k}(G_{k/i}; \theta)\,.$$

References

[1] 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.

[2] 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.

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

[4] Chiquet, J., Mariadassou, M., and Robin, S. Variational inference for sparse network reconstruction from count data. *arXiv preprint arXiv:1806.03120* (2018).

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

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

[7] 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.

[8] 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.

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

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

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

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

[13] Harris, D. J. Inferring species interactions from co-occurrence data with markov networks. *Ecology 97*, 12 (2016), 3308–3314.

[14] Humboldt, A. v., Bonpland, A., et al. Essai sur la géographie des plantes.

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

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

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

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

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

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

[28] Shannon, C. E., and Weaver, W. A mathematical model of communication. *Urbana, IL: University of Illinois Press 11* (1949).

[29] 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.

[30] 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.

[31] 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.