Supporting Information

S1. Description of the four data sets

Plant communities from Kuebbing et al. (2015) The authors selected 8 phylogenetically-paired plant species (4 natives and 4 non-natives) typical of old-fields in the southeastern United States of America. The goal was to test how species richness affects seedling establishment and productivity below- and above-ground, and how the effects differ between native and non-native assemblages. The authors performed a nearly full-factorial design with richness varying from one to four species, performing 14 out of the 15 possible combinations. Each community had 20 replicates and within each replicate there were 12 individuals. We used the data from their biomass assay, in which they randomly selected 10 replicates from each treatment (native versus non-native) and estimate the biomass of all species. The remaining 10 replicates were used for the seedling establishment experiment (data not used here). The species used in each treatment are reported in Table S1.

Table 1: Species selection for the experiments by Kuebbing et al., 2015.

Family	Native species	Non-native species
Asteraceae (as)	*Achillea millefolium*	*Leucanthemum vulgare*
Fabaceae (fa)	*Lespedeza capitata*	*Lespedeza cuneata*
Lamiaceae (la)	*Pycnanthemum virginianum*	*Prunella vulgaris*
Poaceae (po)	*Sorghastrum nutans*	*Phleum pratense*

Phytoplankton communities from Ghedini et al. (2022) The authors cultured assemblages derived from a pool of five phytoplankton species (*Amphidinium carterae* [A], *Tetraselmis* sp [T]., *Dunaliella tertiolecta* [D], *Tisochrysis lutea* [Ti] and *Synechococcus* sp. [S]). They grew each species in isolation (3 replicates), two species together in all possible pairs (10 combinations, 3 replicates each), and all species together (5 replicates) for 10 days (corresponding to approximately ten generations). Time series for these data are reported in Figure S1. Samples were measured every other day. Here we analyze the data for day 8, because some of the species show a marked decline/increase on day 10.

Bacterial communities from Chu et al. (2021) The authors performed experimental evolution by growing a strain of *Pseudomonas fluorescens* (P) along with all possible combinations of four other strains, *Achromobacter* sp. (A), *Ochrobactrum* sp. (O), *Stenotrophomonas* sp. (S), and *Variovorax* sp. (V). The bacteria were grown in six replicates for each community in liquid media. Importantly, these strains form distinct morphological colonies, allowing the authors to estimate the density of each strain by plating an appropriately diluted sample and counting the number of Colony-Forming Units. The density of each species was determined at each transfer to fresh medium (transfers were performed every 7 days, Figure S2). Because *Variovax* sp. went extinct in the majority of communities and replicates, here we analyze the sub-communities measured for experiments that did not include *Variovax* sp.

S2. Rank conditions for fitting all parameters

In order to fit the n^2 parameters of the interaction matrix B in our model, we need to have observed a sufficiently large variety of communities. For simplicity, suppose that we are simulating data from the model as specified by a given matrix B, which details the interactions between the n species in the pool. For each possible set of species k (i.e., all single species, all possible pairs, etc.), we compute the solution $x^{(k)} = \left(B^{(k)}\right)^{-1} 1^{(k)}$ and, if all components of $x^{(k)}$ are positive, we record them in a row of matrix X, along with zeros for all the species that are in the pool but not present in community k. We ask when we can recover all the coefficients in B from X.

Under these conditions, to recover the coefficients of the matrix B we need to be able to find a unique solution for $BX^T = P^T$. We can write this as:

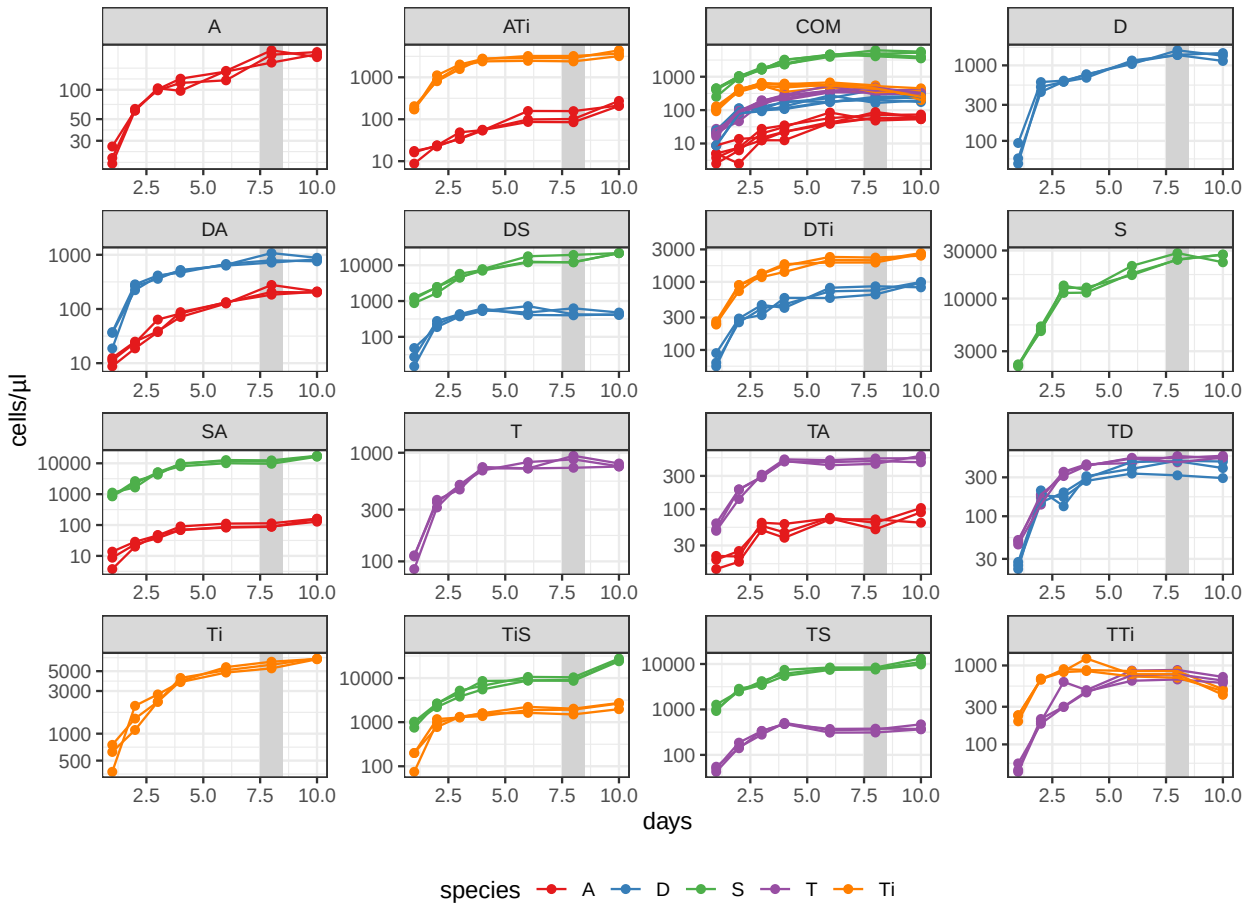

Figure 1: Time series for the growth of the 16 communities considered by Ghedini et al., 2022. For the analysis presented here, we considered the densities measured on day 8 (highlighted in the panels).

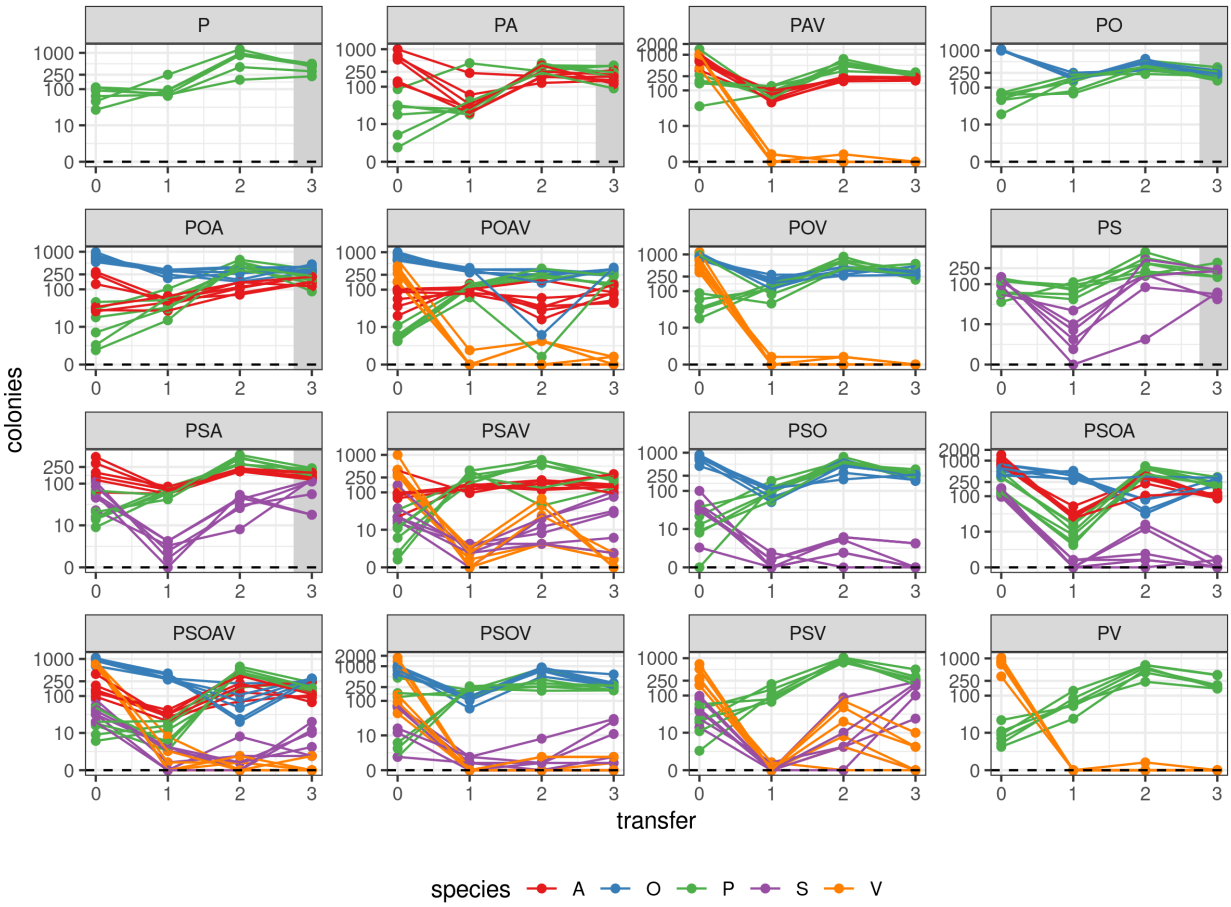

Figure 2: Time series for the growth of the 16 communities considered by Chu et al., 2021. For the analysis presented here, we considered the densities measured at the end of the experiment, for communities that did not include *Variovax* sp., because the strain went extinct in the majority of experiments. We also excluded communities where some of the replicates resulted in the extinctions of some of the species (PSO, PSOA). The points used for the analysis are highlighted in the panels.

$$IBX^T = P^T$$
$$(X \otimes I)\text{vec}(B) = \text{vec}(P^T) \quad ,$$

where vec(A) is the vectorization operator (stacking all the columns in a matrix into a vector), and $\otimes$ is the Kronecker product. Note that (as explained in details in Supporting Information, S4), we cannot determine all elements of P without having access to matrix B. We can however be certain that whenever $X_{ij} > 0$, then $P_{ij} = 1$. If we then call $\tilde{P}$ a matrix that has $\tilde{P}_{ij} = 1$ whenever $X_{ij} > 0$ and $\tilde{P}_{ij} = 0$ elsewhere. We can rewrite the system of equations above as:

$$\text{diag}(\text{vec}(\tilde{P}))(X \otimes I)\text{vec}(B) = \text{vec}(\tilde{P}^T) \quad , \tag{1}$$

where diag(b) creates a diagonal matrix with vector b on the diagonal. It is now apparent that solving the matrix equation for the n^2 entries of B requires the matrix $\text{diag}(\text{vec}(\tilde{P}))(X \otimes I)$ to have rank n^2. This condition will be met only when each of the n species in a system are present in at least n experiments, and each pair of species appears in at least one experiment. Note that if any two species were to always co-occur alongside each other in the experimental data, we would not be able to fit the full B matrix—in this case, we would be unable to distinguish the effects of the two species from one another.

The procedure is the same for observed data; all we need to do is to substitute matrix X in Eq. 1 with the matrix of observed densities obtained by retaining a single row for each community (i.e., removing replicate experiments). These conditions are equivalent to those found in Maynard et al. (2020).

S3. Naïve approach

Here we rewrite the naïve approach of Maynard et al. (2020) in matrix form, thereby showing that it does not minimize the SSQ and therefore does not yield the m.l.e. for matrix B under the simple error model described in the main text.

Taking the equation Eq. 4, and using the matrix $\tilde{P}$ as above, we find:

$$\tilde{P}^T \circ (B\tilde{X}^T) = \tilde{P}^T \circ P^T + \tilde{P}^T \circ S^T = \tilde{P}^T + \tilde{P}^T \circ S^T \quad .$$

where $\circ$ is the element-by-element (Hadamard) product.

If we proceed as above and attempt to solve the system of equations

$$\mathrm{diag}(\mathrm{vec}(\tilde{P}))(\tilde{X} \otimes I)\mathrm{vec}(B) = \mathrm{vec}(\tilde{P}^T) \quad ,$$

for the matrix B by taking the Moore-Penrose pseudoinverse,

$$\mathrm{vec}(\hat{B}) = (\mathrm{diag}(\mathrm{vec}(\tilde{P}))(\tilde{X} \otimes I))^{+}\mathrm{vec}(\tilde{P}^T) \quad ,$$

where $A^{+} = (A^T A)^{-1}A^T$, what we would be minimizing is $\|\tilde{P}^T \circ S^T\| = \|\tilde{P}^T \circ B\mathcal{E}^T\| \neq \|\mathcal{E}\|$. As such, this method will not be the m.l.e. for B for any case in which the measurements are noisy, even if the process generating the data would follow the model precisely. This method instead minimizes the deviation from equilibrium conditions in the corresponding Generalized Lotka-Volterra model (see S4).

S4. Connections to the Generalized Lotka-Volterra model

The Generalized Lotka-Volterra (GLV) model is arguably the simplest nonlinear model for population dynamics, and has been studied for more than a century (Lotka, 1920; Volterra, 1926). It can be written as:

$$\dot{x}(t) = D(x(t))(r - Ax(t)) \quad , \tag{2}$$

where $\dot{x}(t) = dx(t)/dt$, $x(t)$ is a column vector detailing the densities of all species at time t, r is a vector of intrinsic growth/death rates, A is a matrix of species' interactions, and $D(x(t))$ is a diagonal matrix with $x(t)$ on the diagonal. We can divide each element of A by the corresponding growth rate, obtaining:

$$\dot{x}(t) = D(r \circ x(t))(1 - D(r)^{-1}Ax(t)) = D(r \circ x(t))(1 - Bx(t)) \quad . \tag{3}$$

These equations hold whenever all species are present, or when we initialize the system using a subset of species k. The model admits up to 2^n equilibria, of which only one can have all positive components (feasible coexistence equilibrium), and the others are "boundary" equilibria, in which a certain subset of species k can coexist, and the remaining species are absent. For each subset of species k, a feasible equilibrium—when it exists—is found solving $1^{(k)} - B^{(k)}x^{(k)} = 0^{(k)}$, i.e., $x^{(k)} = \left(B^{(k)}\right)^{-1} 1^{(k)}$, which is exactly our Eq. 3. Therefore, as already noted by Maynard et al. (2020), fitting the statistical models presented here is equivalent to finding a GLV model with an equilibrium structure that is as close as possible to the observed data. Note that this also makes clear that, using "static" observations of several sub-communities, we cannot find A and r separately, only the composite parameters $B = D(r)^{-1}A$.

In this context, the predicted matrix X (for a given matrix B) is simply a collection of feasible equilibria for some of the sub-systems we can form from the pool of n species. Take one of the rows of X, $\bar{x}$, containing the density of the $|k|$ species forming the feasible equilibrium, along with zeros for the species that are absent. Because of the equilibrium condition, for all $\bar{x}_i > 0$, we have

$$\sum_j B_{ij}\bar{x}_j = \sum_{j \in k} B_{ij}\bar{x}_j = 1 \quad ,$$

for the remaining species (i.e., the species x_i not present in k), we have

$$\sum_j B_{ij}\bar{x}_j = \sum_{j \in k} B_{ij}\bar{x}_j = \rho_i \quad .$$

The quantity ρ_i can be interpreted as the effect of the "resident" species (i.e., those for which $\bar{x}_i > 0$) on the growth rate of the invader i when entering the community at low abundance. Write the per-capita growth rates:

$$D(x(t))^{-1}\dot{x}(t) = \dot{\log} x(t) = r + Ax(t) = D(r)(1 - Bx(t))$$

If the system is resting at $\bar{x}$, we have that whenever $(D(r)(1 - B\bar{x}))_i > 0$, species i has positive growth rate when invading the system at low abundance. If we assume that the growth rates are positive for all species, we have that the inequality for species i is satisfied whenever $1 - \rho_i > 0$, and therefore $\rho_i < 1$. If the sign of r_i is negative, the inequality is reversed.

Therefore, when we compute $P = XB^T$, we have that $P_{ij} = 1$ whenever the species j is part of the community specified by row i of X, and that (assuming $r_i > 0$) whenever $P_{ij} < 1$ for a species that is not part of the community, then the species can invade the community when rare, and cannot invade whenever $P_{ij} > 1$. If moreover we have that the species in k are resting at a stable equilibrium, then any row of P for which $P_{ij} \geq 1$ for all species represents a feasible, stable and non-invasible equilibrium for the system, also called a "saturated rest point" (Hofbauer et al., 1998; Serván et al., 2018). Serván & Allesina (2021) showed that these equilibria are also final configurations for the assembly paths one can form from the pool of n species.

S5. Derivation of the simplified models from a consumer-resource framework

Having established a parallel between our statistical model and the equilibrium structure of a Generalized Lotka-Volterra model, we use this equivalence as a jumping board to derive simplified versions of our model. In particular, we consider a simple MacArthur's consumer-resource model (MacArthur, 1970), in which we have n consumers and m resources (assuming $m \geq n$, a necessary condition for the coexistence of the n consumers):

$$\begin{cases} \dot{x} = D(x)(Ay - \delta) \\ \dot{y} = D(y)(\rho - y - A^T x) \end{cases}$$

where $\dot{x} = dx(t)/dt$ and $\dot{y} = dy(t)/dt$ are vectors describing the change in the density of the n consumers (stored in x) and the m resources, (y), respectively. The vector δ stores the death (growth) rate of the consumers, ρ that of the resources, and the $n \times m$ matrix A the attack rates of consumers on resources; $D(x)$ defines a diagonal matrix with x on the diagonal. If we assume that the dynamics of the resources are fast compared to those of the consumers, we can solve for the density of the resources attained for a given state of the consumers:

$$y = \rho - A^T x \quad ,$$

and substitute this value in the equations for the consumers, obtaining:

$$\dot{x} = D(x)(Ay - \delta) = D(x)(A\rho - AA^T x - \delta) = D(x)(s - Mx) = D(s \circ x)(1 - Bx) \quad ,$$

where we have defined $s = A\rho - \delta$, $M = AA^T$ and $B = D(s)^{-1}M$. This is a Generalized Lotka-Volterra model, and the term in parenthesis is exactly what we have considered for our statistical model.

In principle, the matrix A can detail the interactions with an arbitrary number of resources $m \geq n$. To simplify the statistical model and reduce the number of free parameters, we assume a particular structure for A. Suppose that each consumer has access to private resources, as well as a shared resources. In the simplest case, with one private resource for each consumer and one resource shared by all consumers, the matrix A has the following structure:

$$A = \begin{pmatrix} \alpha_1 & 0 & 0 & \dots & 0 & \beta_1 \\ 0 & \alpha_2 & 0 & \dots & 0 & \beta_2 \\ 0 & 0 & \alpha_3 & \dots & 0 & \beta_3 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \alpha_n & \beta_n \end{pmatrix} \quad ,$$

where the vector α collects the attack rates for the consumers on their private resources and the vector β the attack rates for each consumer and the shared resource. Then we have that $M = AA^T$ is simply:

$$M = AA^T = \begin{pmatrix} \alpha_1^2 + \beta_1^2 & \beta_1\beta_2 & \beta_1\beta_3 & \dots & \beta_1\beta_n \\ \beta_2\beta_1 & \alpha_2^2 + \beta_2^2 & \beta_2\beta_3 & \dots & \beta_2\beta_n \\ \beta_3\beta_1 & \beta_3\beta_2 & \alpha_3^2 + \beta_3^2 & \dots & \beta_3\beta_n \\ \dots & \dots & \dots & \dots & \dots \\ \beta_n\beta_1 & \beta_n\beta_2 & \beta_n\beta_3 & \dots & \alpha_n^2 + \beta_n^2 \end{pmatrix} = D(\alpha^2) + \beta\beta^T$$

Dividing each row of M by the corresponding s_i, we obtain:

$$B = D(s)^{-1}M = D(\alpha^2/s) + \left(\frac{\beta}{s}\right)\beta^T = D(d) + vw^T \quad ,$$

where $d = \alpha^2/s$, $v = \beta/s$ and $w = \beta$. As such, in this case we can fully define the matrix B using $3n - 1$ parameters (with the -1 due to the fact that elements $v_i w_j$ always appear as products). We can complicate this model by adding extra shared resources, each time adding further parameters. In particular, if we wanted to model off-diagonal (i.e., inter-specific interactions) as the sum of two rank-1 matrices, we could write:

$$B = D(s)^{-1}M = D(\alpha^2/s) + \left(\frac{\beta}{s}\right)\beta^T = D(d) + vw^T + v'w'^T \quad ,$$

adding only $2n - 3$ parameters; in fact, without loss of generality we can assume that $v^T v' = 0$ and $w^T w' = 0$, i.e., the two left vectors are orthogonal and the two right-vectors are also orthogonal. The orthogonality condition allows us to solve for one of the parameters in each vector. Adding another rank-1 matrix, say $v''w''^T$, adds a further $2n - 5$ parameters, and so on.

Can the model be simplified further? If we assume that all the s_i are the same, then the matrix can be rewritten as $B = D(d) + vv^T$, and is therefore symmetric ($2n$ parameters); if further we assume that all consumers have the same attack rate on the resources, the matrix can be written as $B = D(d) + \alpha 11^T$ ($n + 1$ parameters).

We have shown how the consumer-resource framework can be used to derive simplified versions of the statistical model, thereby lifting some of the requirements on the data (see section S7 below). While in a consumer-resource model we would expect all attack rates to be positive, we can relax this condition to have a more general and flexible model.

S6. Fitting the simplified models

The main computational hurdle one faces when fitting the original model with n^2 parameters is the need to invert a sub-matrix of B, $B^{(k)}$ for every species combination k, in order to compute the predicted values. Inverting matrices is computationally costly, requiring on the order of approximately $n^{2.4}$ operations for an $n \times n$ matrix, even using the best available algorithms.

This problem is greatly reduced for the simplified models, as we can easily obtain a closed-form analytical expression for the inverse of a sub-matrix by applying the Sherman-Morrison formula(Sherman & Morrison, 1950). In particular, if we have a matrix:

$$B = D(d) + vw^T \quad ,$$

for arbitrary vectors d, v and w, we can rewrite it without loss of generality as:

$$B = D(\delta)^{-1} + (D(\delta)^{-1}\nu)(D(\delta)^{-1}\omega)^T \quad ,$$

by defining $\delta = 1/d$, $\nu = v/d$ and $\omega = w/d$ and assuming element-by-element (Hadamard) division. With this notation, the inverse of B becomes:

$$B^{-1} = D(\delta) - \frac{\nu\omega^T}{1 + \sum_j \nu_j\omega_j/\delta_j} \quad .$$

The predicted abundance for the species in the community can be computed as:

$$x = B^{-1}1 = \delta - \nu\frac{(\omega^T 1)}{1 + (\nu \circ \omega)^T(1/\delta)} \quad ,$$

Even more conveniently, the predicted abundance for a sub-community be readily computed. Simply define a vector of presence/absence π_k such that $\pi_{ki} = 1$ if species i is part of the sub-community and 0 otherwise. Then, we can write $x^{(k)}$, containing the predicted density of the species when present, and 0 otherwise, as:

$$x^{(k)} = \pi_k \circ \delta - \pi_k \circ \nu\frac{((\pi_k \circ \omega)^T 1)}{1 + (\pi_k \circ \nu \circ \omega)^T(1/\delta)} \quad .$$

Similar (and even simpler) calculations can be performed for the further simplifications of the model. Having access to a linear-time, analytical way to compute the predicted densities speeds up the calculation enormously, allowing our methods to be applied to larger sets of experiments.

S7. Data requirements to fit the simplified model

As detailed in Appendix S2, to fit the full model we need to identify the n^2 coefficients of the matrix B. This can be accomplished by writing a set of equations based on the observed data. In particular, having observed a certain community k in which m species coexist, we can write m equations. For example, having observed species one growing in monoculture (labeled community 1), we can write:

$$B_{11}\tilde{x}_1^{(1)} = 1 \quad \Longrightarrow \quad B_{11} = \frac{1}{\tilde{x}_1^{(1)}}$$

Similarly, when we have observed species one and two growing together (labeled community 2), we can write two equations:

$$\begin{cases} B_{11}\tilde{x}_1^{(2)} + B_{12}\tilde{x}_2^{(2)} = 1 \\ B_{21}\tilde{x}_1^{(2)} + B_{22}\tilde{x}_2^{(2)} = 1 \end{cases} \quad \Longrightarrow \quad \begin{cases} B_{12} = \frac{1-B_{11}\tilde{x}_1^{(2)}}{\tilde{x}_2^{(2)}} \\ B_{21} = \frac{1-B_{22}\tilde{x}_2^{(2)}}{\tilde{x}_1^{(2)}} \end{cases}$$

As such, having observed all monocultures we can write n equations, and having observed all the $\binom{n}{2}$ pairs we can write $n(n-1)$ equations—for a total of n^2 equations, allowing us to solve for all the coefficients of B. Naturally, the same can be accomplished when we have observed all the species growing together (yielding n equations), as well as all the leave-one-out communities (n communities, yielding $n-1$ equations each), or any other combination of experimental observations that allows us to write n^2 linearly independent equations.

Take the simplest of the reduced models presented in section S5 and S6:

$$B = D(d) + \alpha 11^T \quad .$$

Following the same logic, it is obvious that having observed all the monocultures (yielding $B_{ii} = d_i + \alpha = 1/\tilde{x_i}^{(i)}$) and any of the pairs (allowing us to solve for α) would be sufficient to parameterize the model.

The more interesting and complicated case is that of the most complex of the reduced models: $B = D(d) + vw^T$. Here we show that at a minimum we need to be able to write $3n-1$ independent equations, in order to solve for the n diagonal coefficients of B and $2n-1$ off-diagonal coefficients of B. Moreover, we demonstrate that not all combinations of off-diagonal coefficients identify the parameters uniquely, and provide a simple test to determine the identifiability of the parameters.

Because the method to identify the parameters is slightly more complex and abstract than in the case of the full model, we illustrate it with an example. Take the four-species community defined by the parameters:

$$\begin{cases} d = (1, 2, 3, 4)^T \\ v = (1/8, 1/4, 1/2, 1)^T \\ w = (1/32, 1, 7/2, 1)^T \end{cases}$$

Yielding the matrix B:

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} + \begin{pmatrix} \frac{1}{256} & \frac{1}{8} & \frac{7}{16} & \frac{1}{8} \\ \frac{1}{128} & \frac{1}{4} & \frac{7}{8} & \frac{1}{4} \\ \frac{1}{64} & \frac{1}{2} & \frac{7}{4} & \frac{1}{2} \\ \frac{1}{32} & 1 & \frac{7}{2} & 1 \end{pmatrix} = \begin{pmatrix} \frac{257}{256} & \frac{1}{8} & \frac{7}{16} & \frac{1}{8} \\ \frac{1}{128} & \frac{9}{4} & \frac{7}{8} & \frac{1}{4} \\ \frac{1}{64} & \frac{1}{2} & \frac{19}{4} & \frac{1}{2} \\ \frac{1}{32} & 1 & \frac{7}{2} & 5 \end{pmatrix}$$

This system allows for the coexistence for all combinations of species. For simplicity, in the examples below, we will only use the "observed" data for monocultures and pairs:

$$\tilde{X} = \begin{pmatrix} \frac{256}{257} & 0 & 0 & 0 \\ 0 & \frac{4}{9} & 0 & 0 \\ 0 & 0 & \frac{4}{19} & 0 \\ 0 & 0 & 0 & \frac{1}{5} \\ \frac{16}{17} & \frac{15}{34} & 0 & 0 \\ \frac{48}{53} & 0 & \frac{11}{53} & 0 \\ \frac{104}{107} & 0 & 0 & \frac{83}{428} \\ 0 & \frac{31}{82} & \frac{7}{41} & 0 \\ 0 & \frac{19}{44} & 0 & \frac{5}{44} \\ 0 & 0 & \frac{9}{44} & \frac{5}{88} \end{pmatrix}$$

Clearly, if we had access to all the rows of $\tilde{X}$, we would be able to infer an arbitrary B (as in the full model), and therefore also any B with the special form $B = D(d) + vw^T$. Our goal is to show that we do not need to have observed all the monocultures and pairs to infer the parameters of this simplified form.

Before we start, we note that $G = vw^T$ is defined uniquely by $2n - 1$ (rather than $2n$) parameters. In fact, we can take any $\theta \neq 0$ and define $v' = \theta v$ and $w' = \frac{1}{\theta} w$ such that $G = vw^T = v'w'^T$.

Now consider the case in which we have observed all the monocultures, all the pairs involving species one, and the pair involving species two and three. I.e., we have observed all the monocultures, but only four pairs of the six potential. We can use this data to solve for the coefficients B_{11}, B_{12}, B_{13}, and B_{14}:

$$\begin{pmatrix} \frac{256}{257} & 0 & 0 & 0 \\ \frac{16}{17} & \frac{15}{34} & 0 & 0 \\ \frac{48}{53} & 0 & \frac{11}{53} & 0 \\ \frac{104}{107} & 0 & 0 & \frac{83}{428} \end{pmatrix} \begin{pmatrix} B_{11} \\ B_{12} \\ B_{13} \\ B_{14} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} B_{11} \\ B_{12} \\ B_{13} \\ B_{14} \end{pmatrix} = \begin{pmatrix} \frac{257}{256} \\ \frac{1}{8} \\ \frac{7}{16} \\ \frac{1}{8} \end{pmatrix}$$

Using the appropriate observations, we can also identify B_{22}, B_{21}, B_{23}, B_{31}, B_{33}, B_{41} and B_{44}. In summary, we have identified $3n - 1$ coefficients of B: the n elements on the diagonal, as well as $2n - 1$ elements in the off-diagonal part of the matrix.

Note that $B_{ii} = d_i + G_{ii}$, while $B_{ij} = G_{ij}$ when $i \neq j$. In the equation above, we have therefore observed a partial matrix G:

$$G^* = \begin{pmatrix} G_{11} & \frac{1}{8} & \frac{7}{16} & \frac{1}{8} \\ \frac{1}{128} & G_{22} & \frac{7}{8} & G_{24} \\ \frac{1}{64} & G_{32} & G_{33} & G_{34} \\ \frac{1}{32} & G_{42} & G_{43} & G_{44} \end{pmatrix} \tag{4}$$

Is there a unique way to "complete" the matrix G, knowing that it must be rank-1? This problem is known in mathematics as the "matrix completion" problem, and has been studied extensively. In particular, a partially-specified matrix admits at least one rank-1 completion whenever it meets two conditions (Hadwin et al., 2006). The first condition, called the "zero row or column property" is very simple: if one of the specified coefficients is zero, then all the specified coefficients in either the corresponding row or column must also be zero. In our case, we have no zero coefficients, and thus the property is trivially satisfied.

The second condition is based on the bipartite graph $\mathcal{G}$ that we can construct by taking the labels for the rows and columns of G^*, r_i and c_i respectively, as nodes and drawing an undirected edge between r_i and c_j whenever G^*_{ij} is specified. The second condition, called the "cycle property", has to do with the consistency of the undirected cycles in the bipartite graph $\mathcal{G}$. For the partially-specified matrix in Eq. 4, the graph has no undirected cycles (Fig. S3), and therefore the property is trivially satisfied. Finally, if the graph $\mathcal{G}$ is connected (i.e., there is a path connecting any node to any other), the rank-1 completion is unique. Thus, by inspecting the graph corresponding to matrix G^* in Eq. 4, we conclude that a matrix completion of G^* exists and is unique. Having established that a completion exists and is unique, we then ask how we can find

the coefficients needed to complete the matrix.

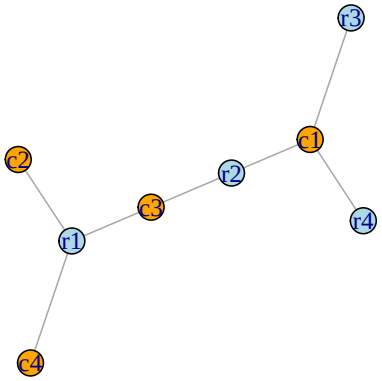

Figure 3: Bipartite graph associated with the partially-specified matrix in Eq. 8. The graph represents a connected tree, and as such a completion of matrix G^* exists and is unique.

For this final step, one can devise several alternative algorithms. One of the simplest approaches, from the conceptual point of view, is based on the fact that any $k \times k$ submatrix of a rank-1 matrix, obtained selecting a certain set of rows and corresponding columns, must also be rank-1. As such, the determinant of any submatrix of G^* with at least two rows and columns must be zero. We can therefore form equations to solve for the remaining coefficients. For example, considering the submatrices induced by the rows and columns $\{1,2\}$, $\{1,3\}$, $\{2,3\}$, and $\{1,2,3\}$, we can write:

$$\begin{cases} \det(G^{*(1,2)}) = G_{11}G_{22} - \frac{1}{1024} = 0 \\ \det(G^{*(1,3)}) = G_{11}G_{33} - \frac{7}{1024} = 0 \\ \det(G^{*(2,3)}) = G_{22}G_{33} - \frac{7\,G_{32}}{8} = 0 \\ \det(G^{*(1,2,3)}) = G_{11}G_{22}G_{33} - \frac{7}{4096} - \frac{7\,G_{22}}{4096} + \frac{7\,G_{32}}{2048} - \frac{7\,G_{11}G_{32}}{8} - \frac{G_{33}}{1024} = 0 \end{cases}$$

Yielding $G_{11} = \frac{1}{256}$, $G_{22} = \frac{1}{4}$, $G_{33} = \frac{7}{4}$ and $G_{32} = \frac{1}{2}$. Having identified these coefficients, we write equations for $\det(G^{*(1,4)})$ to find $G_{44} = 1$, and then the equations for $\det(G^{*(2,4)})$ and $\det(G^{*(1,2,4)})$ to solve for $G_{24} = \frac{1}{4}$ and $G_{42} = 1$. Finally, to complete the matrix G^*, we solve the equations for $\det(G^{*(3,4)})$ and $\det(G^{*(1,3,4)})$, yielding $G_{34} = \frac{1}{2}$ and $G_{43} = \frac{7}{2}$.

In summary, provided with $2n-1$ nonzero coefficients of G satisfying the connectedness of the corresponding graph $\mathcal{G}$, we can complete the matrix and the completion will be unique. Note that we are specifying only $2n-1$ coefficients, which implies that $\mathcal{G}$ has only $2n-1$ edges. As such, if $\mathcal{G}$ is connected, it is necessarily a tree. A tree automatically satisfies the cycle condition in (Hadwin et al., 2006), and if the coefficients are nonzero, connectedness is necessary and sufficient for the existence of a unique solution. Because the solution is unique, numerical methods could easily substitute the algebra we performed above.

Finally, having identified all of the coefficients G_{ii}, and provided with the estimates of B_{ii} we obtain d_i by difference.

To demonstrate that not all choices of $2n-1$ off-diagonal coefficients would satisfy the conditions for existence and uniqueness, consider the same system and use the monocultures and the experiments in which we grow species (1, 2), (2,3), (3,4) and (4,1) together. In total, we should be able to identify $3n$ coefficients of B, including all the diagonal coefficients and $2n$ off-diagonal coefficients. Excluding one of them arbitrarily (this has no effect on the result), we end up with the partially-specified matrix:

$$G^* = \begin{pmatrix} G_{11} & \frac{1}{8} & G_{13} & \frac{1}{8} \\ \frac{1}{128} & G_{22} & \frac{7}{8} & G_{24} \\ G_{31} & \frac{1}{2} & G_{33} & \frac{1}{2} \\ \frac{1}{32} & G_{42} & G_{43} & G_{44} \end{pmatrix} \tag{5}$$

In Figure S4 we show that the induced graph $\mathcal{G}$ is not connected (and in fact, not a tree). As such, if the matrix were to satisfy the cycle condition (it does), we would have multiple solutions (in this case, infinitely many), while if it did not, we would have no possible completion.

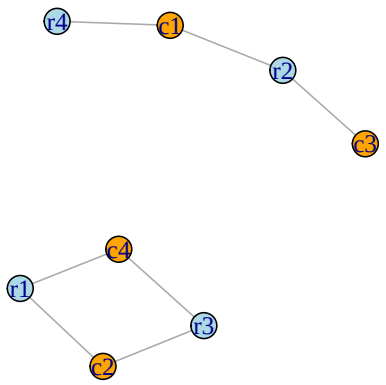

Figure 4: Bipartite graph associated with the partially-specified matrix in Eq. 9. The graph does not represent a connected tree, and as such a completion of matrix either does not exist, or is not unique.

In summary, to identify the parameters of the simplified model we need to have sufficient data to determine both the n diagonal coefficients of B as well as $2n-1$ off-diagonal coefficients, ensuring that the induced graph $\mathcal{G}$ associated with the matrix G^* is connected. This condition is both necessary and sufficient to guarantee the existence and the uniqueness of a solution. Finally, because we need to identify n^2 parameters (quadratic in n) for the full model, and only $3n-1$ (linear in n) for the simplified model, the benefits of the simplified model will be markedly greater for larger n.

S8. Full results

Iterative Algorithm. The iterative method proposed in the main text improves over the naïve approach reported in section S2 in all cases (Figures S3, S4, and S5).

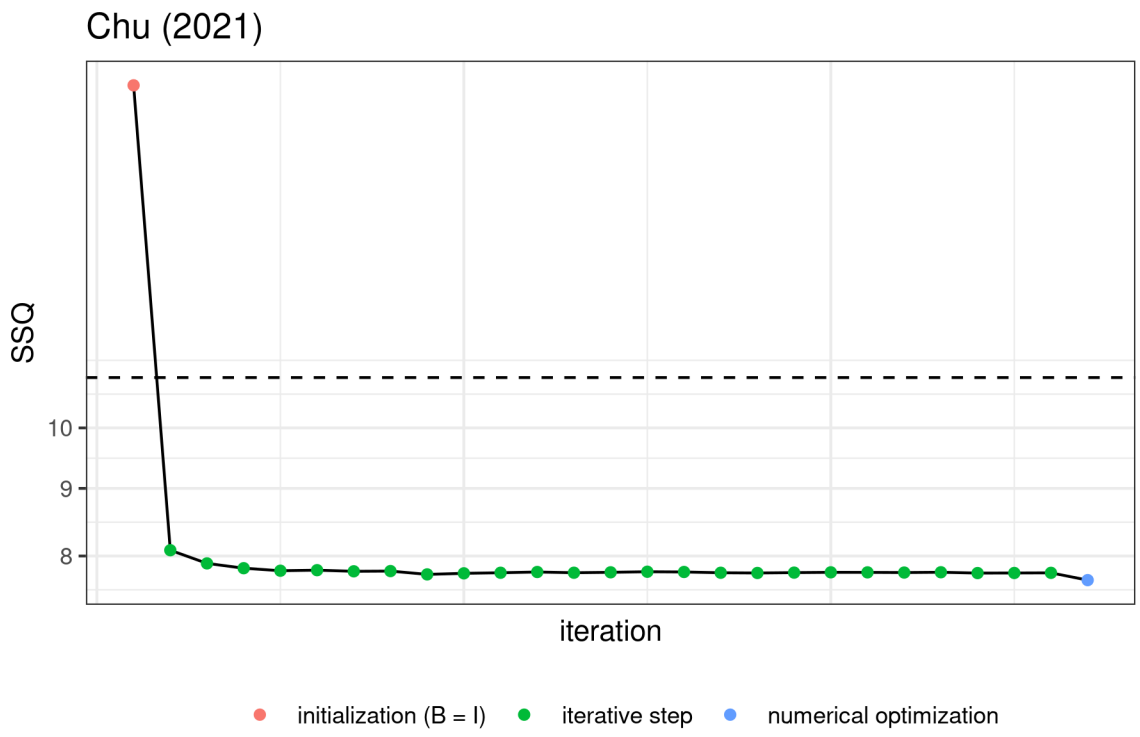

Figure 5: As Figure 1, but for the data of Chu et al., 2021.

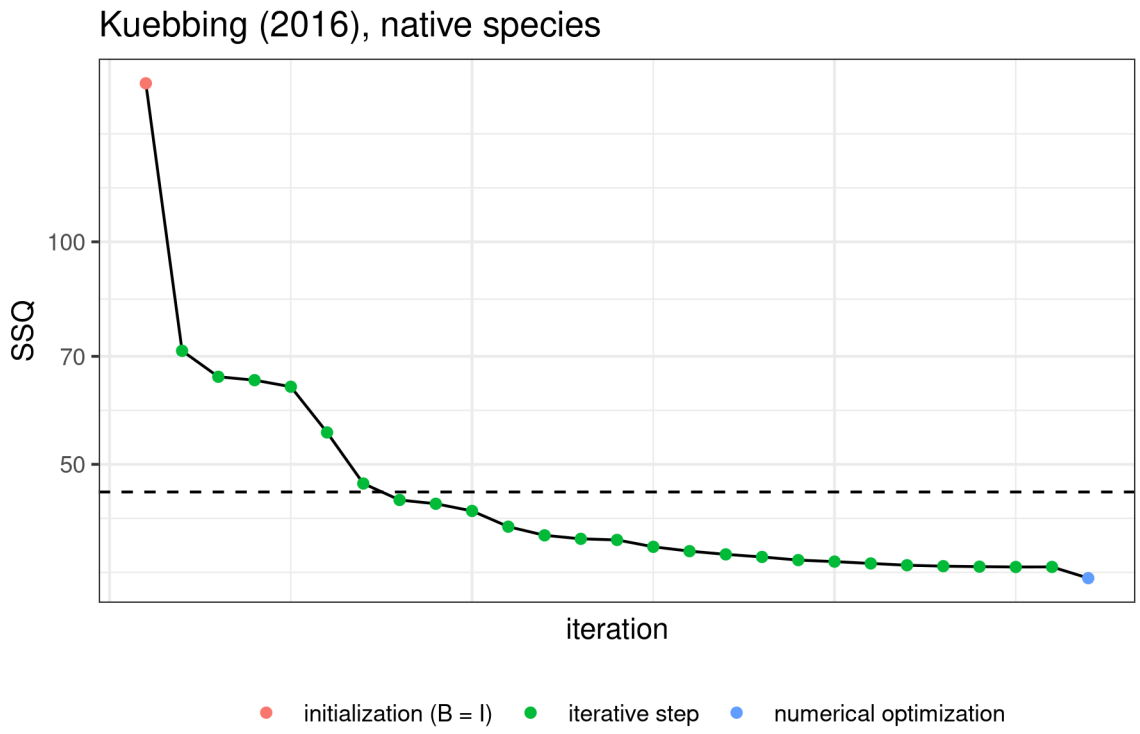

Figure 6: As Figure 1, but for the data of Kuebbing et al. (2015), native plants).

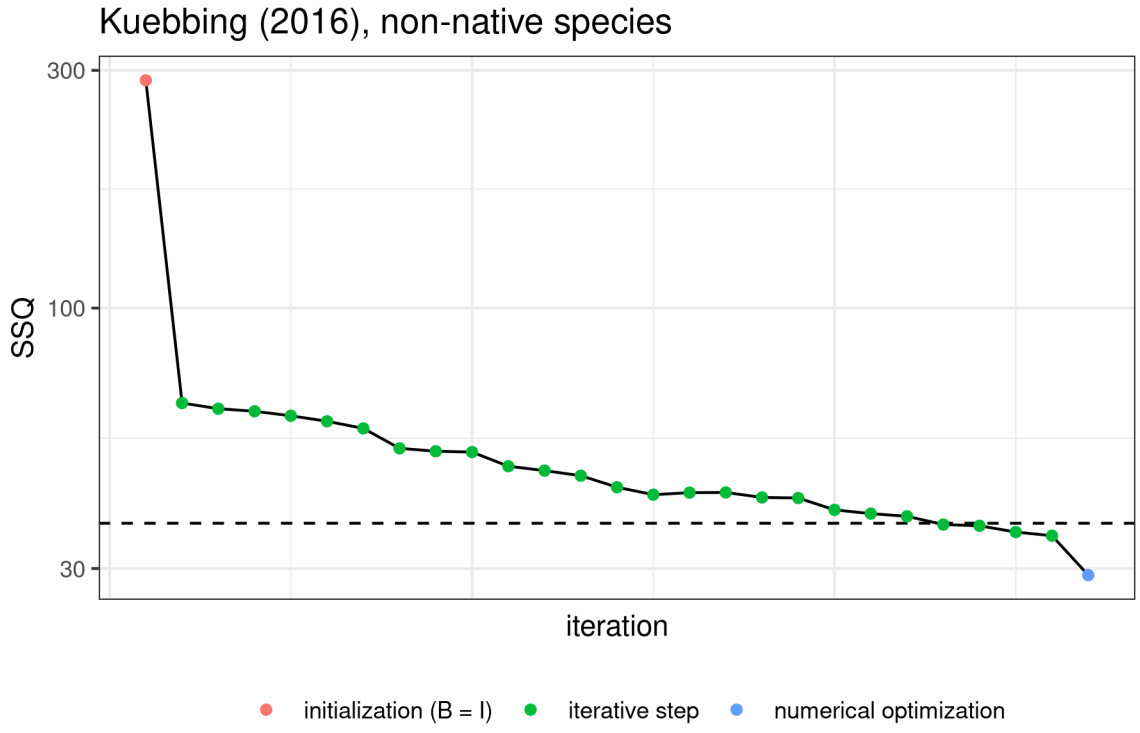

Figure 7: As Figure 1, but for the data of Kuebbing et al. (2015), non-native plants.

Model comparison. For all data sets, we find that the simplified model with $3n - 1$ parameters, in which the interaction matrix is constrained to have form $B = D(d) + vw^T$, performs similarly to the full model with n^2 parameters (Figures S6, S7, and S8).

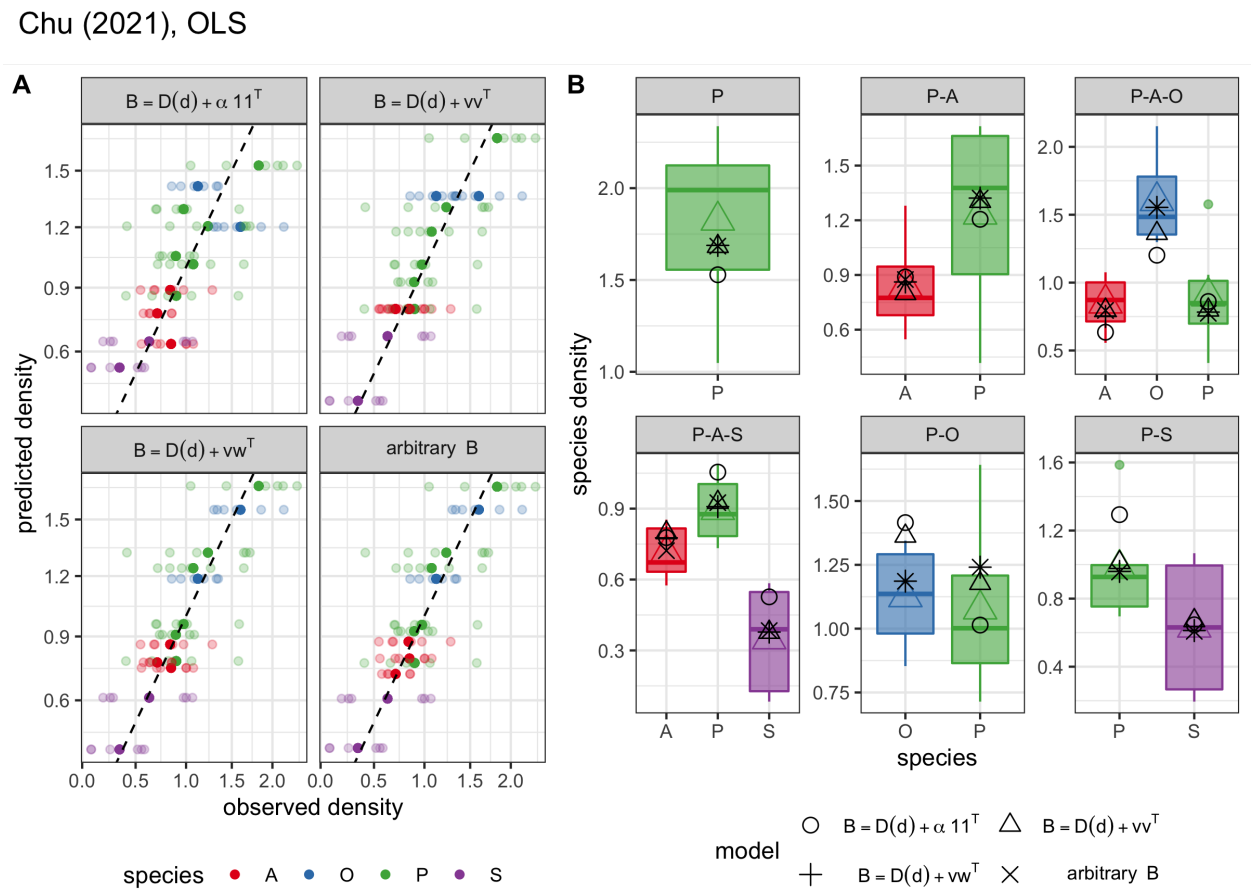

Figure 8: As Figure 2, but for the data of Chu et al., 2021.

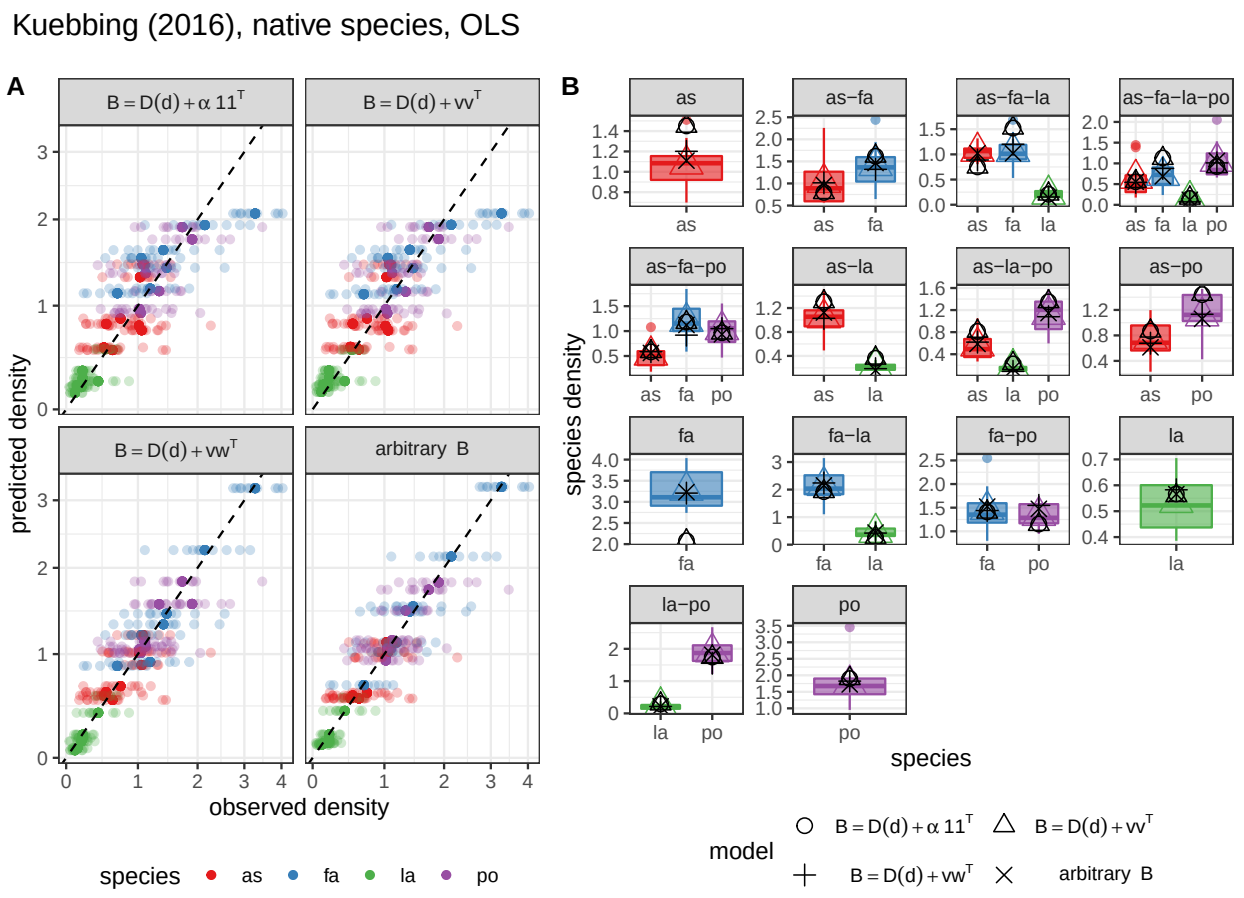

Figure 9: As Figure 2, but for the data of Kuebbing et al. (2015), native plants.

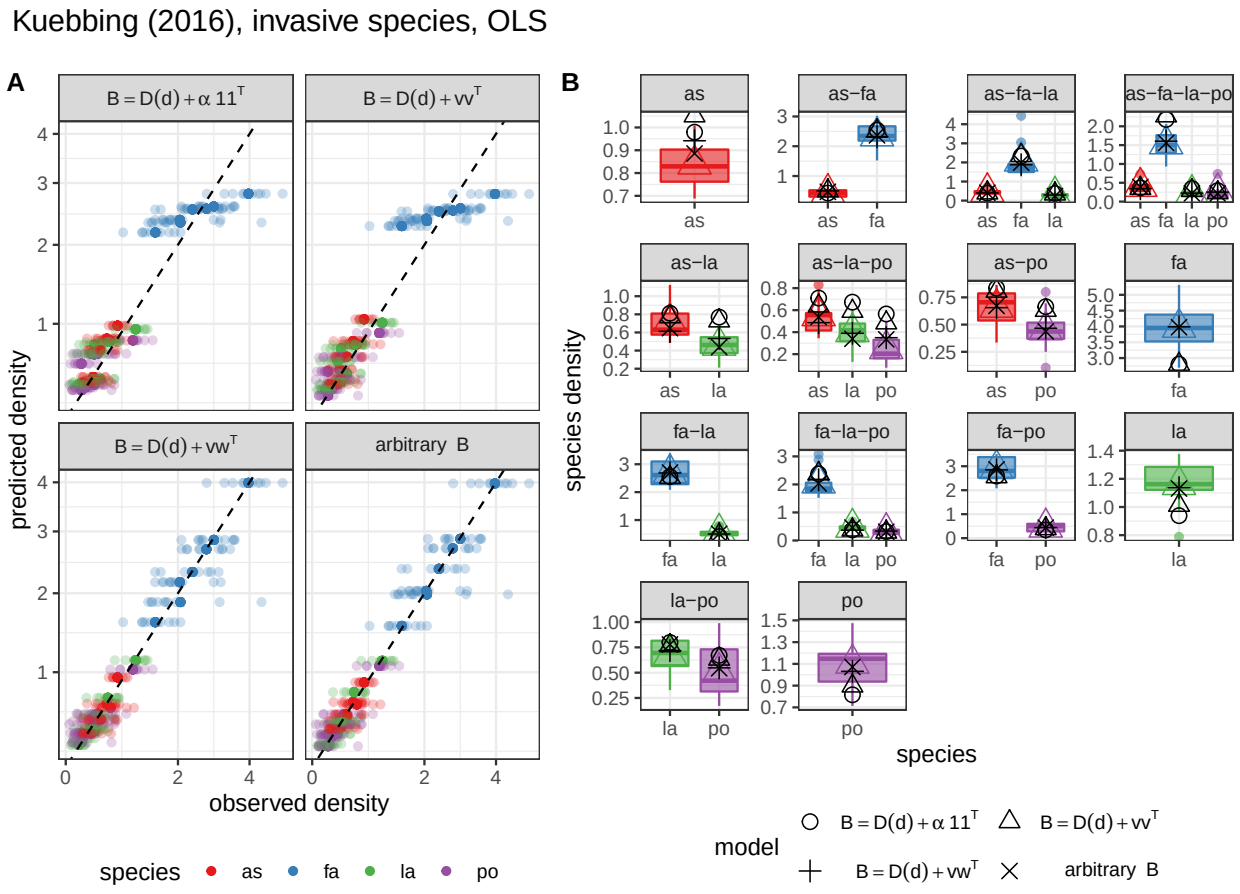

Figure 10: As Figure 2, but for the data of Kuebbing et al. (2015), non-native plants.

Ordinary Least Squares vs. Weighted Least Squares. When attempting to minimize the weighted sum of squared deviations, rather than the sum of squared deviations, we obtain a better fit for the species with low abundance, at the cost of a slightly less precise fit for the highly abundant species (Figures S9, S10, and S11).

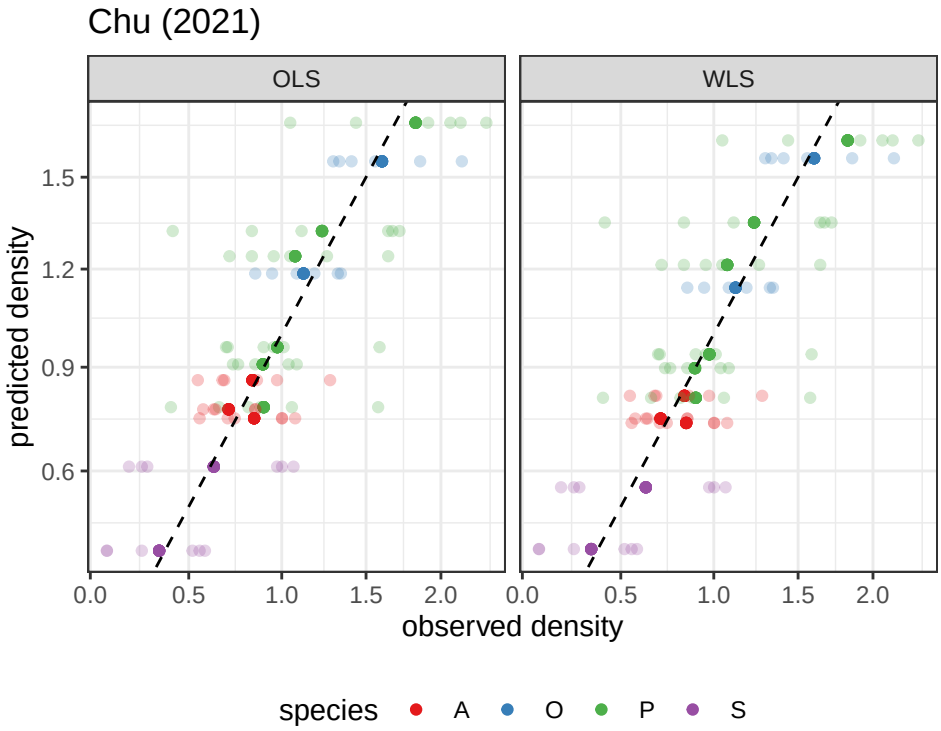

Figure 11: As Figure 5, but for the data of Chu et al., 2021.

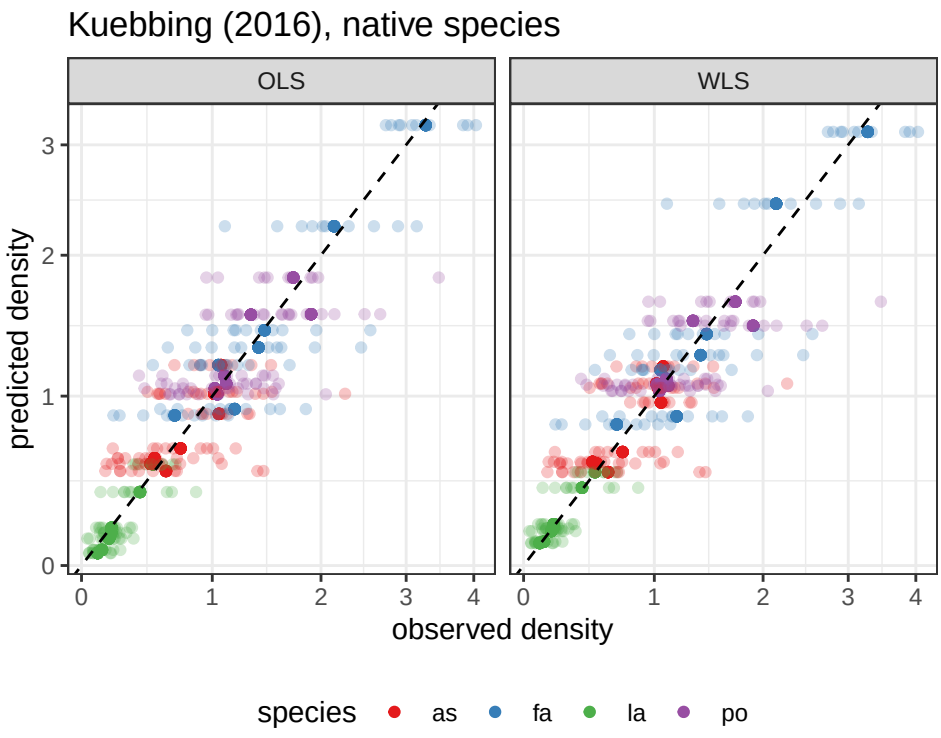

Figure 12: As Figure 5, but for the data of Kuebbing et al. (2015), native plants.

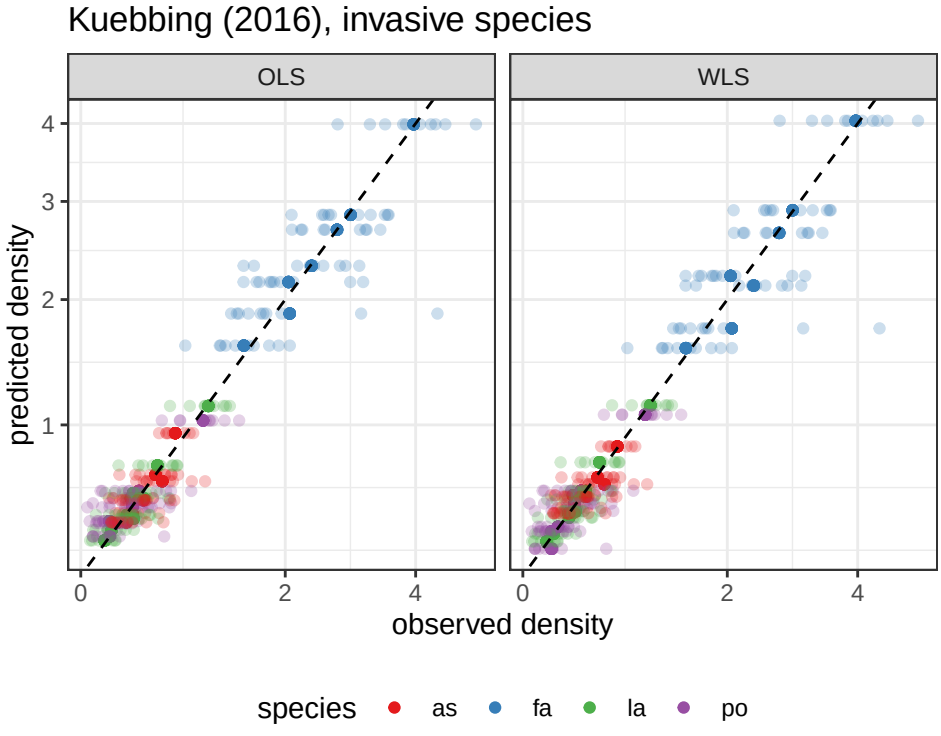

Figure 13: As Figure 5, but for the data of Kuebbing et al. (2015), non-native plants.

Out-of-sample predictions for the simplfied model $B = D(d) + vw^T$. For the data sets of Kuebbing et al. (2015), even using the simplified model $B = D(d) + vw^T$ we obtain excellent out-of-sample predictions (Figures S13 and S14): we correctly predict that the species coexist in the observed communities, and predicted densities that match the observed values quite closely. For the data from Chu et al. (2021), on the other hand, we make one qualitatively incorrect prediction (lack of coexistence, when the community is observed to coexist empirically), and also the quantitative predictions are worse, especially for the *Ochrobactrum* sp., which appears in only two communities (Figure S12). In all cases, using OLS or WLS yields similar results.

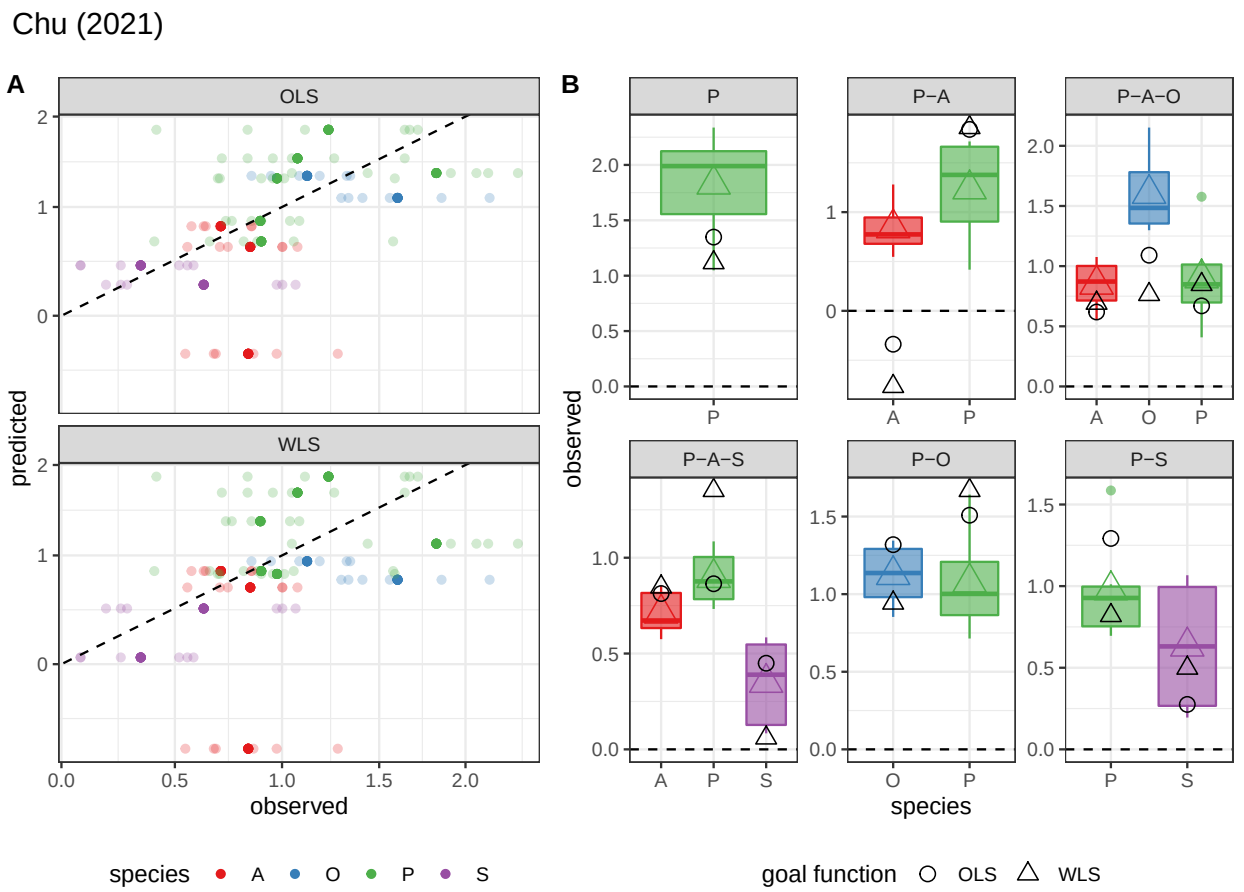

Figure 14: As Figure 6, but for the data of Chu et al., 2021. When we fit the model without the experiments in which *Pseudomonas fluorescens* and *Achromobacter* sp. are grown together, we obtain a qualitatively wrong prediction (lack of coexistence). *Ochrobactrum* sp. appears in only two communities, and when one of the communities is removed the fit is greatly reduced.

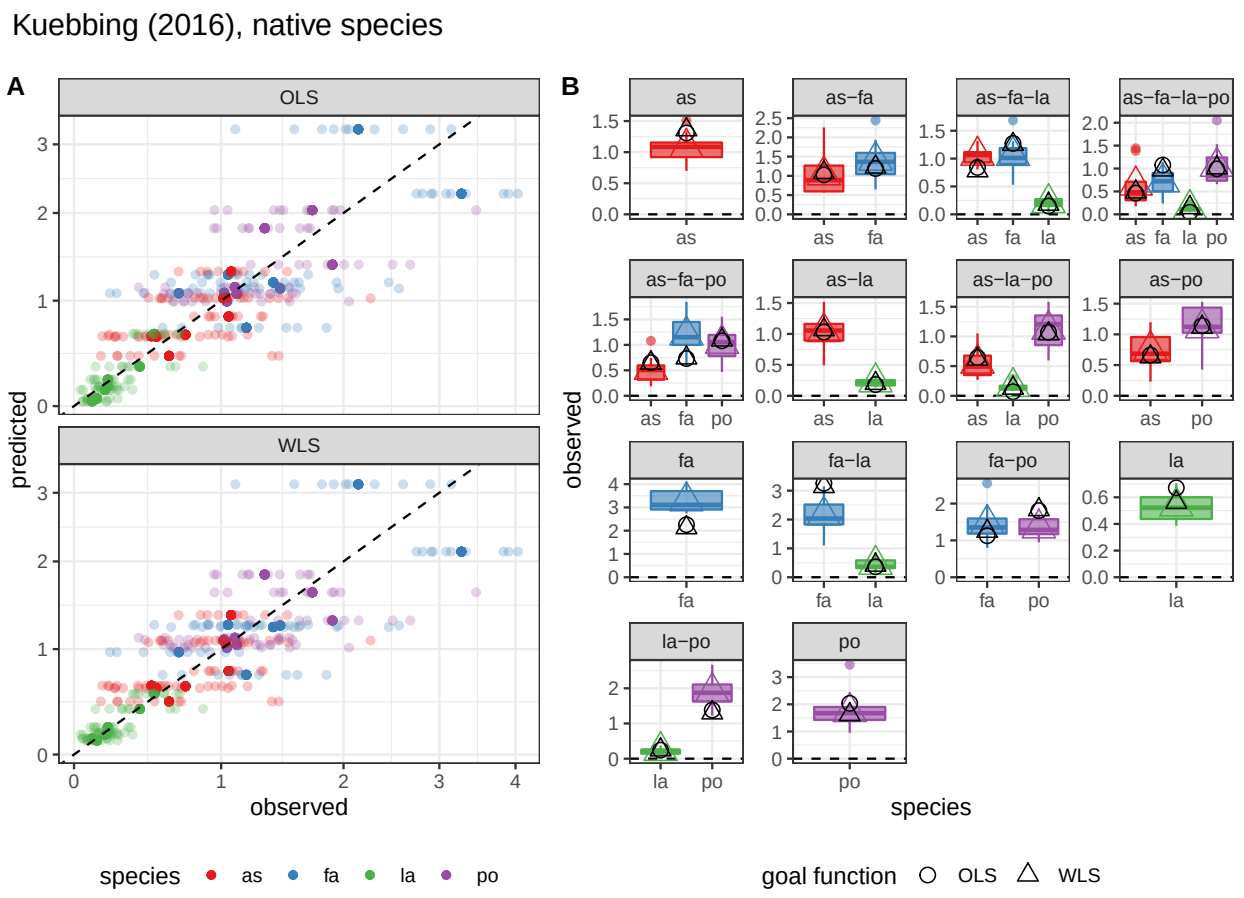

Figure 15: As Figure 6, but for the data of Kuebbing et al. (2015), native plants.

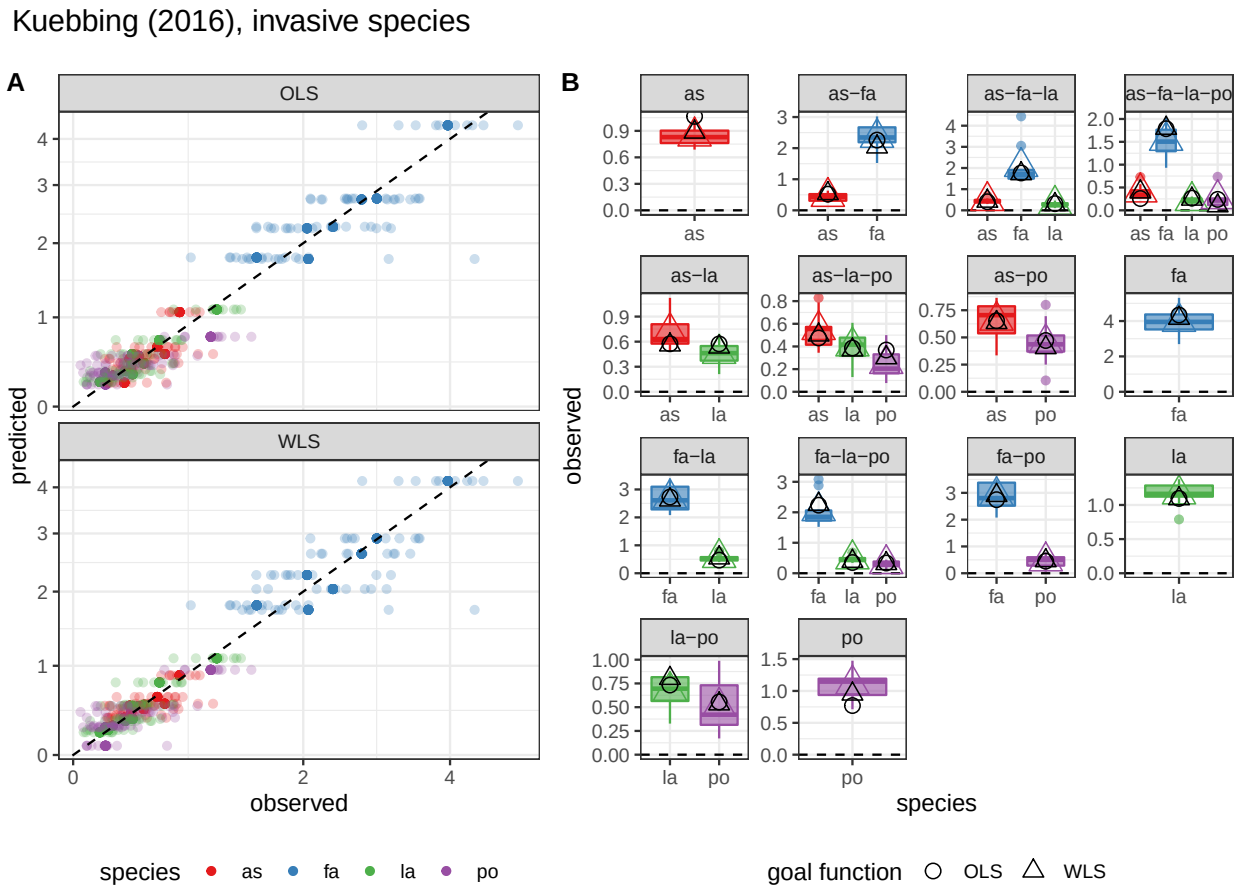

Figure 16: As Figure 6, but for the data of Kuebbing et al. (2015), non-native plants.

S9. Simulations

In this section, we present results obtained from simulations in which we first simulate noisy data, and then attempt to fit the "observed" points and recover the original parameters used to generate the data. These simulations allow us to highlight two aspects of the approach presented here: a) WLS is superior to OLS whenever the observations on abundances display a marked mean-variance relationship; and b) while the approach presented in the main text is fundamentally based on normally-distributed deviations from the theoretical mean (as found in both OLS and WLS), one could easily extend the approach to other distributions—in which case the function to maximize would be a more generic likelihood (rather than simply minimizing the sum of squared deviations).

All the simulations presented below are based on a simple, 3-species "true" matrix of interactions:

$$B = \begin{pmatrix} 1 & -0.05 & -0.1 \\ -0.4 & 0.1 & -0.07 \\ 0.05 & 0.05 & 0.001 \end{pmatrix}$$

which was chosen for three reasons: a) it yields coexistence for all species in all combinations; b) it yields "true" abundances $x_i^{(k)}$ spanning a large range of values (from $0.41\bar{6}$ to 1000); c) it contains a mix of "mutualistic/facilitation" and "consumer-resource" relationships (signs are reversed), showing that the approach presented here extends to systems in which competition is not the only or primary mode of interaction between species.

This matrix of interactions yields a matrix of "true" values X, detailing the abundance each species would achieve when grown in one of the seven possible combinations of species. We then simulate the matrix of "observed" abundances, $\tilde{X}$, by sampling each abundance $\tilde{x}_i^{(k,r)}$ independently from a given distribution with mean $x_i^{(k)}$. To explore distributions with prescribed mean-variance relationships, here we examine the case in which, for each community, we sample 50 noisy replicates and take $\tilde{x}_i^{(k,r)} \sim \text{Gamma}(x_i^{(k)}, 1)$, or $\tilde{x}_i^{(k,r)} \sim \text{InverseGaussian}(x_i^{(k)}, 1)$. For this parameterization of the Gamma distribution, we have that the expectation $\mathbb{E}[\tilde{x}_i^{(k,r)}] = x_i^{(k)}$, and the variance $\mathbb{V}[\tilde{x}_i^{(k,r)}] = x_i^{(k)}$, yielding a linear mean-variance relationship; for the Inverse Gaussian distribution, on the other hand, we have the same expectation $\mathbb{E}[\tilde{x}_i^{(k,r)}] = x_i^{(k)}$, but a variance that grows rapidly with the mean, $\mathbb{V}[\tilde{x}_i^{(k,r)}] = (x_i^{(k)})^3$.

Naturally, even if we were to choose the correct matrix B to fit the data, the residuals would not be normally-distributed around the corresponding means, because the shapes of these distributions depart considerably from normality (e.g., they are not symmetric around the mean). Because of this fact, and because the variance changes considerably with the mean, we expect the OLS to perform poorly—in particular, as explained in the main text, we expect low abundance values, which contribute less to the sum of squared deviations, to be fit especially poorly, while the high abundance measurements should be predicted with more accuracy. This problem should be considerably eased by performing WLS—in this case the problem of the different mean-variance relationships should be removed, while the fact that we are implicitly (and incorrectly) assuming normally-distributed (weighted) residuals remains. We can, however, attempt a different fitting routine: instead of searching for parameters minimizing the (weighted) squared deviations from the mean, we can specify an alternative form of the distribution and attempt to identify the maximum-likelihood parameters by searching for the matrix B maximizing the product of the densities of all observations, when either assuming that samples are taken from a Gamma distribution with shape parameter $x_i^{(k)} = ((B^{(k)})^{-1}1)_i$ and rate parameter 1, or alternatively from an Inverse Gaussian distribution with mean parameter $x_i^{(k)}$ and shape parameter 1. Note that because these are exactly the distributions used to generate the data, when given a sufficiently large number of replicates per community, our search should converge to the true matrix B.

The results are presented in Figs. 17-20. First, we examine the fitted abundances vs. "observed" abundances for the two distributions and three fitting routines (OLS, WLS, or likelihood-based). In Fig. S17 we show the results for simulations using a Gamma distribution. Here, as expected, the OLS approach performs poorly,

failing especially to fit the low-abundance points; the fit is greatly improved by WLS, which in fact fits the data almost as well as when maximizing the likelihood choosing the "right" distribution. The same is true for data sampled from the Inverse Gaussian distribution (Fig. S18).

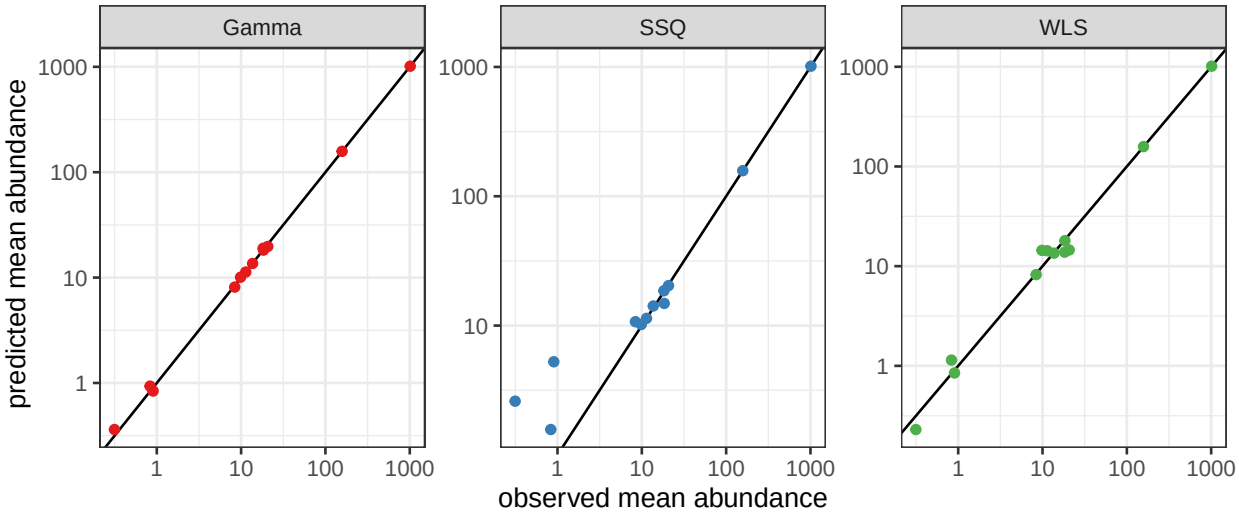

Figure 17: Fitted vs. observed densities for simulations in which replicates for a given system are sampled from a Gamma distribution. For each community, we plot the average abundance taken over the fifty replicates. The 1:1 line is indicated in black, for reference.

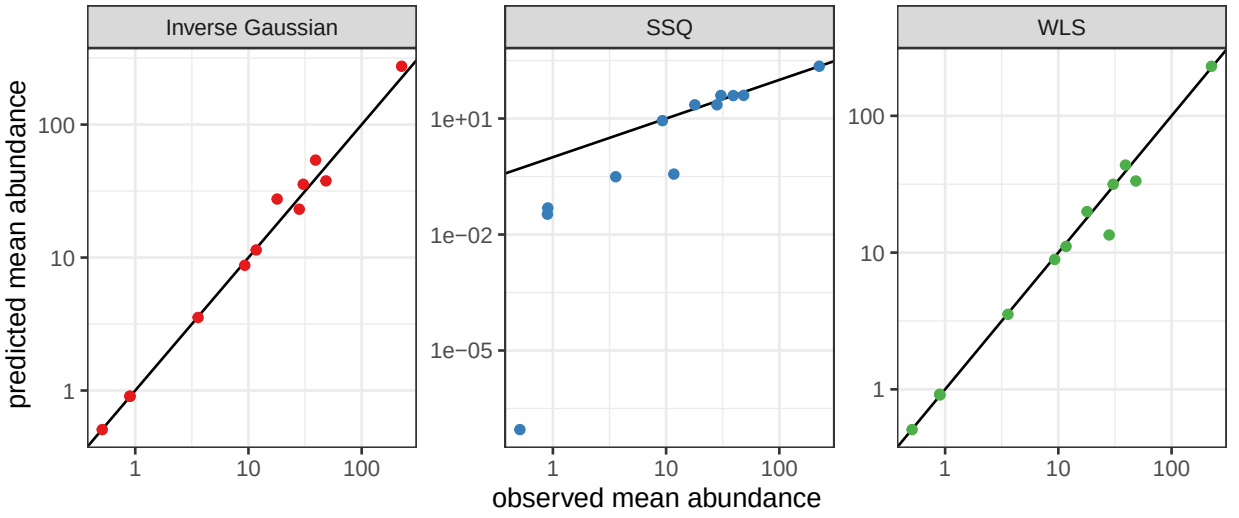

Figure 18: As Fig. S17, but simulating data sampled from an Inverse Gaussian distribution.

Next, we consider the quality of the parameter inference: how does the inferred B compare to that used to generate the data? In Fig. S19 (Gamma) and S20 (Inverse Gaussian), we show that indeed the parameter inference using WLS is much superior to OLS, further demonstrating the utility of this approach.

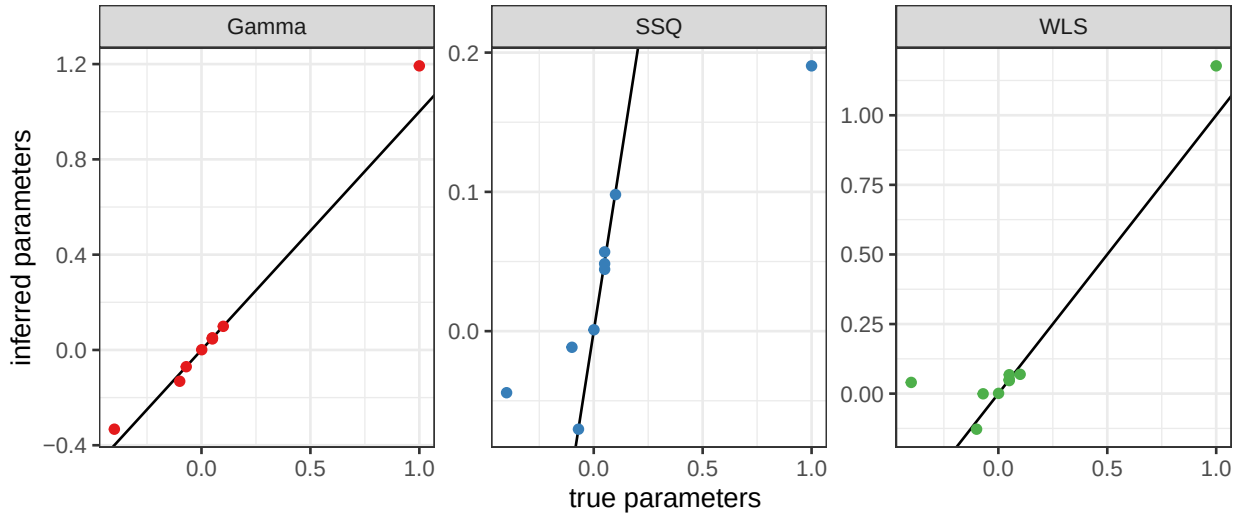

Figure 19: Inferred vs. true parameters for simulations in which replicates for a given system are sampled from a Gamma distribution. For each parameter, we plot the inferred value using the specified approach and fifty replicates for each community. The 1:1 line is indicated in black, for reference.

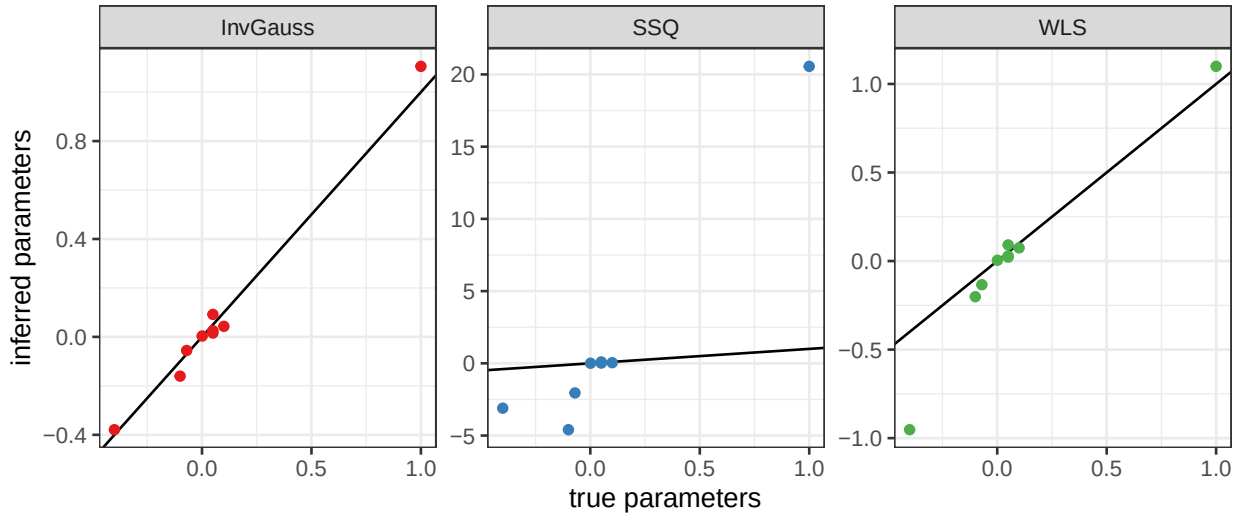

Figure 20: As Fig. S19, but simulating data sampled from an Inverse Gaussian distribution.

References

Chu, X.-L., Zhang, Q.-G., Buckling, A., & Castledine, M. (2021). Interspecific niche competition increases morphological diversity in multi-species microbial communities. *Frontiers in Microbiology*, 2103.

Ghedini, G., Marshall, D. J., & Loreau, M. (2022). Phytoplankton diversity affects biomass and energy production differently during community development. *Functional Ecology*, *36*, 446–457.

Hadwin, D., Harrison, K., & Ward, J. (2006). Rank-one completions of partial matrices and completely rank-nonincreasing linear functionals. *Proceedings of the American Mathematical Society*, *134*(8), 2169–2178.

Hofbauer, J., Sigmund, K., et al. (1998). *Evolutionary games and population dynamics*. Cambridge university press.

Kuebbing, S. E., Classen, A. T., Sanders, N. J., & Simberloff, D. (2015). Above-and below-ground effects of plant diversity depend on species origin: An experimental test with multiple invaders. *New Phytologist*, *208*(3), 727–735.

Lotka, A. J. (1920). Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences*, *6*(7), 410–415.

MacArthur, R. (1970). Species packing and competitive equilibrium for many species. *Theoretical Population Biology*, *1*(1), 1–11.

Maynard, D. S., Miller, Z. R., & Allesina, S. (2020). Predicting coexistence in experimental ecological communities. *Nature Ecology & Evolution*, *4*(1), 91–100.

Serván, C. A., & Allesina, S. (2021). Tractable models of ecological assembly. *Ecology Letters*, *24*(5), 1029–1037.

Serván, C. A., Capitán, J. A., Grilli, J., Morrison, K. E., & Allesina, S. (2018). Coexistence of many species in random ecosystems. *Nature Ecology & Evolution*, *2*(8), 1237–1242.

Sherman, J., & Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, *21*(1), 124–127.

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature*, *118*, 558–560.