Learning bidirectional ecological associations from co-occurrence and environmental data

Supplementary Materials

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

February 27, 2020

Contents

1 Supplements to framework description

1.1 Inference procedure in details

We outline the inference process, i.e. the procedure for training the model. For a site k, we define the vector of observed species' abundances:

$$y_k = (y_{ki})_{i \in \mathcal{S}} \,.$$

We gather as positive and negative observations (instances), the species that are present and absent at each site, respectively:

$$\begin{aligned} \mathcal{S}_k^+ &= \{i \in \mathcal{S}, y_{ki} > 0\}\,, \\ \mathcal{S}_k^- &= \{i \in \mathcal{S}, y_{ki} = 0\}\,. \end{aligned}$$

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 $\mathcal{S}_k^-$ are sub-sampled randomly (at rate $r\%$) at each training iteration. By introducing randomness into the objective function, this sub-sampling procedure also improves the robustness of the estimations and prevents over-fitting. Furthermore, we promote the sparsity of the embeddings and of the resulting association matrices by adding regularizers on the embedding vectors.

In summary, given the training data (abundances Y and abiotic variables X) together with the user-defined hyper-parameters (including hyper-parameters for the abiotic suitability model ϕ_h, embedding dimension d, vector of species offsets O, negative examples subsample rate r and regularization coefficient λ), the model training procedure aims to infer the values 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 $\mathcal{L}$. In other words, the model is trained to maximize the penalized likelihood $\mathcal{L}$ of the observed abundances on each site for all positive instances and the sampled subset of negative instances.

$$\begin{aligned} \mathcal{L}(\theta_h, P, Q) &= \sum_{k \in \mathcal{K}} \Big(\sum_{i \in \mathcal{S}_k^+} \mathcal{L}_{ki}^+ + \sum_{i \in \tilde{\mathcal{S}}_k} \mathcal{L}_{ki}^- \Big) + \lambda(|P|_1 + |Q|_1) \\ \text{where} \quad \mathcal{L}_{ki}^+ &= \log\left[h_i(x_k;\theta_h)\right] + \log\left[f_{\mathcal{E}}\big(\eta_{ki}(P,Q), \tau(y_{ki})\big)\right] \\ \text{and} \quad \mathcal{L}_{ki}^- &= \log\left[\big(1 - h_i(x_k;\theta_h)\big) + h_i(x_k;\theta_h) f_{\mathcal{E}}\big(\eta_{ki}(P,Q), \tau(0)\big)\right] \end{aligned}$$

To do so, we use Stochastic Gradient Descent [1] as our optimization algorithm of choice. The resulting model can then be applied on hold-out data for validation, and on previously unseen records to predict the abundance of a species given other species' abundances.

1.2 Extensions of the biotic context definition

1.2.1 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.

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

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.

1.2.2 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:

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

1.2.3 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.

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

1.2.4 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, *contextual network*.

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

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

2 Supplements to the simulation experiment

2.1 Simulation how-to

We used 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. 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. The final communities' composition is returned at the end Fig. 2.

2.1.1 Notation

- We start by sampling n locations uniformly on a single environmental gradient E.
- All locations have the same carrying capacity of K individuals from a common pool of m species $S = \{S_j/j \in [1, m]\}$.
- Each species has its own optimal environmental value $\mu_j \in E$ as well as a niche breadth $\delta_j \in E$.
- Biotic interactions are described by a full interaction matrix $I = (I_{jk})/j, k \in [1, m]^2$; $-1 \leq I_{jk} \leq 1$ where I_{jk} represents the effect of the interaction between the pair (S_j, S_k) on species S_k. We also write: $I = I^+ - I^-$ such that:
  - $I^+ = (I^+_{jk})/j, k \in [1, m]^2; 0 \leq I^+_{jk} \leq 1$ represents the matrix of positive effects (facilitation matrix)
  - $I^- = (I^+_{jk})/j, k \in [1, m]^2; -1 \leq I^+_{jk} \leq 0$ represents the matrix of negative effects (competition matrix)

2.1.2 Assembly rules

At each timestep (epoch), given an actual composition c, the probability that an individual from a given species i to replace any other individual of c is given by the following equation ; such that:

- B_{env}: weights of the abiotic filter.
- B_{comp}: weight of the competition.
- B_{fac}: weight of the facilitation.
- B_{abun}: weight of the reproduction filter, can be interpreted in terms of growth rate.
- $P_{env,i,c}$: the probability of species i to occur under the environmental value E_c is given by the normalized density on E_c of a Gaussian distribution parameterized by its optimum and niche breadth. The closer to its optima, the higher the probability of the species' occurrence.
- $P_{comp,i,c}$: the probability for an individual of species i to join the community given the aggregated effect of its competitors in c.
- $P_{fac,i,c}$: the probability for an individual of species i to join the community given the aggregated effect of its facilitators in c.
- $P_{abund,i,c}$: probability of an individual of species i to join the community as a result of the reproduction of some of the $N_{i,c}$ conspecifics in c.

The unnormalized weights $W_{i,c}$ for each species are then normalized by dividing each one of them by their sum. The result is a vector of probabilities W that sums to 1. Finally, we sample from a multinomial distribution, parameterized with W, K individuals to compose the new community.

2.2 Simulation configuration

We summarize the simulation parameters as well as the graphic codes in Fig. 3. Throughout our simulation experiment, we use symbols to picturally and concisely represent combinations of density and directionality. Hollow shape represent a sparse setting whereas a filled shape represents a dense setting. A triangle represent an asymmetric setting whereas a diamond represent a symmetric setting, since they intuitively evoke single-directional and a bi-directional arrows, respectively. Density and directionality of associations is irrelevant for configuration with the abiotic filter only, that are hence represented by a simple dot.

2.3 Framework adaptation and training

For each simulated dataset, we counted the number of individuals of each species in each site to produce a site-by-species abundance matrix and binarized these counts to produce a site-by-species occurrence matrix.

We used a Generalized Linear Model (GLM) with a logistic link and a quadratic term to model the response of species to the environmental gradient to produce the habitat suitability map for each of the simulated species. We then applied the 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 across its presence sites (i.e. $o_i = \bar{y}_i$). We also added 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 used a 5-fold cross-validation scheme where we monitored the deviance of the predicted abundances (see Supplementary Materials).

Given the observed count y and the predicted mean count μ, the deviance is computed as two times the difference between the predicted and the maximum achievable likelihood. The latter is simply the likelihood on the observed count value

$$d(y, \mu) = 2\big(\mathcal{L}(y) - \mathcal{L}(\mu)\big)\,.$$

Having set d to the value that minimizes the average deviance over all species, we trained the model on 1000 bootstrap samples from the training set. Finally, we used the bootstrap estimates to compute the 95% confidence interval of the inferred association matrix's mean.

2.4 Model selection

The model performances for each simulation and for each embedding dimension are reported on Table 4, the embedding dimension was selected based on the best score in terms of deviance.

3 Supplements to the empirical application

3.1 Environmental data preparation

The plant dataset contained the following set of environmental variables:

slope : the slope inclination in degrees,
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 initially applied a one-hot encoding scheme to the two categorical features (aspect and form) and we scaled the numerical features.

3.2 Framework adaptation and training

We split the observations into a training and a test dataset using a multi-label stratification scheme[1] to ensure that all species were covered and their proportions were preserved in both sets.

For each plant species, we pre-trained a generalized linear model (GLM) with a logit link to relate species occurrences to the environmental variables. We used the learnt weights as initial parameter values in the habitat suitability component of our framework.

We defined the biotic context for a target species as the set of plants observed on the location of interest. We used a negative binomial distribution to fit the plant counts. The embedding vectors were 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 was set to its average count on occurrence points.

We trained the full model using stochastic gradient descent (with a learning rate of 0.01 and momentum of 0.8) on the training dataset using a subsampling rate of 25% for the negative examples. We monitored 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 stopped when the loss stops decreasing or when 200 epochs have elapsed.

---

[1] Python library `scikit-multilearn`: `http://scikit.ml/`

3.3 Embedding dimension and lasso parameter selection

The first step in this evaluation was to find appropriate values for the hyperparameters of our model. For a species pool of size m, the embedding dimension d is 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 (Fig. 2).

3.4 Habitat suitability

Performances The model predicts habitat suitability with a $87.7 \pm 0.17\%$ AUC score for all genera (Fig. 5). The analysis of environmental variable importance showed the dominance of snow duration followed by zoogenic disturbances, the site form and aspect. Physical disturbance and slope weights were negligible, probably due to their correlation with snow.

Relationship between performances and prevalence . We illustrate in Fig. 6 the relationship between the habitat suitability score (Area under the curve) and the species prevalence. Overall, the performance decreased with prevalence as would be expected since more prevalence meant larger environmental range, hence more difficulty to identify suitable habitats.

3.5 Analyzing the functional meaning of plant embeddings

We investigated the functional determinants of the associations diversity. To do so, we compute the mutual information between the learnt embeddings and the plant traits (reported in [2]). The Mutual Information [4] 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 mass, 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 such traits to have a higher mutual information with effect embeddings than with response embeddings. Conversely, traits related to nutrient uptake and biomass accumulation potential 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.

There was a relatively significant contribution of the leaf nitrogen mass and spread to the plants response, whereas leaf angle was found independent (Fig. 7). The Specific Leaf Area contributes significantly to the effect in addition to the Nitrogen mass and on a lesser extent Spread. Height is reported as related to both parameters.

4 Discussion of the offset choice

4.1 Interplay of offset choice, niche overlap and carrying capacity

The way we simulated data tended to produce high co-occurrence probabilities due to high niche overlap between species and large carrying capacities. This was especially true for the largest pool sizes that further leaded to large biotic contexts. When applied to these simulated data, our proposed inference model detected a number of spurious weak positive associations. There are two complementary explanations for this. First, a large difference between the chosen offsets (baseline abundances) and the observed abundances forced the model to compensate with a strong positive signal. Second, since the biotic contexts were large, hence the large amount of positive associations, the required positive effect had to be distributed amongst the biotic context members, explaining the low strengths. Conversely, negative associations had to be strong to counterbalance the strong positive signal to provoke absence where habitat suitability was satisfied. The simulations with 5 species departed from this rule because the abiotic filter outweighed the few associated pairs.

4.2 Alternative offset configurations

Unquestionably, the offset choice had an impact on both the interpretation of the associations and their inferred strength. We chose to fix the offsets as the average abundance of species over their presence points. This definition affords a convenient interpretation for biotic associations as explanatory variables of the deviation from the mean abundance. However, it averaged many points where species occurred, albeit with low abundance, outside their abiotic niche mainly due to their assumed unlimited dispersal abilities, leading to underestimated offset abundances.

To overcome this problem, we could restrict the offset computation to suitable habitats only providing they are correctly estimated by the HSM. Alternatively, instead of using a fixed offset, we could fit the expected abundance under no biotic influence as a fixed bias term, as a location-dependent function of the abiotic environment for each species [3] or as the output of a population dynamic model [5].

References

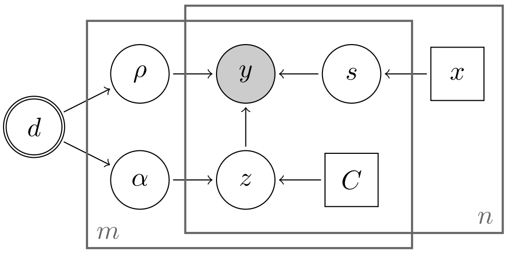
[1] Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

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

[3] Ehrlén, J., and Morris, W. F. Predicting changes in the distribution and abundance of species under environmental change. *Ecology Letters 18*, 3 (2015), 303–314.

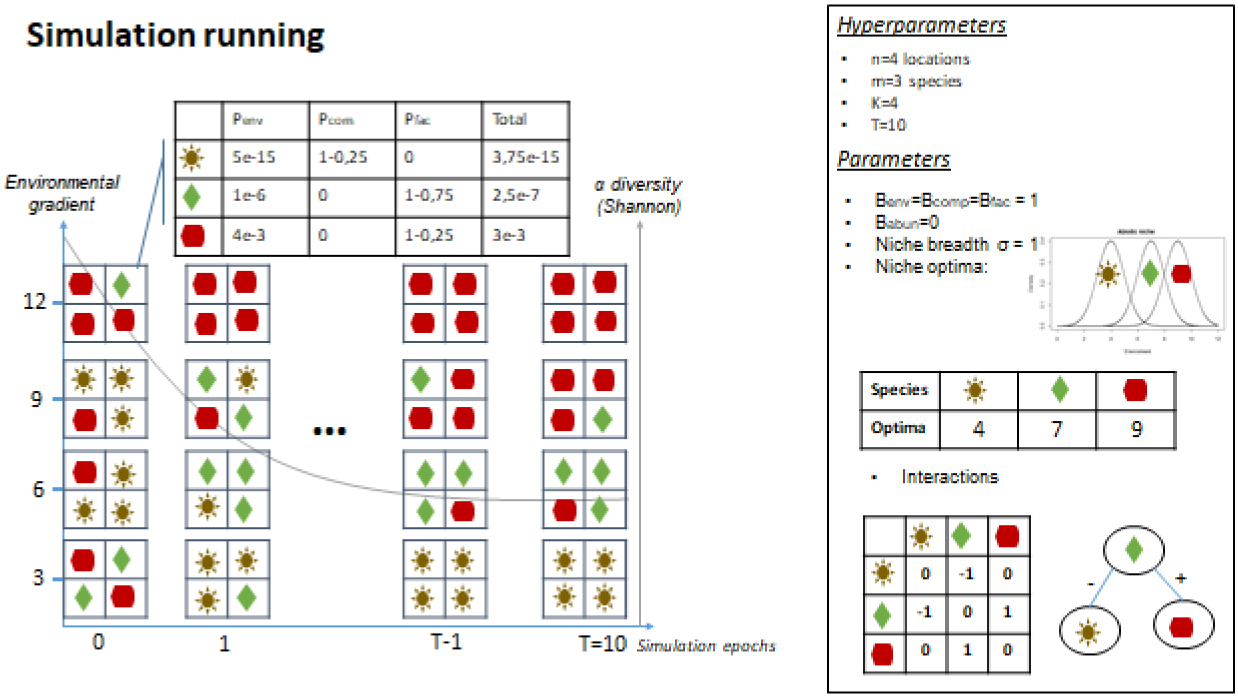
[4] Shannon, C. E., and Weaver, W. A mathematical model of communication. *Urbana, IL: University of Illinois Press 11* (1949).

[5] Soubeyrand, S., Neuvonen, S., and Penttinen, A. Mechanical-statistical modeling in ecology: from outbreak detections to pest dynamics. *Bulletin of Mathematical Biology 71*, 2 (2009), 318.



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
n	Number of observation sites
m	Species pool size

Fig. 1: Plate diagram of the generative model of abundance.



	P_{env}	P_{com}	P_{fac}	Total
☀	5e-15	1-0,25	0	3,75e-15
◆	1e-6	0	1-0,75	2,5e-7
⬣	4e-3	0	1-0,25	3e-3

Hyperparameters

- n=4 locations
- m=3 species
- K=4
- T=10

Parameters

- $B_{env}=B_{comp}=B_{fac} = 1$
- $B_{abun}=0$
- Niche breadth $\sigma = 1$
- Niche optima:

Species	☀	◆	⬣
Optima	4	7	9

- Interactions

	☀	◆	⬣
☀	0	-1	0
◆	-1	0	1
⬣	0	1	0

Fig. 2: Simulation procedure.

pool size	*density*	*directionality*	
Abiotic filter only			
5 species			•
10 species			•
20 species			•
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 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	◆

Fig. 3: Design of the simulation experiment. We list the different configurations corresponding to combinations of poolsize, density and directionality. On the right-hand side we indicate the symbols that are used throughout the simulation experiment to represent combinations of density (sparse:hollow shape/dense:filled shape) and directionality (asymmetric:triangle/symmetric:diamond).

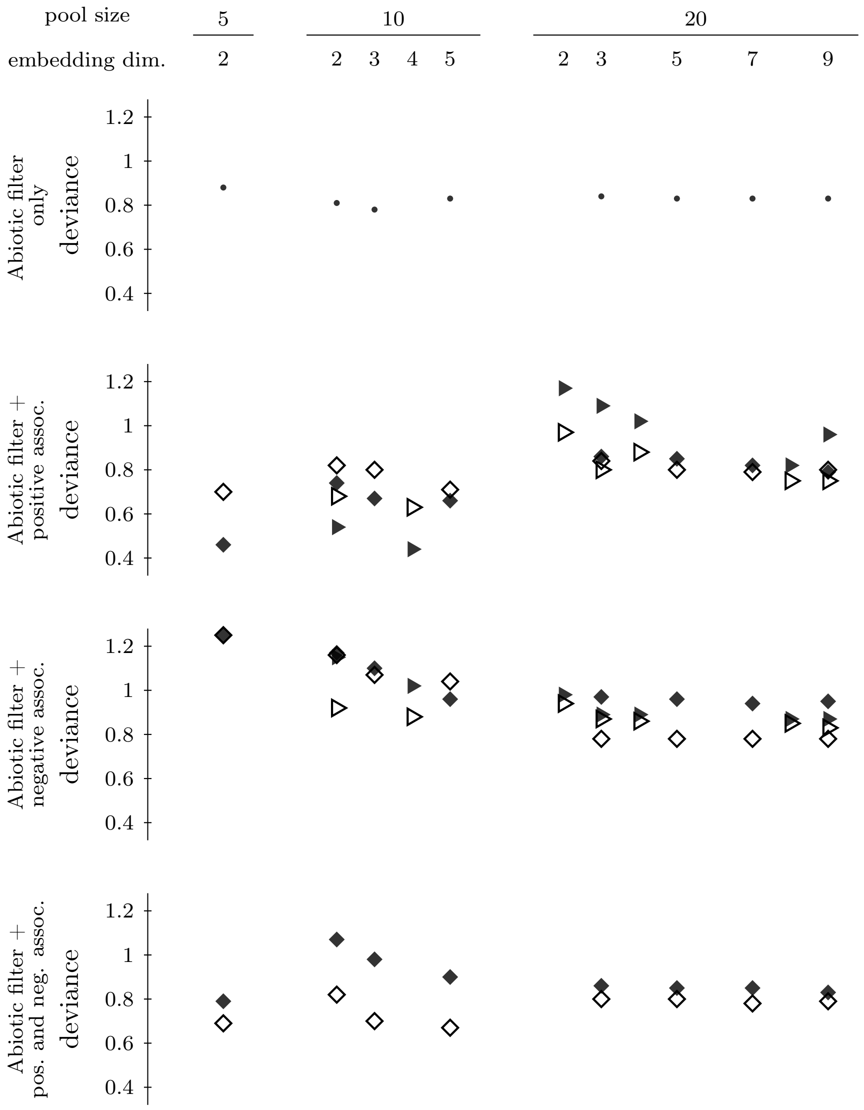


Fig. 4: Deviance of predicted abundances as a function of the embedding dimension for different number of species and embedding dimensions. The symbols represent combinations of density (sparse:hollow shape/dense:filled shape) and directionality (asymmetric:triangle/symmetric:diamond)

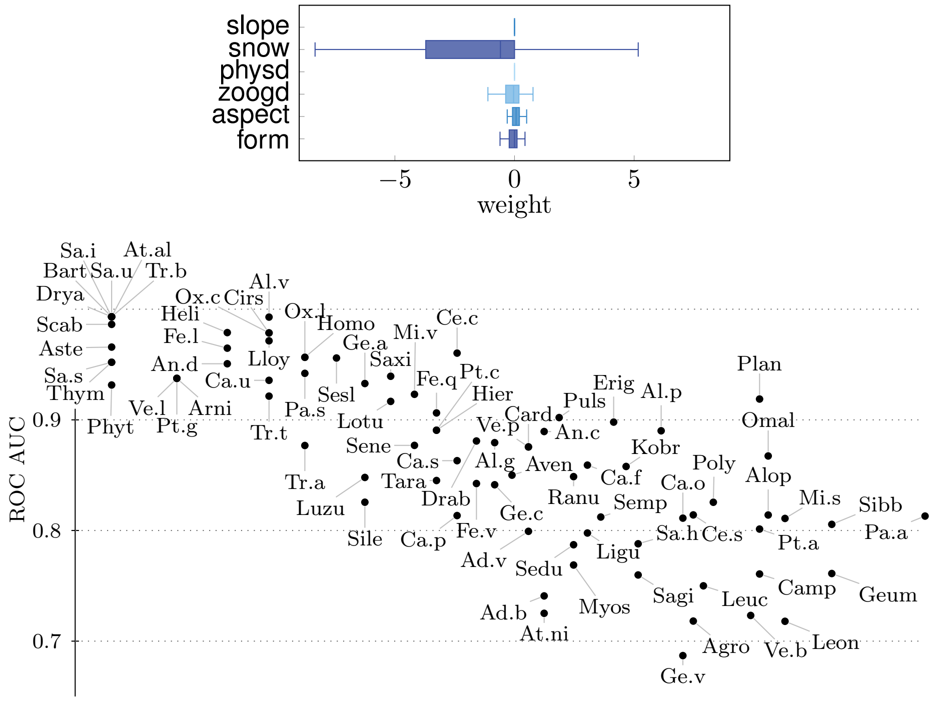


Fig. 5: Habitat Suitability Model variable importance and prediction performances per genus.

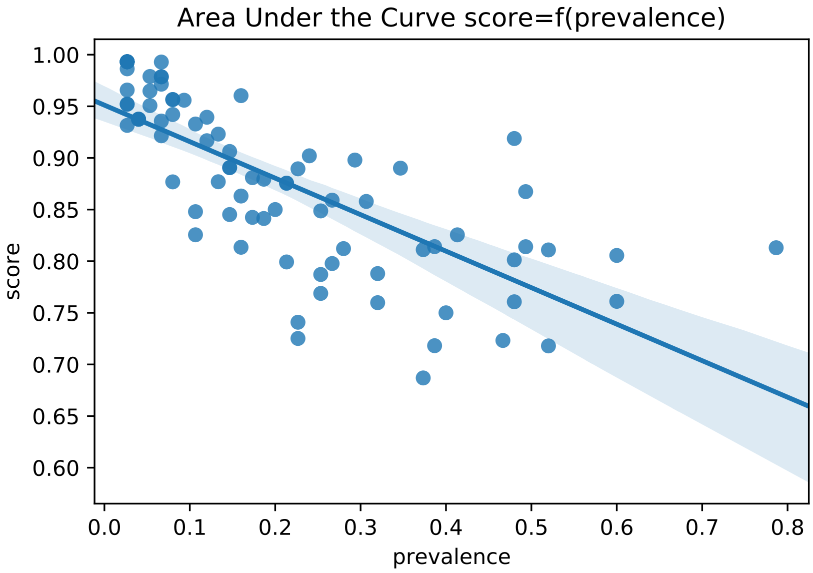


Fig. 6: Area Under the Curve score of the habitat suitability model as a function of the species prevalence.

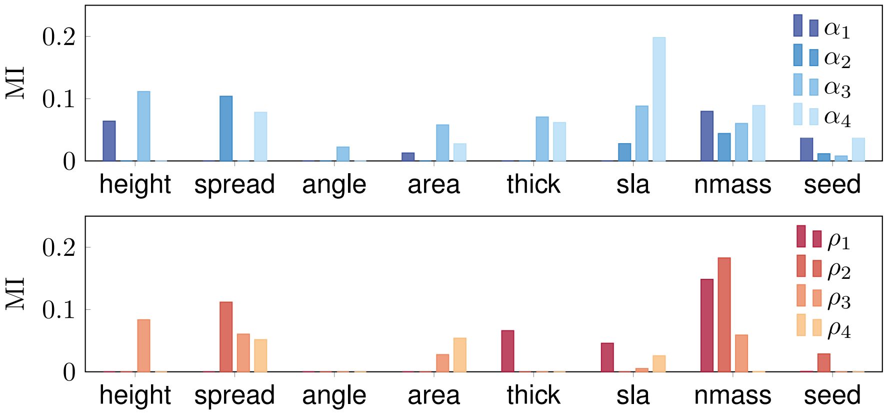


Fig. 7: Mutual information between plant traits and their latent representations. Each bar concerns a specific trait, it represents the stack of mutual information scores from the first to the last (fourth) embedding dimension. The lower (resp. upper) figure shows the results for the response (resp. effect) embeddings.

Sim. ID	**Association types**	**Pool size**	**Density**	**Direction**	**Dimension**	**Deviance of predictions**
1	A	5	/	/	**2**	**0.88±0.88**
2	A	10	/	/	**3**	**0.78±0.54**
3	A	20	/	/	**5**	**0.83±0.4**
4	A+F	5	Sparse	Symetric	**2**	**0.7±0.74**
5	A+F	5	Dense	Symetric	**2**	**0.46±0.59**
6	A+F	10	Sparse	Symetric	**5**	**0.71±0.6**
7	A+F	10	Sparse	Asymetric	**4**	**0.63±0.47**
8	A+F	10	Dense	Symetric	**5**	**0.66±0.4**
9	A+F	10	Dense	Asymetric	**4**	**0.44±0.43**
10	A+F	20	Sparse	Symetric	**7**	**0.79±0.32**
11	A+F	20	Sparse	Asymetric	**8**	**0.75±0.54**
12	A+F	20	Dense	Symetric	**9**	**0.79±0.35**
13	A+F	20	Dense	Asymetric	**8**	**0.82±0.76**
14	A+C	5	Sparse	Symetric	**2**	**1.25±0.88**
15	A+C	5	Dense	Symetric	**2**	**1.25±1.13**
16	A+C	10	Sparse	Symetric	**5**	**1.04±0.67**
17	A+C	10	Sparse	Asymetric	**4**	**0.88±0.61**
18	A+C	10	Dense	Symetric	**5**	**0.96±0.57**
19	A+C	10	Dense	Asymetric	**4**	**1.02±0.79**
20	A+C	20	Sparse	Symetric	**5**	**0.78±0.3**
21	A+C	20	Sparse	Asymetric	**9**	**0.83±0.45**
22	A+C	20	Dense	Symetric	**7**	**0.94±0.55**
23	A+C	20	Dense	Asymetric	**8**	**0.87±0.48**
24	A+F+C	5	Sparse	Symetric	**2**	**0.69±0.85**
25	A+F+C	5	Dense	Symetric	**2**	**0.79±0.86**
26	A+F+C	10	Sparse	Symetric	**5**	**0.67±0.51**
27	A+F+C	10	Dense	Symetric	**5**	**0.9±0.7**
28	A+F+C	20	Sparse	Symetric	**7**	**0.78±0.32**
29	A+F+C	20	Dense	Symetric	**9**	**0.83±0.4**

Table 1: Selected embedding dimension and its corresponding deviance over predicted abundances for each simulation.

$\lambda \setminus k$	2	4	8	16	32
Effective dimension					
0.010	2	4	8	16	32
0.020	2	3	5	11	21
0.030	0	0	1	3	0
0.040	0	0	0	0	0
Deviance					
0.010	0.300	0.295	0.296	0.290	0.287
0.020	0.302	0.298	0.298	0.295	0.295
0.030	0.579	0.579	0.304	0.305	0.579
0.040	0.579	0.579	0.579	0.579	0.579
AIC					
0.010	1148.960	1804.961	3116.962	5740.960	10988.957
0.020	1148.946	1804.952	3116.954	5740.951	10988.947
0.030	1148.704	1804.704	3116.936	5740.920	10988.704
0.040	1148.704	1804.704	3116.704	5740.704	10988.704

Table 2: Effective dimension (number of non-zero components), positive deviance and Akaike Information Criterion (AIC) as a function of the embedding size (k) and the lasso penalty parameter λ.