Modeling ecological communities when composition is manipulated experimentally

Abigail Skwara[1,2,*] Paula Lemos-Costa[1] Zachary R. Miller[1]
Stefano Allesina[1,3]

[1] Department of Ecology & Evolution, University of Chicago, 1101 E. 57th st., Chicago, IL 60637, USA
[2] Department of Ecology & Evolutionary Biology, Yale University, New Haven, CT.
[3] Northwestern Institute for Complex Systems, Northwestern University, Evanston, IL.

[*] Correspondence: Abigail Skwara <abby.skwara@yale.edu>

Abstract

1. In a laboratory setting, the composition of ecological communities can be manipulated directly. Starting from a pool of n species, one can co-culture species in different combinations, spanning monocultures, pairs of species, and all the way up to the full pool. Here we advance methods aimed at inferring species interactions from data sets reporting the density attained by species in a variety of sub-communities formed from the same pool.

2. First, we introduce a fast algorithm to estimate parameters for simple statistical models describing these data, which can be combined with likelihood maximization approaches. Second, we derive from consumer-resource dynamics statistical models with few parameters, which can be applied to study systems where only a small fraction of the potential sub-communities have been observed. Third, we show how a Weighted Least Squares (WLS) framework can be used to account for the fact that species abundances often display a strong relationship between means and variances.

3. To illustrate our approach, we analyze data sets spanning plants, bacteria and phytoplankton, recovering a good fit to the data and demonstrating the ability to predict experiments out-of-fit.

4. We greatly extend the applicability of recently proposed methods, opening the door for the analysis of larger pools of species.

Keywords. Generalized Lotka-Volterra model; Generalized Linear model; weighted least squares; species interactions.

Running title. Modeling experimental ecological communities

Introduction

Efforts to relate empirical measurements of population growth to dynamical models can be traced back to the origins of ecology. For example, in his famous study proposing the logistic growth model, Verhulst (1838) presented the fitted time-series for the human population growth of France and Belgium. Interest in contrasting empirical data with models for population dynamics has only grown through the years (Gilpin, 1973; Powell & Steele, 2012), with the development of sophisticated approaches (Ellner et al., 2002) accounting for different sources of errors (Carpenter et al., 1994; De Valpine & Hastings, 2002), and important applications such as forecasting population trajectories (Clark et al., 2001; Sugihara et al., 2012), modeling the evolution of disease epidemics (Du et al., 2017), and detecting chaos in natural systems (Perry et al., 2012; Sugihara & May, 1990).

Because of the intrinsic difficulty of manipulating natural systems, much of the literature on these issues has historically focused on inferring the parameters of dynamical models from time-series observations (Downing et al., 2020; Ives et al., 2003; Vandermeer, 1969). With the advent of high-throughput laboratory techniques, however, a different approach has become viable. Instead of attempting to learn the parameters of a model by analyzing the fluctuations of several interacting populations through time, it is instead possible to use steady-state abundances recorded for a variety of community compositions, each captured at a single time point, to infer model parameters (Ansari et al., 2021; Fort, 2018; Maynard et al., 2020; Voit et al., 2021; Xiao et al., 2017). For a species pool of interest, the initial community composition is manipulated in a series of experiments, and then the resulting set of final community compositions can be used to estimate the parameters of a statistical model. For example, to infer the interactions among n species, one could perform a series of experiments in which either a single species is grown in isolation, or in which pairs, triplets, or larger subsets of species are co-cultured (Dormann & Roxburgh, 2005; Friedman et al., 2017). Once any transient dynamics have elapsed, densities of all surviving species are recorded. Provided that a sufficiently large and diverse set of sub-systems (i.e., distinct species compositions) have been observed, it is possible to infer the parameters of a statistical model for species interactions from these static measurements. Such statistical models can be derived from corresponding dynamical models, thereby linking static estimates to models for population dynamics.

We build upon previous work (Carrara et al., 2015; Maynard et al., 2020; Xiao et al., 2017) in which the density of each species is assumed to be linearly related to the densities of the other species in a community. This statistical structure arises naturally from the ubiquitous Generalized Lotka-Volterra (GLV) dynamical

model (Lotka, 1920; Volterra, 1926)—although it is not necessary for dynamics to obey this simple model to yield a good fit to data (Maynard et al., 2020). While very easy to formulate, this type of statistical model is both difficult and computationally expensive to fit. Here, we extend the approach of Maynard et al. (2020) in three main ways to overcome issues that have limited the application of these methods. First, we introduce a fast iterative algorithm that finds parameters yielding a good fit to data. This algorithm can be used in conjunction with more rigorous fitting approaches by providing a starting point for numerical maximum likelihood optimization. We show that this hybrid approach is computationally very efficient, and produces better results than current methods. Second, we derive simplified versions of the statistical model, by exploiting the parallel between the model structure and Lotka-Volterra dynamics. These versions of the model require the estimation of fewer parameters, while still providing a good fit to data and retaining a clear ecological interpretation. Finally, we show that a more sophisticated error model for the observations results in a Weighted Least Squares problem that can be solved efficiently. Extending the error structure offers more flexibility to fit empirical data, especially when the variance of experimental observations changes with the mean, as typically seen in ecological data (Grilli, 2020; Taylor, 1961). To illustrate these improvements, we examine four recently published data sets spanning communities of plants, bacteria, and phytoplankton (Chu et al., 2021; Ghedini et al., 2022; Kuebbing et al., 2015).

Materials and Methods

Experimental setup and data

Given a pool of n species, suppose we have performed a large set of experiments in which different combinations of species are co-cultured for a suitably long time. At the end of each experiment (once transient dynamics have elapsed), we measure and record the biomass/density of each extant species. The experiments encompass a variety of initial compositions and initial densities, and are conducted with replication. Provided that our measurements, after any species extinctions, span a sufficient variety of sub-communities, we can fit a simple statistical model to these empirical "endpoints", which may in turn be used to predict whether any particular subset of species k can coexist, and, if so, at what densities.

To test our models, we use four recently-published data sets that have a suitable experimental design. In particular, we consider two data sets from Kuebbing et al. (2015), who selected two pools (natives, non-natives) of four plants each, and grew them in 14 out of 15 possible combinations of species. Similarly, Ghedini et al. (2022) considered five phytoplankton species, and grew them in monoculture, in all possible pairs, and all together. Finally, we consider a subset of the data from Chu et al. (2021), consisting of communities of four bacterial strains co-cultured along with *Pseudomonas fluorescens* in different combinations. A detailed

description of each data set is reported in the Supporting Information S1.

Given that species' densities in these experiments span several orders of magnitude, we rescale all densities such that the average normalized density in each data set is exactly one. While this normalization does not affect the mathematical analysis, it might reduce the impact of numerical errors when performing computations.

A simple statistical framework

To develop our statistical framework we make two main assumptions. First, we take each observed measurement to be a noisy estimate of a "true" value:

$$\tilde{x}_i^{(k)} = x_i^{(k)} + \epsilon_i^{(k)} \quad . \tag{1}$$

That is, the observed density of species i when grown in community k, $\tilde{x}_i^{(k)}$, is given by a "true" value, $x_i^{(k)}$, modified by an error term, $\epsilon_i^{(k)}$. We can arrange our empirical data in a matrix $\tilde{X}$, with n columns (one for each species in the pool) and as many rows as the number of observed communities. In particular, $\tilde{X}_{ki} = \tilde{x}_i^{(k)}$ if species i is present in community k, and zero otherwise. Similarly $\mathcal{E}_{ki} = \epsilon_i^{(k)}$ if species i is in community k and zero otherwise. Then, we can rewrite Eq. 1 in matrix form:

$$\tilde{X} = X + \mathcal{E} \quad . \tag{2}$$

When we have several replicates, we assume that the density recorded for species i in community k and replicate r follows $\tilde{x}_i^{(r,k)} = x_i^{(k)} + \epsilon_i^{(k,r)}$, i.e., all replicate measurements of species i in community k stem from the same true value. To recover the notation in Eq. 2, we simply stack the replicates in matrix $\tilde{X}$ and the corresponding true means in matrix X (with repeated rows). This assumption amounts to ruling out "true multistability" (Xiao et al., 2017)—if a set of species coexists, it must always reach the same attractor (be it an equilibrium, cycle, or chaotic attractor). This also implies that community dynamics have a deterministic skeleton.

Second, following Maynard et al. (2020), we assume the true species' densities in a given community are linearly related to each other. For any species i in community k, we can express these densities by writing:

$$X_{ki} = \alpha_i - \sum_{j \neq i} \beta_{ij} X_{kj} \quad ,$$

where α_i is the density that species i would attain when grown in isolation, and the coefficients β_{ij} model the effects of the other species in community k on the density of i. Importantly, note that β_{ij} depends only on the identities of the species, and not on the community we are modeling—this assumption amounts to ruling out higher-order interactions or other nonlinearities that would make the per-capita effect of species j on i dependent on the state of the system. Rearranging, we obtain:

$$\begin{aligned} \sum_{j \neq i} \frac{\beta_{ij}}{\alpha_i} X_{kj} + \frac{1}{\alpha_i} X_{ki} = 1 \\ \sum_j B_{ij} X_{kj} = 1 \end{aligned} \quad ,$$

with $B_{ii} = 1/\alpha_i$ and $B_{ij} = \beta_{ij}/\alpha_i$.

Because $X_{kj} = 0$ whenever species j is not present in community k, we can define the sub-matrix $B^{(k)}$ obtained by retaining only the rows and columns of B corresponding to species that are present in community k. We similarly take $X^{(k)}$ to be a vector containing only the densities of the species in k, and $1^{(k)}$ to be a vector of ones with as many elements as $X^{(k)}$. Then $B^{(k)} X^{(k)} = 1^{(k)}$ for each community k.

Suppose that we have obtained a matrix B and we want to predict whether a particular set of species k can coexist. We solve the corresponding equation for $X^{(k)}$,

$$X^{(k)} = \left(B^{(k)}\right)^{-1} 1^{(k)} \quad , \tag{3}$$

with two possible outcomes: a) all the components in $X^{(k)}$ are positive, in which case we conclude that the species can coexist, with predicted densities $X^{(k)}$; or b) some of the components in $X^{(k)}$ are negative, which we interpret as the impossibility of coexistence of this combination of species (Maynard et al., 2020).

Arguably, the simplest version of this statistical model is obtained by assuming that errors are independent, identically-distributed random values sampled from a normal distribution, such that $\tilde{X}_{ki} \sim \mathcal{N}(X_{ki}, \sigma^2)$ whenever species i is in community k. Then, fitting the model requires minimizing the sum of squared deviations between the observed data and model predictions (Ordinary Least Squares, OLS).

This suggests a straightforward method to fit the model: propose a matrix B, compute the predicted densities for all species in all observed communities using Eq. 3, and search for the matrix B that minimizes the deviations between the observed data $\tilde{X}$ and the predicted X. Unfortunately, this method is quite inefficient and computationally very expensive, as it requires inverting a sub-matrix of B for each observed community, and there may be up to $2^n - 1$ unique community compositions that can be built from a pool of n species.

To circumvent this problem, Maynard et al. (2020) proposed a simple analytical approach (dubbed the "naïve method") to find a rough estimate of B from observed data; this initial estimate of B can then be used as a starting point for more sophisticated fitting routines. However, as discussed in their study, this method suffers from a number of issues (Supporting Information, S3), and one of the goals of the present work is to build an iterative algorithm that is not only computationally efficient, but that also improves upon the naïve estimate for B and yields a superior starting point for subsequent parameter search.

Results

A fast, iterative algorithm to estimate parameters

We want to find the maximum likelihood estimate for the matrix B, i.e., the choice of B minimizing the sum of squared deviations:

$$\mathrm{SSQ} = \sum_k \sum_r \sum_i \left(\tilde{x}_i^{(k,r)} - x_i^{(k)} \right)^2 \quad .$$

The computational bottleneck we face is that determining the predicted abundances $x_i^{(k)}$ from the matrix B (via Eq. 3) for all community compositions is very expensive. We therefore develop an algorithm in which this expensive calculation is performed only at defined intervals, and in between these steps optimization is carried out without having to re-calculate the predicted $x_i^{(k)}$. To achieve this goal, we divide the process of optimizing B into two steps: a prediction step (computationally expensive), and a numerical optimization step (with a computationally-cheaper goal function). By alternating between the two steps, we quickly converge to a good draft matrix B.

To construct this algorithm, we derive two "auxiliary" matrices that are useful for computation. Having arranged our data in the matrix $\tilde{X}$ as detailed above, we take $\tilde{X} = X + \mathcal{E}$ (Eq. 2), transpose each side, and multiply by B. We obtain the sum of two new matrices:

$$B\tilde{X}^T = BX^T + B\mathcal{E}^T = P^T + S^T \quad . \tag{4}$$

While in the reminder of this section we use P and S simply as a convenient means to build our algorithm, the matrix P has an interesting interpretation in terms of invasion growth rates when our statistical model is viewed as yielding all the possible feasible equilibria of a Generalized Lotka-Volterra model (Maynard et al., 2020) (Supporting Information, S4). From Eq. 4, we have $B^{-1}P^T = X^T$ and $B^{-1}S^T = \mathcal{E}^T$. Our sum of

squared deviations is simply the squared Frobenius norm of $\mathcal{E}$, $\sum_{ij} \mathcal{E}_{ij}^2 = \|\mathcal{E}\|_F^2 = \|\mathcal{E}\|$, and from Eq. 2 and 4, we can write:

$$\begin{aligned} \|\mathcal{E}^T\| &= \|B^{-1}S^T\| \\ &= \|\tilde{X}^T - B^{-1}P^T\| \quad . \end{aligned}$$

Our goal is therefore to find a matrix B such that $B^{-1}P^T$ is as close as possible to the observed data $\tilde{X}^T$. However, neither B nor P are known, although P can be calculated from X and B, and hence from B. We therefore attempt to minimize the sum of squared deviations through an iterative algorithm reminiscent of the Expectation-Maximization (EM) approach (Moon, 1996):

1. Propose a candidate matrix B;
2. Consider B to be fixed, and use it to compute X via Eq. 3 (prediction step);
3. Compute $P^T = BX^T$;
4. Consider P to be fixed, and find an updated B^{-1} by numerically minimizing $\|\tilde{X}^T - B^{-1}P^T\|$ (optimization step). Invert it to recover an updated B;
5. Repeat steps 2-4 until convergence.

In principle, this iterative algorithm depends on the proposal of the initial matrix: for our numerical experiments, we always start from the identity matrix (i.e., no interactions between species). Finally, we perform an additional numerical optimization to refine the results produced by the algorithm. The algorithm above does not guarantee that all the values of $X_{ki} \geq 0$, because the solution minimizing the sum of squared deviations does not need to contain only non-negative densities for the species. Clearly, such solutions would be biologically unattainable, and thus, when numerically refining the solution, we only accept proposals that yield non-negative predictions.

While this algorithm is not guaranteed to find the maximum likelihood estimate of B, we observe that in practice, the combination of our iterative algorithm and the final numerical optimization step yields very good solutions with reasonable computational runtime. As shown in Figure 1 (and Supporting Information, S7), the SSQ typically decays rapidly with the number of iterative steps, and the final numerical search provides an additional improvement. While for the data presented here the algorithm converges smoothly, in principle, one could observe the SSQ oscillating as the algorithm progresses. As for gradient descent and similar methods, this problem can be alleviated by introducing a "momentum" (Polyak, 1964; Qian, 1999), i.e., when updating matrix B in Step 4, one could take a weighted average of the matrix used to compute P

(the current estimate of B), and the matrix that maximizes the likelihood given P (the new estimate of B). The addition of momentum would help achieve a smooth, monotonic decay of the SSQ, at the cost of having to take a larger number of steps before convergence.

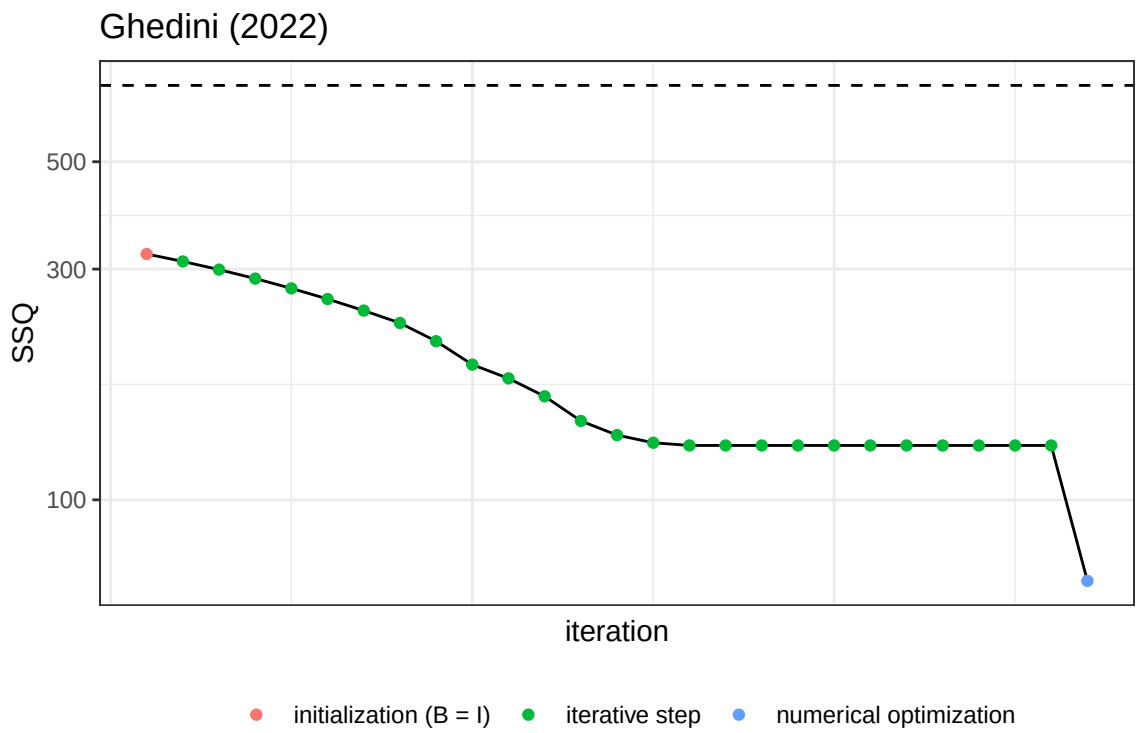

Figure 1: Sum of squared deviations (y-axis) for the initial matrix $B = I$, at each successive iteration of the algorithm, and after the numerical maximization (colors). The data from Ghedini et al., 2020. is used as an example. The dashed line marks the SSQ obtained using the naïve method of Maynard et al., 2020.

Simplified versions of the model

Fitting the n^2 parameters of the interaction matrix B requires a sufficiently varied set of experimental community compositions. It is necessary to observe each of the n species in at least n distinct communities, and to observe each possible pair of species together in at least one community (for a full derivation of these conditions, see Supporting Information, S2). These are very demanding requirements, and therefore many published data sets do not contain a sufficient variety of communities to be able to fit all parameters. These conditions grow more onerous with the number of species in the pool, making the approach unfeasible for species pools of even moderate size. To address this issue, we propose a nested set of simplified versions of the statistical model that use fewer parameters, but retain the basic model structure and a straightforward ecological interpretation. These simpler models have the added benefit of being extremely efficient to fit from a computational standpoint, and useful for model selection, especially when one suspects that the full model

with n^2 parameters may be susceptible to overfitting.

We follow a disciplined approach to develop these simpler models: we consider a version of the MacArthur's consumer-resource model (MacArthur, 1970) in which each species has access to its own private resources, and all species have access to a shared resource (Supporting Information, S5). By iteratively reducing the number of free parameters in the model, we obtain simpler structures for the matrix B (Table 1).

Table 1: Simpler versions of the model derived by considering a MacArthur's consumer-resource model in which species have access to a private pool of resources as well as a shared pool of resources. See Supporting Information, S5 for a derivation.

Model	Number of parameters	Interpretation
B arbitrary	n^2	Arbitrary interactions between the species.
$B = D(d) + vw^T$	$3n - 1$	Arbitrary intraspecific interactions; interspecific interactions $B_{ij} = v_i w_j$ given by the product of the "resource transformation" of species i (v_i), and the "attack rate" of species j (w_j).
$B = D(d) + vv^T$	$2n$	Arbitrary intraspecific interactions; resource transformation is the same for all species.
$B = D(d) + \alpha 11^T$	$n + 1$	Arbitrary intraspecific interactions; resource transformation and attack rates are the same for all species.

Given that the matrix B can be interpreted as detailing pairwise interactions (i.e., the effect of the density of species j on species i), the model $B = D(d) + vw^T$ corresponds to the following ecological picture: each species interacts with conspecifics through their private resources (coefficients d_i) and with all species via the shared resource. Interactions arising from the shared resource are given by the product of a resource utilization vector w (i.e., attack rates), and a resource transformation vector v, where each species is characterized by its v_i and w_i values. The interpretation of the simpler models is similar: by considering equal transformation rates one makes $vw^T = vv^T$ symmetric, and by assuming that all species also have the same attack rates $vw^T = \alpha 11^T$.

In Figure 2, we show the fit of these four models when analyzing the data reported in Ghedini et al. (2022).

For each community (different panels in the figure), we plot the fitted species' abundances for all four models against the observed species' abundances, with the boxplots showing the range of observed experimental values.

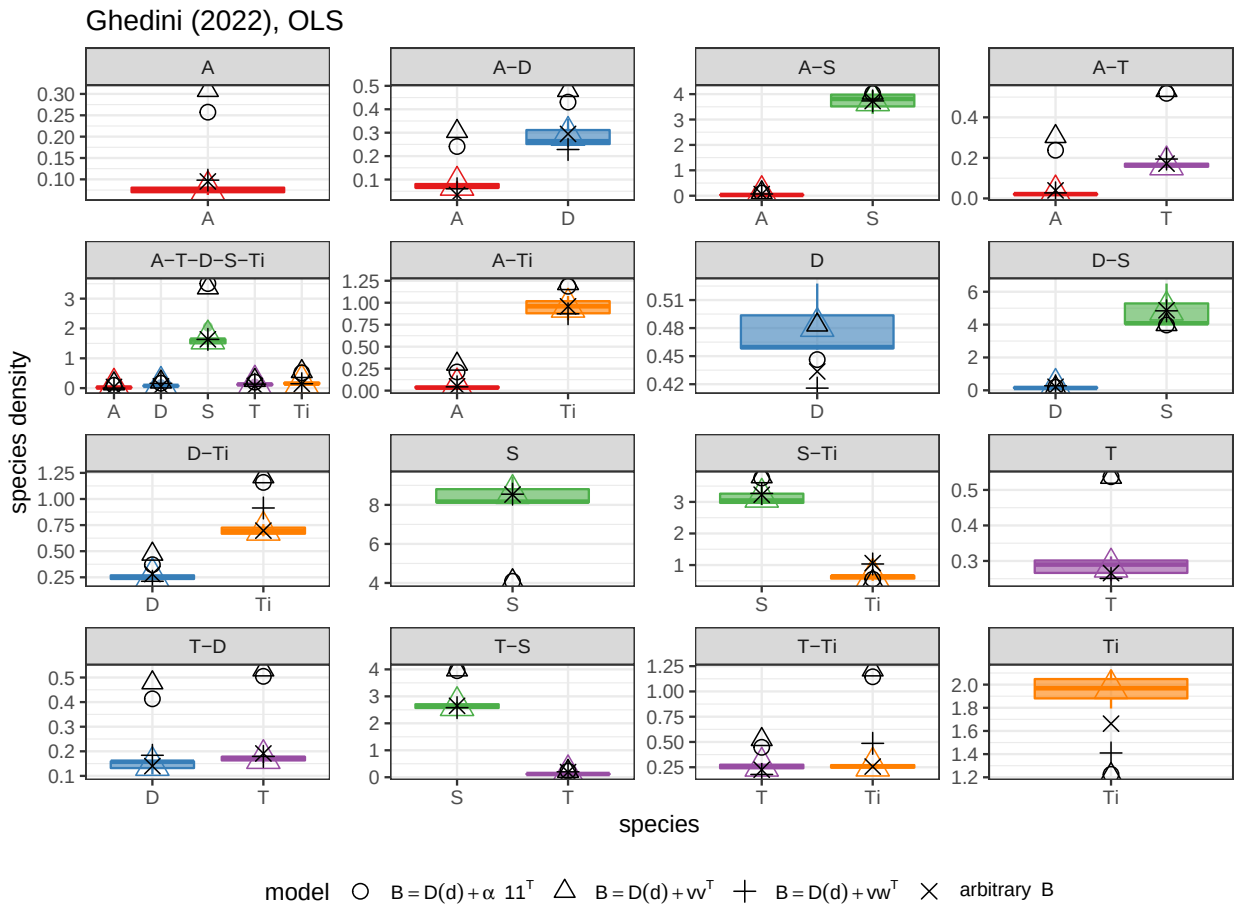

Figure 2: Observed species (color, x-axis) densities (y-axis) for each of the 16 communities cultured by Ghedini et al., 2022. Boxplots show the distribution of the species densities across replicates, with the median density reported as a solid colored line; the mean density is represented by the colored triangle. Predicted values for each species in each community are represented by black open symbols (one for each of the four versions of the model).

Importantly, the simplified models are computationally much less demanding, resulting in a much quicker runtime (Supporting Information, S6). Because all the simplified models involve a rank-one update of a diagonal matrix, the inverse of any sub-matrix can be computed analytically via the Sherman-Morrison formula (Sherman & Morrison, 1950), thereby removing the main computational bottleneck. For all data sets we find the same qualitative results: the full model (n^2 parameters, 25 parameters for the dataset of Ghedini et al. (2022)) and the diagonal plus rank-one matrix model (with $3n - 1$ parameters, 14 for Ghedini et al. (2022)) have very similar performance (Supporting Information, S7), while any further simplification results

in a marked loss of fit. These trends are also evident when contrasting the total SSQ across models and data sets (Figure 3).

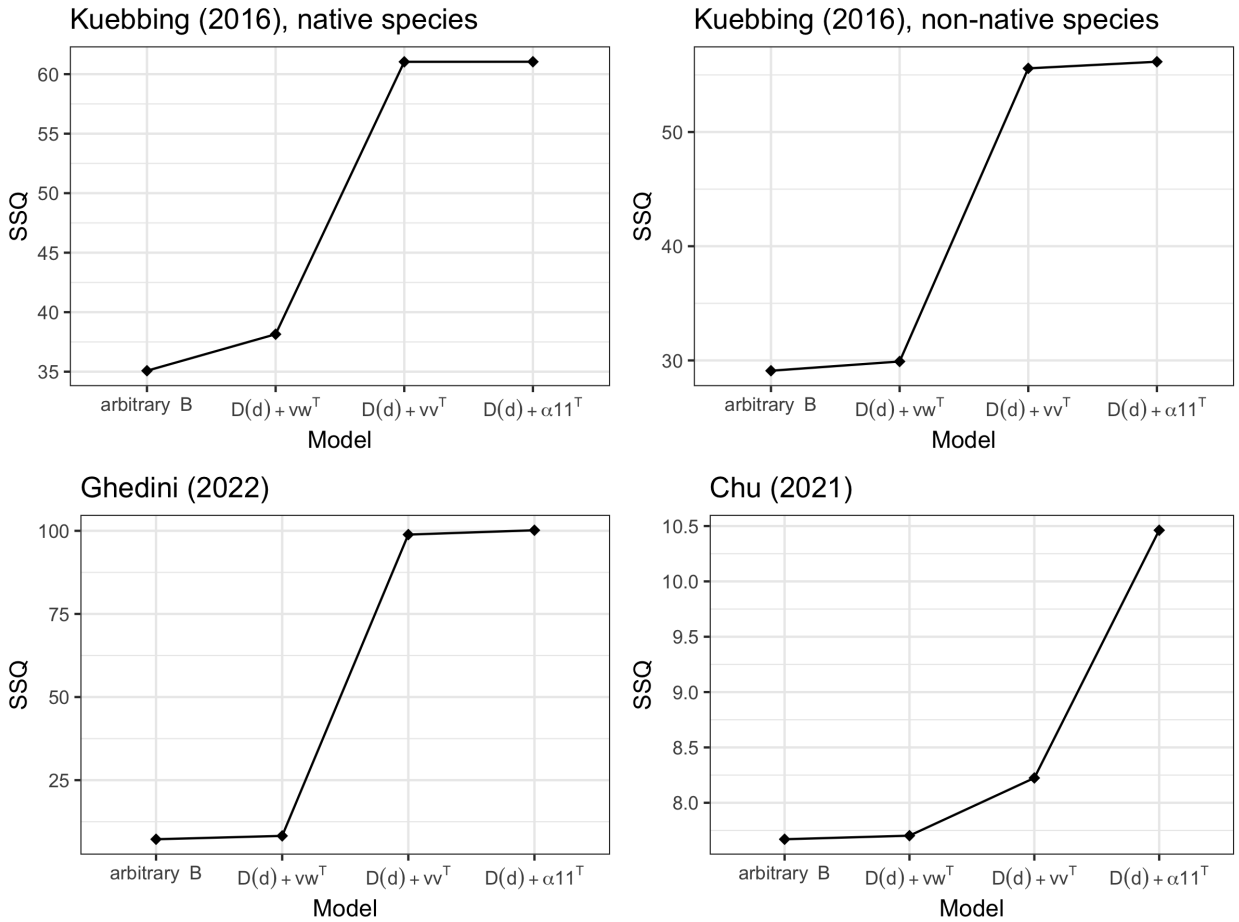

Figure 3: Sum of squared deviations between observed and predicted densities for the four data sets considered, and the four versions of the model.

Allowing the variance to change with the mean

So far, we have assumed that errors are independent and identically-distributed for all measurements. In many situations, however, this assumption would be quite unrealistic. For example, some species could systematically grow to much higher density than others—resulting in potentially larger absolute errors in their measurement—or measurements might be made on a small sub-plot or sample volume and then extrapolated to the whole plot or volume through multiplication. In such cases, the data would display marked heteroskedasticity (i.e., the variance would change with species density).

In Figure 4, we show that the variance scales with a power of the mean in the phytoplankton data from Ghedini et al. (2022). This is in fact the expected behavior of many ecological systems, as posited by Taylor's

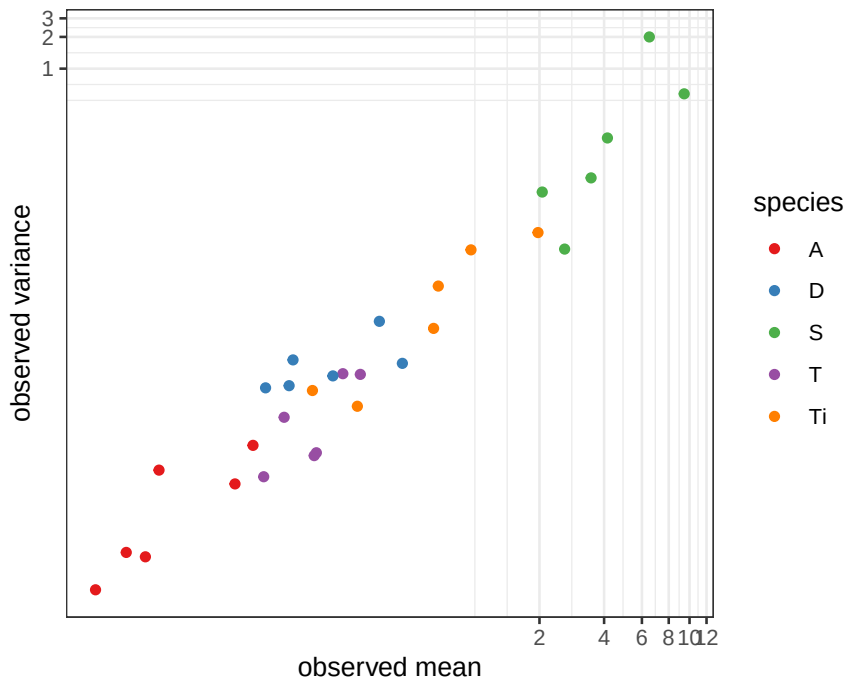

Figure 4: Mean (x-axis) vs. variance (y-axis) for the data published by Ghedini et al., 2022. For each species (colors) and community combination, the mean and variance of the observed species densities are computed across replicates. Note that both axes have a logarithmic scale, and thus the strong linear trend displayed by the data corresponds to a power-law relationship between the mean and the variance.

law (Gaston & McArdle, 1994; Routledge & Swartz, 1991; Taylor, 1961).

A possible approach to dealing with the systematic heteroskedasticity observed in these data sets is to use Weighted Least Squares (WLS) instead of OLS. To implement this approach, each squared residual is divided by the variance of the corresponding distribution, which is equivalent to measuring residuals as standard deviations of the standardized data. Here for simplicity we use the variance in the observed $\tilde{x}_i^{(r,k)}$ to estimate the variance ${\sigma_i^{(k)}}^2$. The WLS approach can be interpreted as reweighting the relative importance of errors made in our predictions, such that small errors made when estimating low species abundances are penalized as much as larger errors made when predicting higher species abundances. This is in contrast to OLS, where a one percent error in estimating the density of a species with high abundance contributes much more to the sum of squared deviations than the same proportional error for a species with low abundance.

Since experimental replicates are necessary to estimate these variances, we must have replicates for all communities in order to implement the WLS approach in this straightforward manner. All of the data sets considered here include such replicates, allowing us to estimate the variance for each species in each community.

If replicates are unavailable, or as an alternative to WLS, one could postulate a specific mean-variance relationship, and model abundances as noisy estimates taken from a distribution satisfying the required relationship. For example, sampling abundances from a Gamma distribution would produce variances that are proportional to the corresponding means; a Tweedie distribution could be used for power-law relationships with powers between 1 and 2, etc. This allows for additional flexibility in fitting data across a range of different ecological systems and with distinct experimental protocols that may have generated different error structures in the observed data.

In Figure 5 (and Supporting Information, S7), we compare these two error structures and find that we indeed observe a better fit for species with lower abundances when the WLS approach is used. This has important implications for the qualitative prediction of species' coexistence, because for a species present at low abundance in a community, the numerical difference between a positive abundance and a negative abundance (equivalent to a prediction of lack of coexistence) could be quite small.

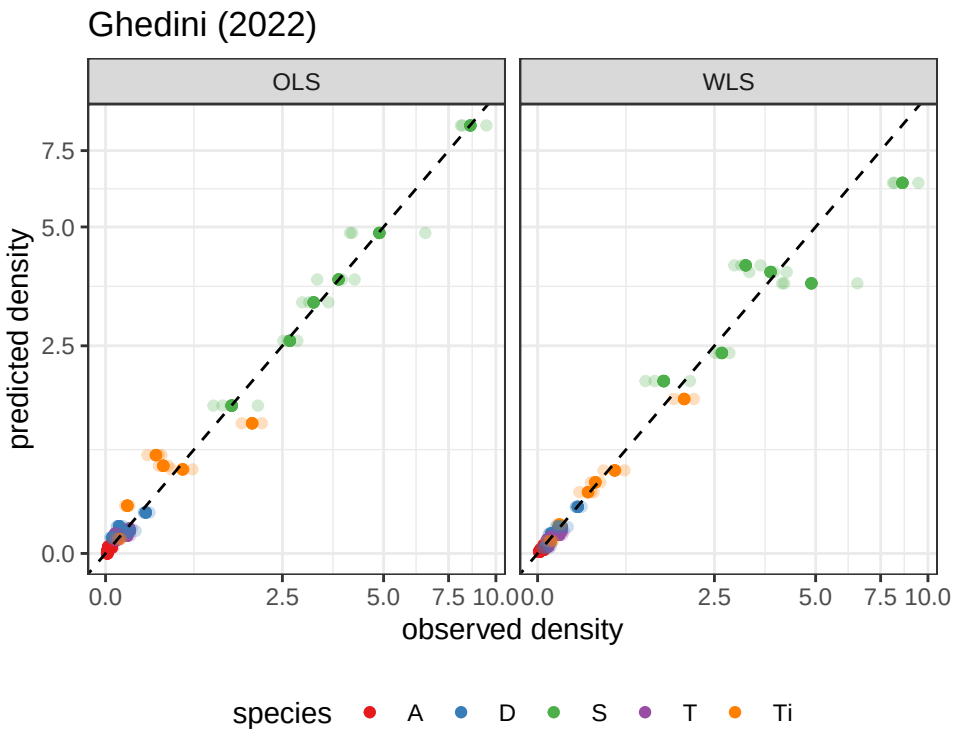

Figure 5: Oserved (x-axis) vs. predicted (y-axis) species densities in all communities using the data by Ghedini et al. (2022), when fitting the simplified model $B = D(d) + vw^T$ and either minimizing the sum of squared deviations (OLS) or the sum of standardized squared deviations (WLS). Replicate measurements for each species/community are reported as semi-transparent points; the mean for each species/community combination is reported as a solid point. In OLS, small (proportional) deviations of highly abundant species (e.g., *Synechococcus* sp., in green) are penalized more than larger (proportional) deviations of lower-abundance species (e.g., *Tisochrysis lutea*, in yellow). In contrast, when performing WLS each data point is standardized by its corresponding variance, leveling the importance of each measurement.

Predicting experiments out-of-fit

For the data sets considered here, we always find good agreement between the observed and the fitted values when using the full model under either an OLS or WLS scheme. We also consider a leave-one-out (LOO) cross-validation approach, to verify that the models capture real features of the interactions between species in the communities, and are not simply over-fitting the data. For a data set in which m experimental communities have been measured, we implement the LOO approach by designating one of the communities (along with any replicates) as out-of-fit, and fitting our model on the remaining $m-1$ communities. This process can be repeated for each of the m communities in turn, and we then compare the predicted species' abundances with their experimentally observed values, as shown in Figure 6 (and Supporting Information, S7). While the quality of the predictions is necessarily worse, in almost all cases we would have correctly predicted the experimental outcome both qualitatively (i.e., possibility of coexistence) and quantitatively.

Discussion

In this study, we have further developed a simple, extensible framework to infer species' interactions from experimental data where community composition has been manipulated, and showcased its performance by testing it on data sets spanning plant, phytoplankton, and bacterial communities. We have improved the computational efficiency of the fitting routines, derived simplified versions of the model, and extended the error structure used to capture the variability in the data.

Previous studies have used a variety of approaches to model this type of experimental data, including bottom-up "assembly rules", in which the focus is on observing all possible pairs of species and using the outcomes to predict coexistence of larger species assemblages (Dormann & Roxburgh, 2005; Friedman et al., 2017); methods relying on monocultures and leave-one-out communities (Ansari et al., 2021; Venturelli et al., 2018), using the extremes of very simple and highly speciose communities to infer pairwise interactions; and methods based on time-series that use fluctuations to estimate interaction strengths and the effects of environmental variables on ecological dynamics (Downing et al., 2020; Ives et al., 2003).

Our framework bridges the dynamical and statistical perspectives of these approaches. Importantly, though the models presented here are statistical in nature, they have a straightforward connection to the Generalized Lotka-Volterra dynamical model. By exploiting this parallel, we have derived a family of simplified models that put a premium on the ecological interpretability of the parameters. While an alternative approach to model regularization would be data-driven, machine learning methods (e.g., enforcing the sparsity or parsimony of B through a penalized regression), we have shown that ecologically-motivated model constraints

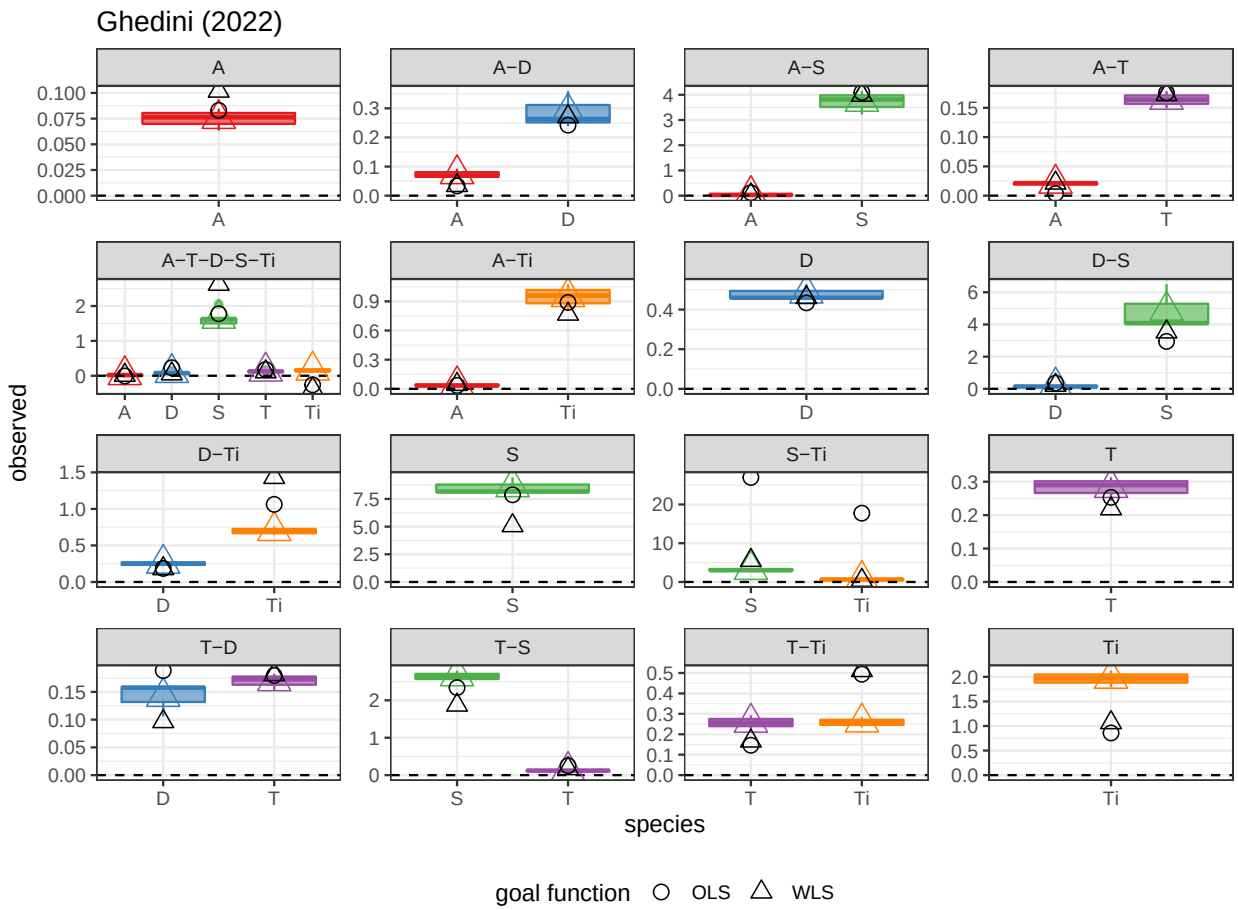

Figure 6: As Figure 2, but showing only out-of-fit predictions. Each panel is obtained by excluding the corresponding data, fitting the model using the remaining data, and then predicting the data reported in the panel out-of-fit. Results are presented for the model $B = D(d) + vw^T$ under either OLS or WLS. Note that in the panel displaying the measurements for the full community, *Tisochrysis lutea* is predicted at a negative abundance—without having observed these data, our models would suggest a lack of coexistence.

are a viable option. By clearly relating statistical and dynamical models, we retain the ability to probe other properties of these systems, for example in relation to invasibility, assembly and dynamical stability (Maynard et al., 2020).

Using our framework, we are able to obtain quantitative predictions for species abundances in arbitrary sub-communities both in- and out-of-fit, even for data sets that comprise just a fraction of the possible sub-communities that may be formed from a fixed pool of species. Here, our simpler models are used to fit the relatively sparse data sets when fitting the full model is impossible. As the number of communities that can be formed from a set of n species grows exponentially with n, this ability to predict species abundances out-of-fit is critical for for exploring larger systems. Additionally, a good out-of-fit performance indicates that the models are capturing meaningful information, rather than merely over-fitting the data.

Surprisingly, we find that a model in which the interactions between all pairs of species is given by the outer product of two vectors achieves a comparable level of accuracy to the full model for the data sets, suggesting that one could successfully predict experiment outcomes out-of-fit even when the observed data are not sufficient to determine all the parameters of the full model. This result also suggests that the ecological interactions in the systems analyzed here are not completely idiosyncratic, but rather are largely characterized by a simpler structure. This observation agrees with the finding that simple rules govern the structure of interactions in microbial communities (Friedman et al., 2017), as well as recent work suggesting that plant-plant competitive interactions are characterized by low dimensionality (Stouffer et al., 2021). Similarly, these patterns are consistent with the success of a sparse-modeling approach demonstrating that, when considering only a few focal species of plants, many heterospecific interactions can be captured by a generic interaction term (Weiss-Lehman et al., 2022). Our results suggest that effective interspecific interactions sit in between extremely low-dimensional (i.e., characterized by a generic interaction term, as in the model $B = D(d) + \alpha 11^T$, which has a poor performance across all data sets), and fully structured pairwise interactions (arbitrary B).

Naturally, to develop these simple models we need to make certain strong assumptions about the community dynamics. In particular, we ruled out "true multistability" (which, however, may be common in ecological systems), as well as higher-order interactions or other strong non-linearities that would make the effect of one species on another context-dependent. While these two simplifying assumptions are supported by the good agreement with experimental data found in our results, it is important to keep these caveats in mind. The main outstanding problem with our approach is the assumption that dynamics have reached a steady state, and thus the observed community composition is the true final composition for the system. Suppose, for example, that we have two species, A and B, and that A excludes B asymptotically. If we sample this

community before the extinction of B, we force the model to find parameters consistent with the robust coexistence of A and B, thereby biasing the results considerably. This problem of "spurious coexistence" is especially troublesome for microbial communities, where species' presence is frequently determined by sequencing. Sequencing-based methods often detect some species at very low densities (Venturelli et al., 2018), potentially spuriously, making it difficult to discriminate between coexistence of rare species and actual extinctions masked by "background noise". Because this issue is driven by uncertainty about species with low abundances, our Weighted Least Squares approach, which is more sensitive to these species, might reduce the severity of this problem. Ideally, one could go further and jointly estimate the composition of the community (i.e., is this community truly formed by A and B, or only by A?) and the species densities—which could be accomplished in a Bayesian framework with latent variables for species presence/absence. The main obstacle to such a development is the combinatorial nature of community composition, which makes predictions computationally very expensive. In this respect, the simplified models presented here could open the door for a more comprehensive solution to address spurious coexistence.

This general framework can be further extended in a number of directions. For example, one could introduce a more sophisticated error model assuming that species abundances are correlated within communities (e.g., overestimating the density of a predator could be associated with an underestimate of the densities of its prey). More generally, once we have determined the means as encoded by Eq. 3, the statistical model takes the form of a simple Generalized Linear Model—the large toolbox and body of knowledge available for GLMs directly translates into possible improvements for the statistical structure in this framework. Similarly, while OLS and WLS are intrinsically tied to the normal distribution (and thus support over the entire real line), one could improve upon the approach presented here by sampling from distributions that have support only on the positive half-line (e.g., Gamma or Log-Normal distribution), automatically guaranteeing that predictions are non-negative.

Overall, the methods presented make it easier to contrast experimental data with simple models for population dynamics, returning parameters that have clear ecological interpretation, and allowing us to test predictions in a straightforward manner. The type of ecological data examined here is appearing with increasing frequency in the ecological literature, and these methods provide a complementary (or alternative) approach to model-fitting via time-series analysis. The minimal data needed to fit the simplified models and the fact that only a single measurement per community is required make this framework especially appealing and cost-effective.

Acknowledgments. This work was supported by the National Science Foundation, award DEB-#2022742 to SA. We thank P. Lechon-Alonso and K. Della Libera for comments.

Authorship statement. All authors contributed to the development of the methods. AS and SA wrote the code, analyzed the data, and wrote the manuscript. All authors edited the manuscript.

Data accessibility statement. All data and model code necessary to recreate the analyses presented in this manuscript are archived at `https://github.com/StefanoAllesina/skwara_et_al_2022`

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 S1: 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 declined/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 strains, *Achromobacter* sp. (A), *Ochrobactrum* sp. (O), *Stenotrophomonas* sp. (S), and *Variovorax* sp. (V). The

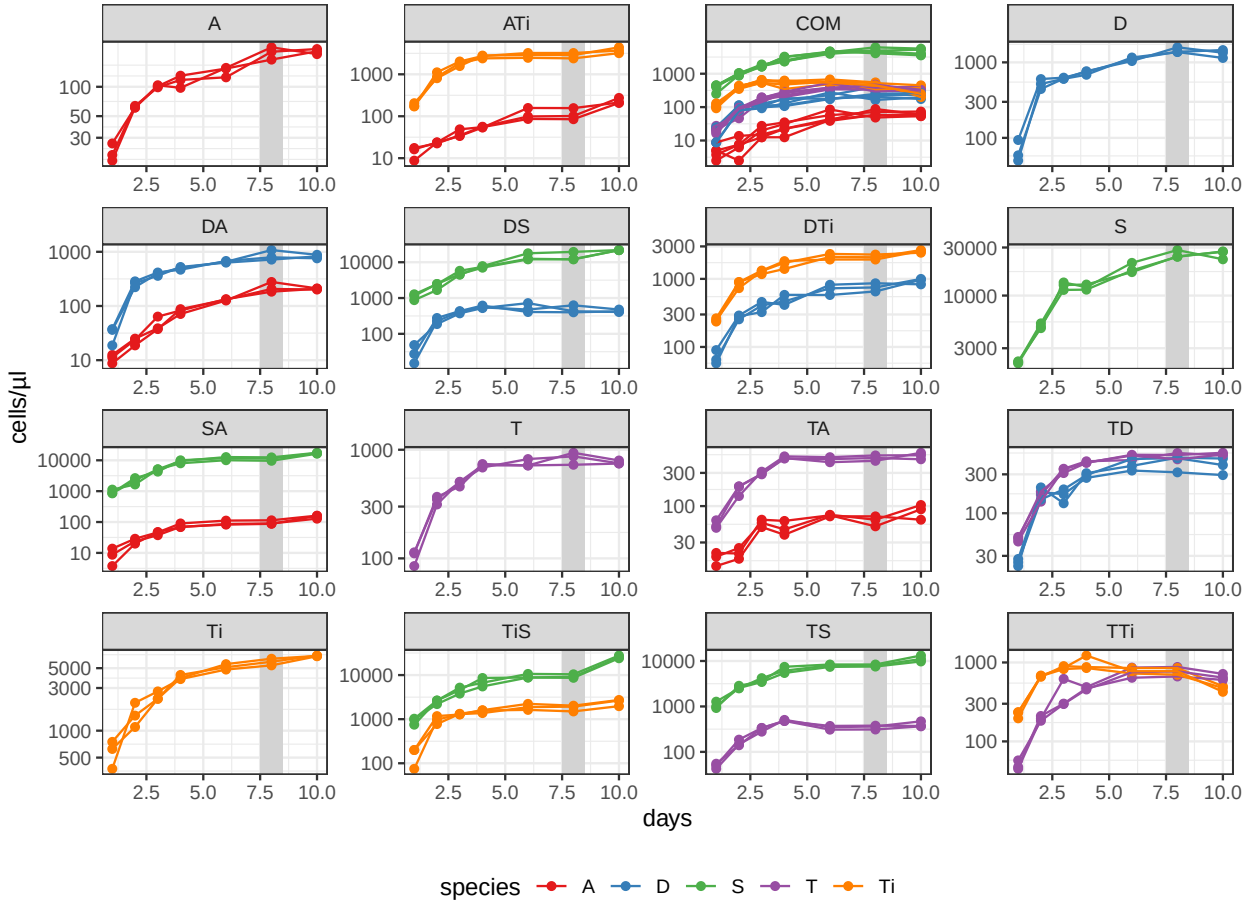

Figure S1: Time series for the growth of the 16 communities considered by Ghedini et al., 2022. For the analysis presented here, we considered the densities measured on day 8 (highlighted in the panels).

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.

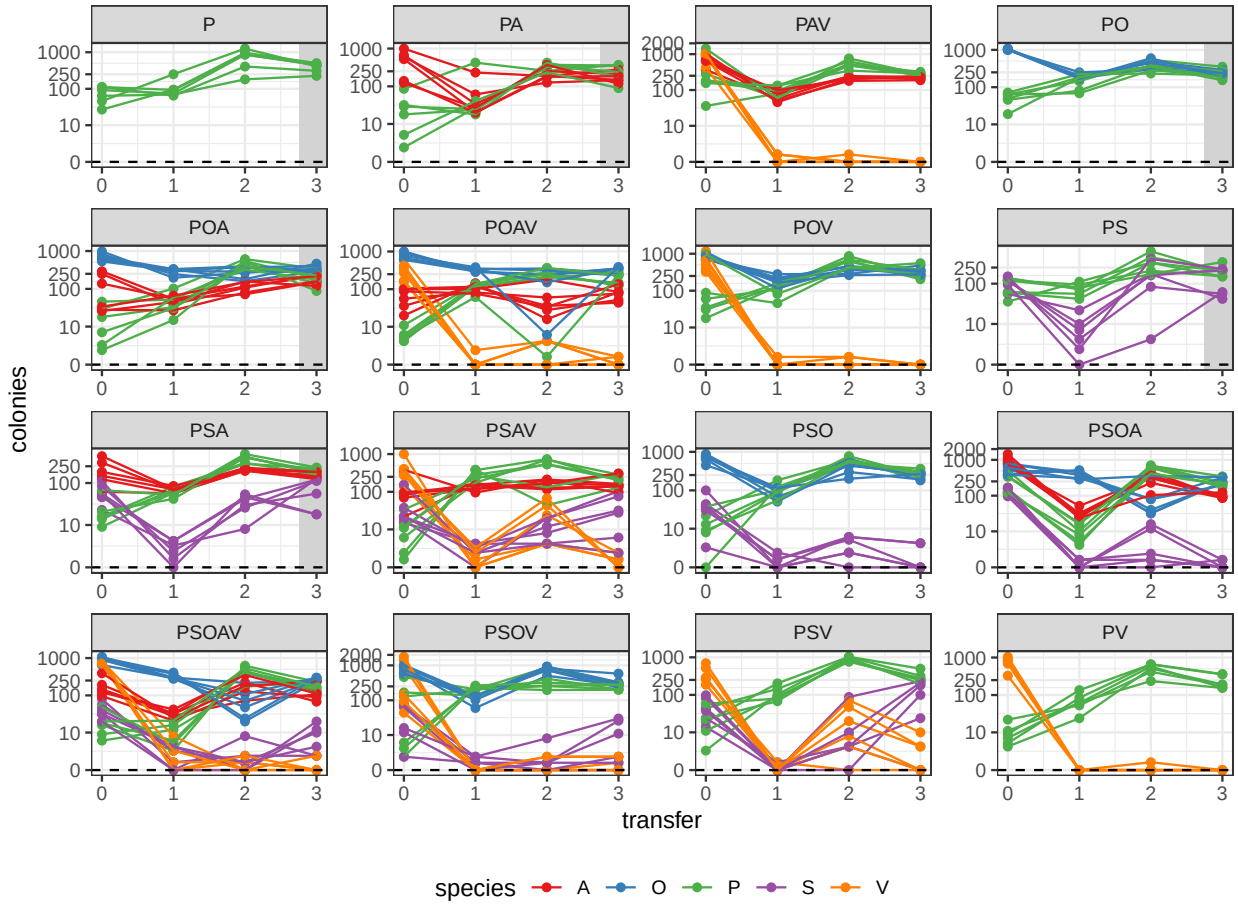

Figure S2: Time series for the growth of the 16 communities considered by Chu et al., 2021. For the analysis presented here, we considered the densities measured at the end of the experiment, for communities that did not include *Variovax* sp., because the strain went extinct in the majority of experiments. 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.

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 sub-community 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 with 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:

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

where $\text{vec}(A)$ 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:

$$\text{diag}(\text{vec}(\tilde{P}))(X \otimes I)\text{vec}(B) = \text{vec}(\tilde{P}^T) \quad , \tag{5}$$

where $\text{diag}(b)$ 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. 5 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).

S2. 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.

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 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.

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:

$$\dot{x}(t) = D(x(t))(r - Ax(t)) \quad , \tag{6}$$

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:

$$\dot{x}(t) = D(r \circ x(t))(1 - D(r)^{-1}Ax(t)) = D(r \circ x(t))(1 - Bx(t)) \quad . \tag{7}$$

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 separately A and r, but 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., those that are absent), 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 (at least) another n parameters. 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. 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 community 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. 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).

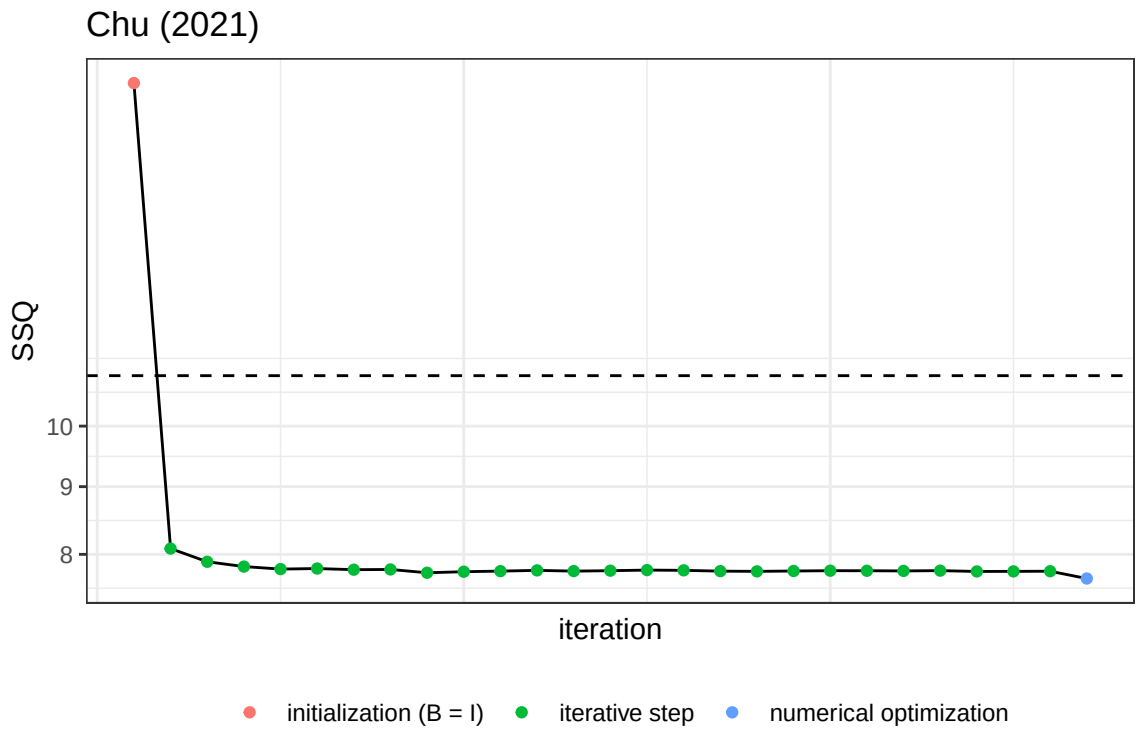

Figure S3: As Figure 1, but for the data of Chu et al., 2021.

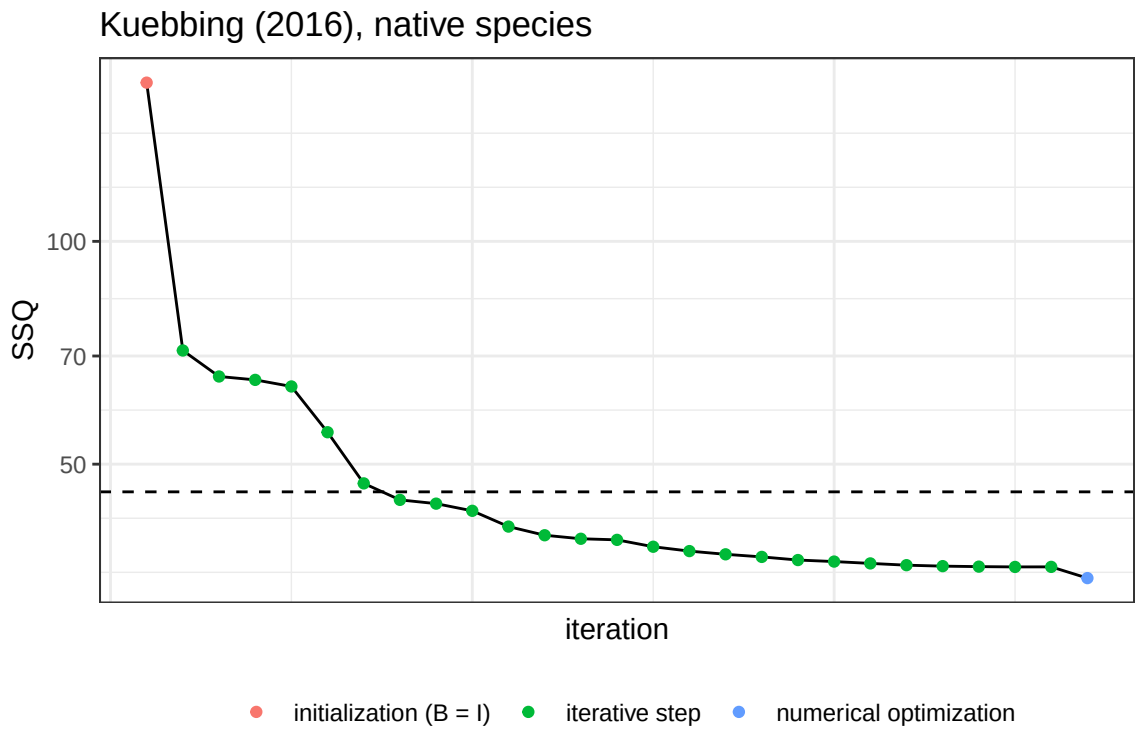

Figure S4: As Figure 1, but for the data of Kuebbing et al. (2015), native plants).

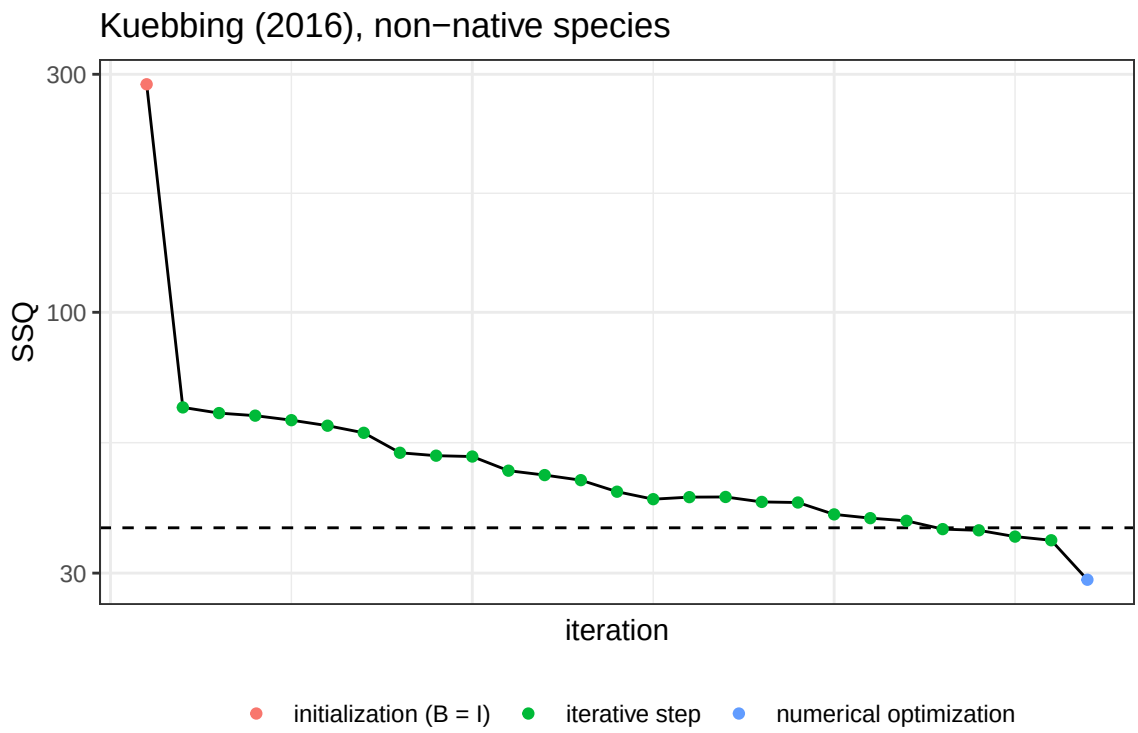

Figure S5: As Figure 1, but for the data of Kuebbing et al. (2015), non-native plants.

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 $B = D(d) + vw^T$ perform similarly to the full model with n^2 parameters (Figures S6, S7, and S8).

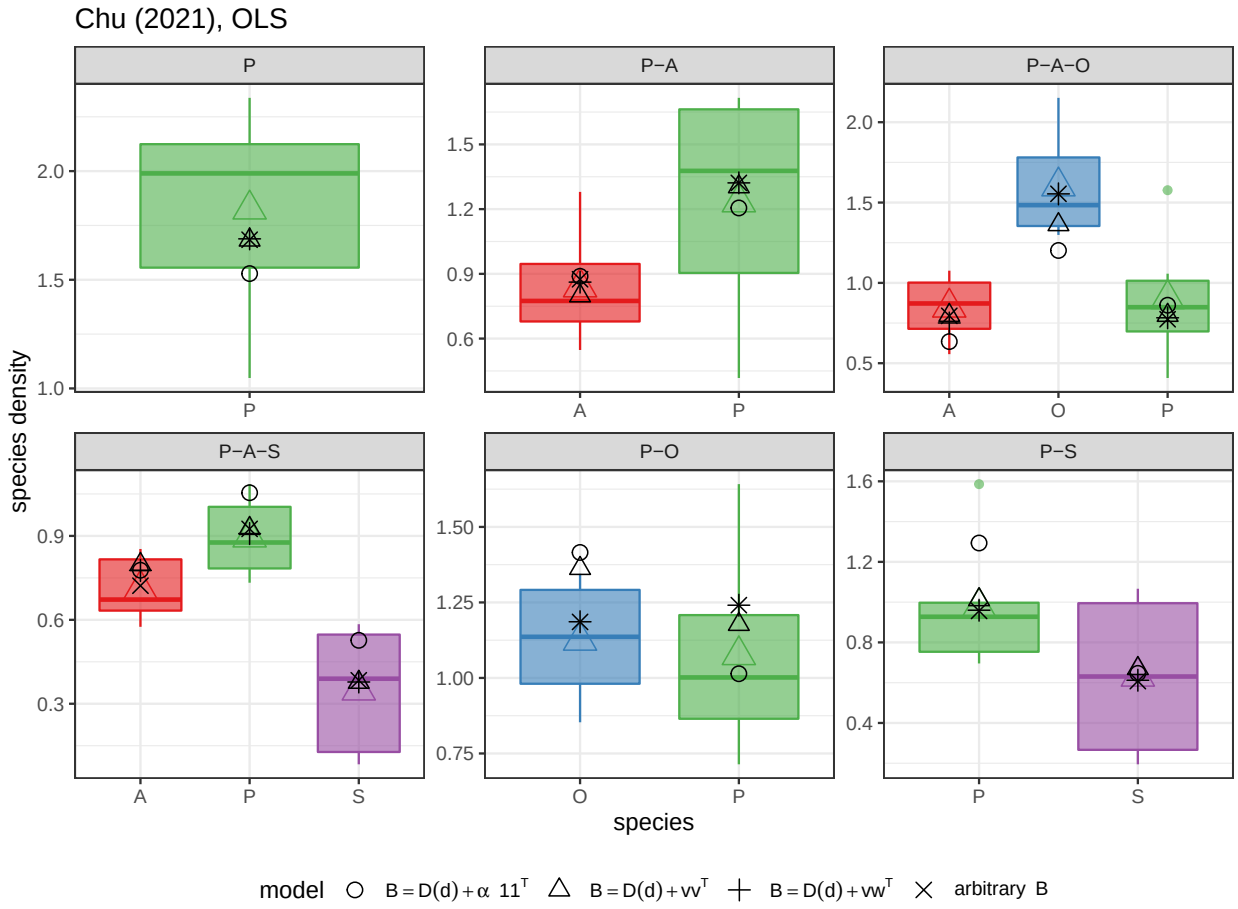

Figure S6: As Figure 2, but for the data of Chu et al., 2021.

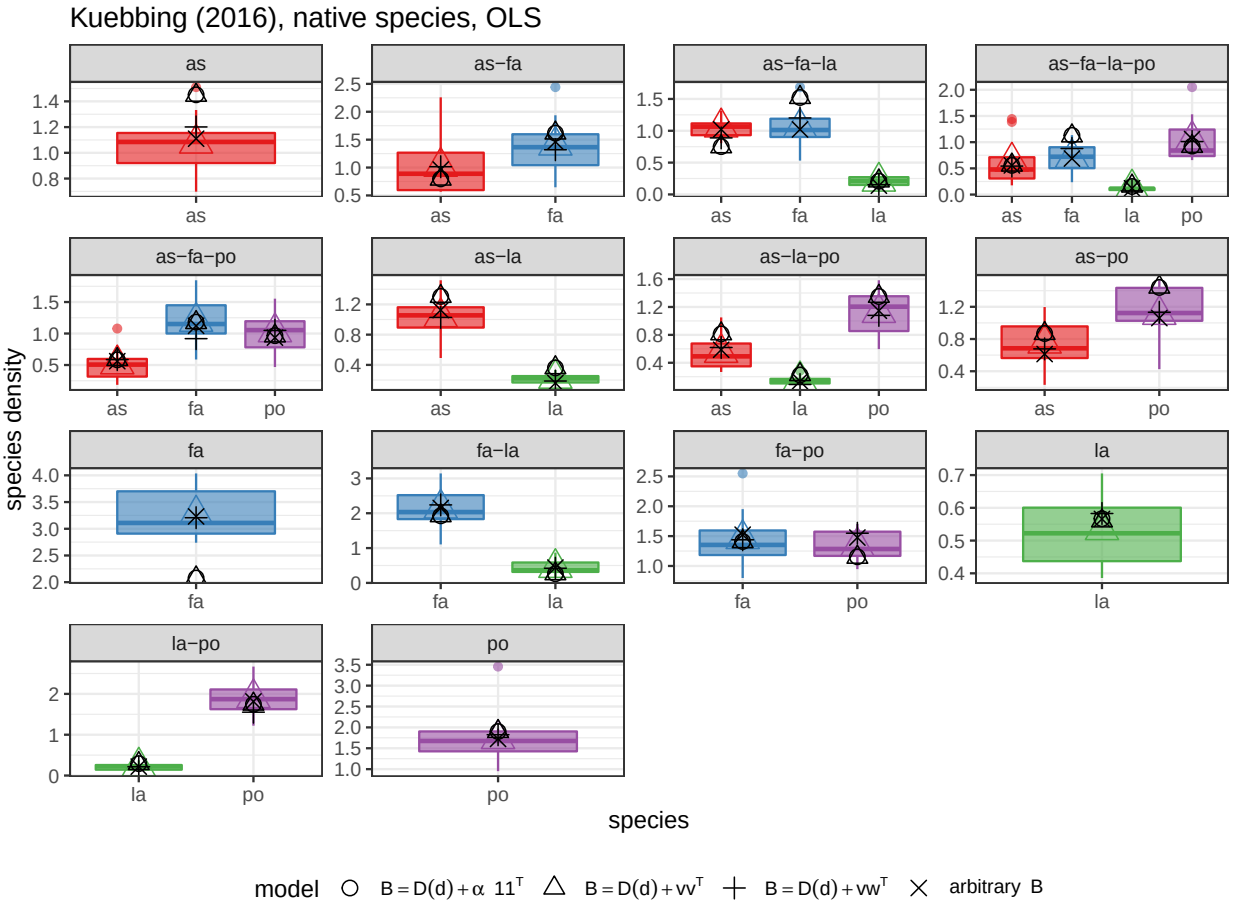

Figure S7: As Figure 2, but for the data of Kuebbing et al. (2015), native plants.

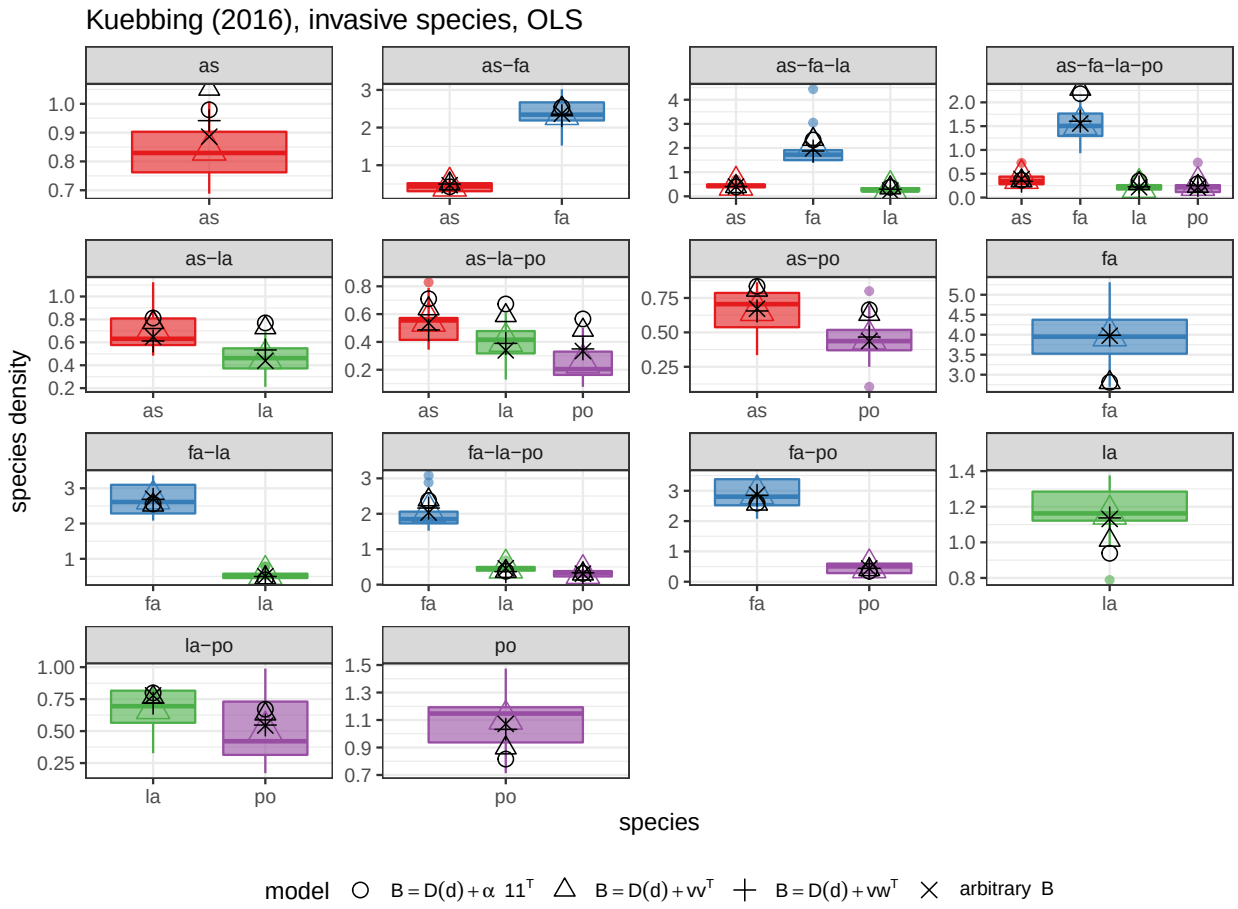

Figure S8: As Figure 2, but for the data of Kuebbing et al. (2015), non-native plants.

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).

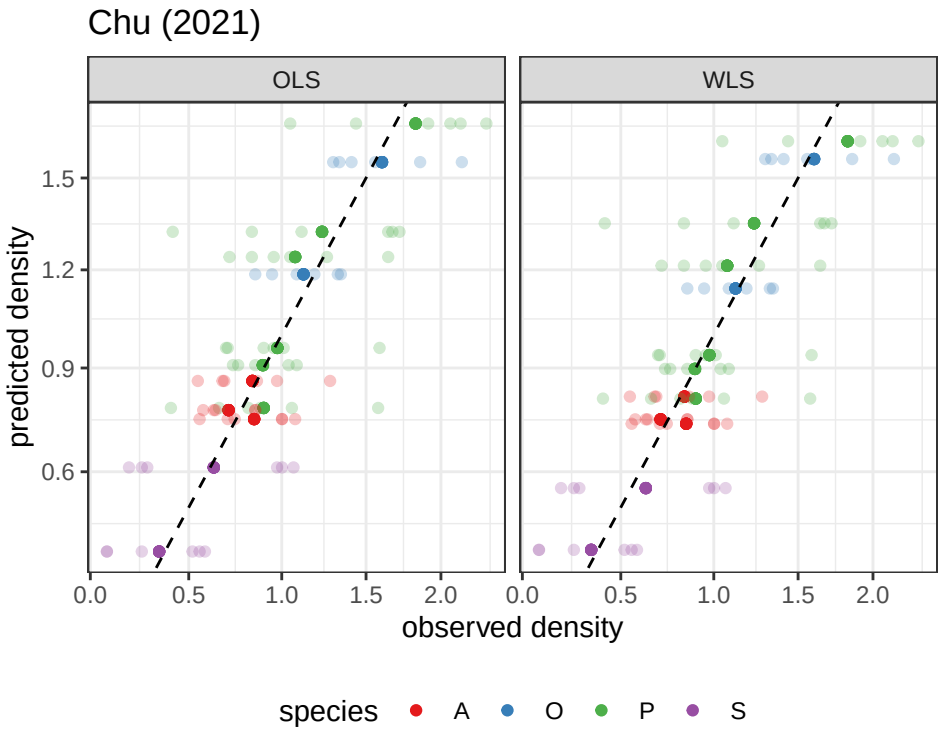

Figure S9: As Figure 5, but for the data of Chu et al., 2021.

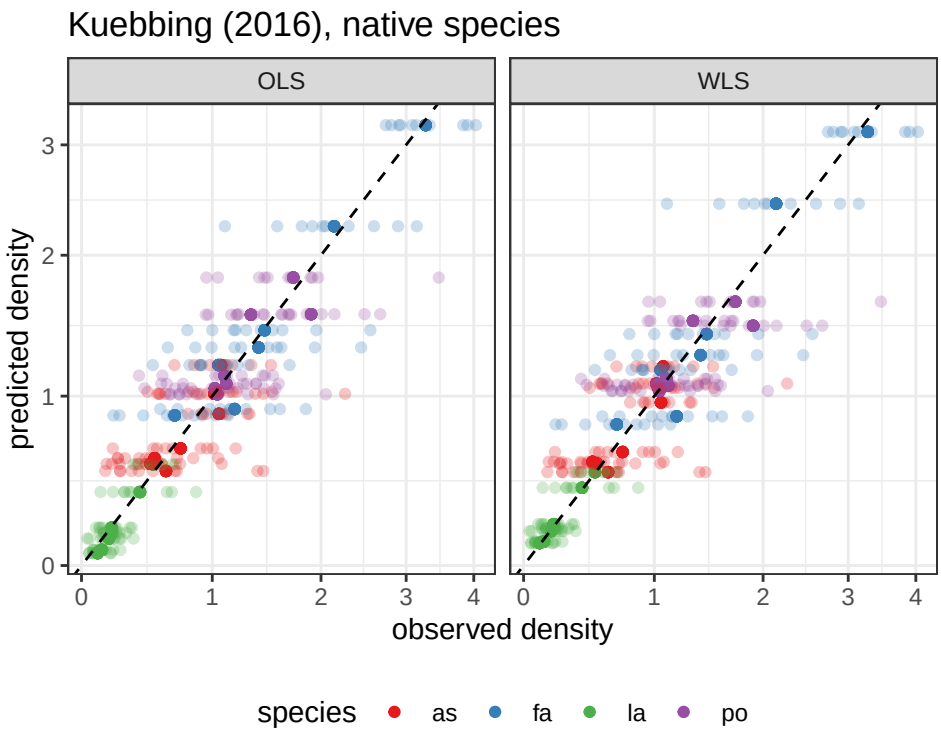

Figure S10: As Figure 5, but for the data of Kuebbing et al. (2015), native plants.

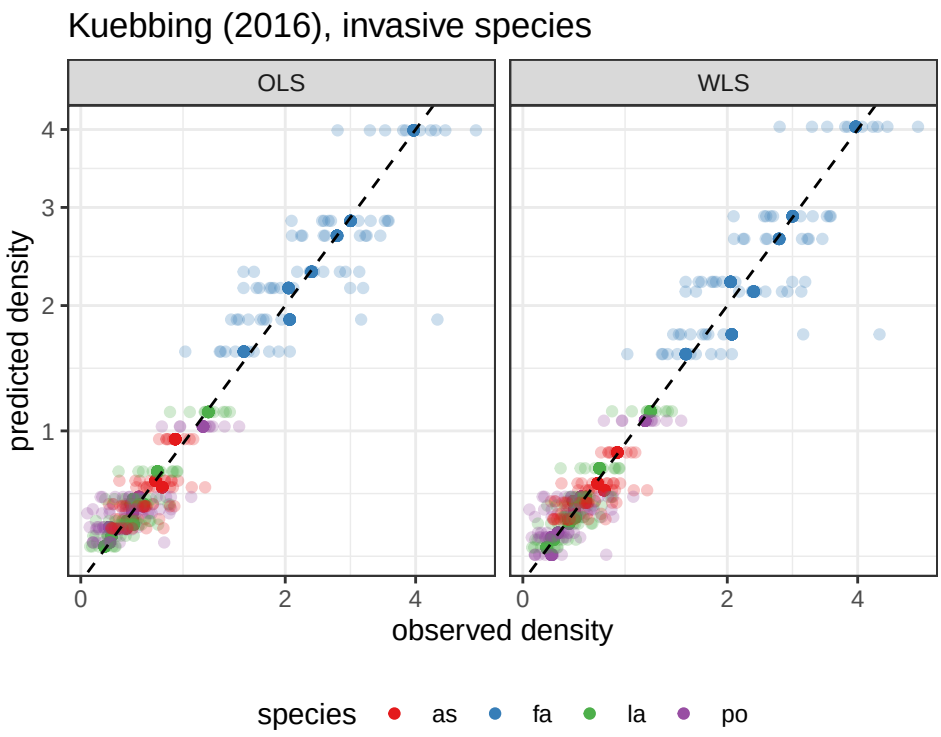

Figure S11: As Figure 5, but for the data of Kuebbing et al. (2015), non-native plants.

Out-of-fit 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-fit 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.

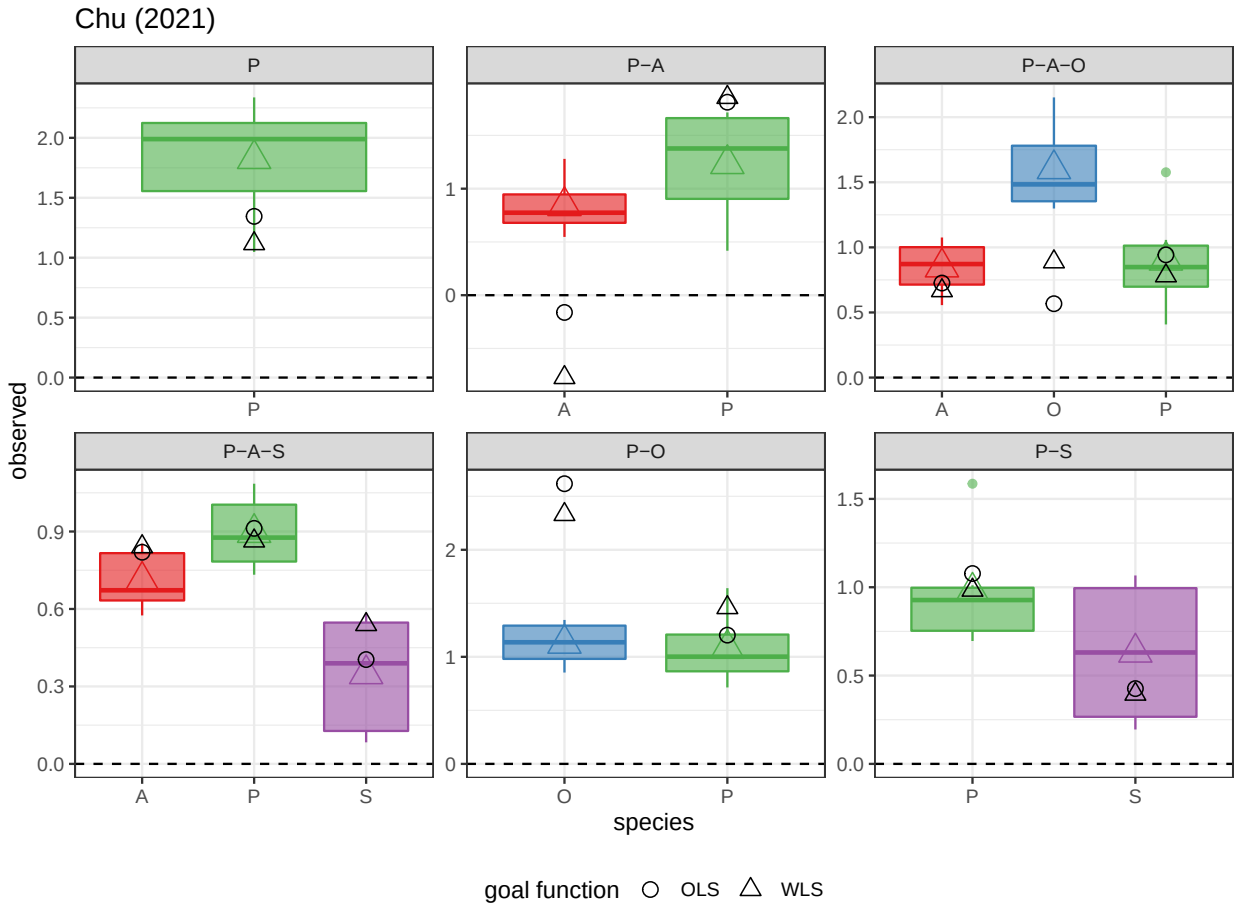

Figure S12: As Figure 6, but for the data of Chu et al., 2021. When we fit the model without the experiments in which *Pseudomonas fluorescens* and *Achromobacter* sp. are grown together, we obtain a qualitatively wrong prediction (lack of coexistence). *Ochrobactrum* sp. appears in only two communities, and when one of the communities is removed the fit is greatly reduced.

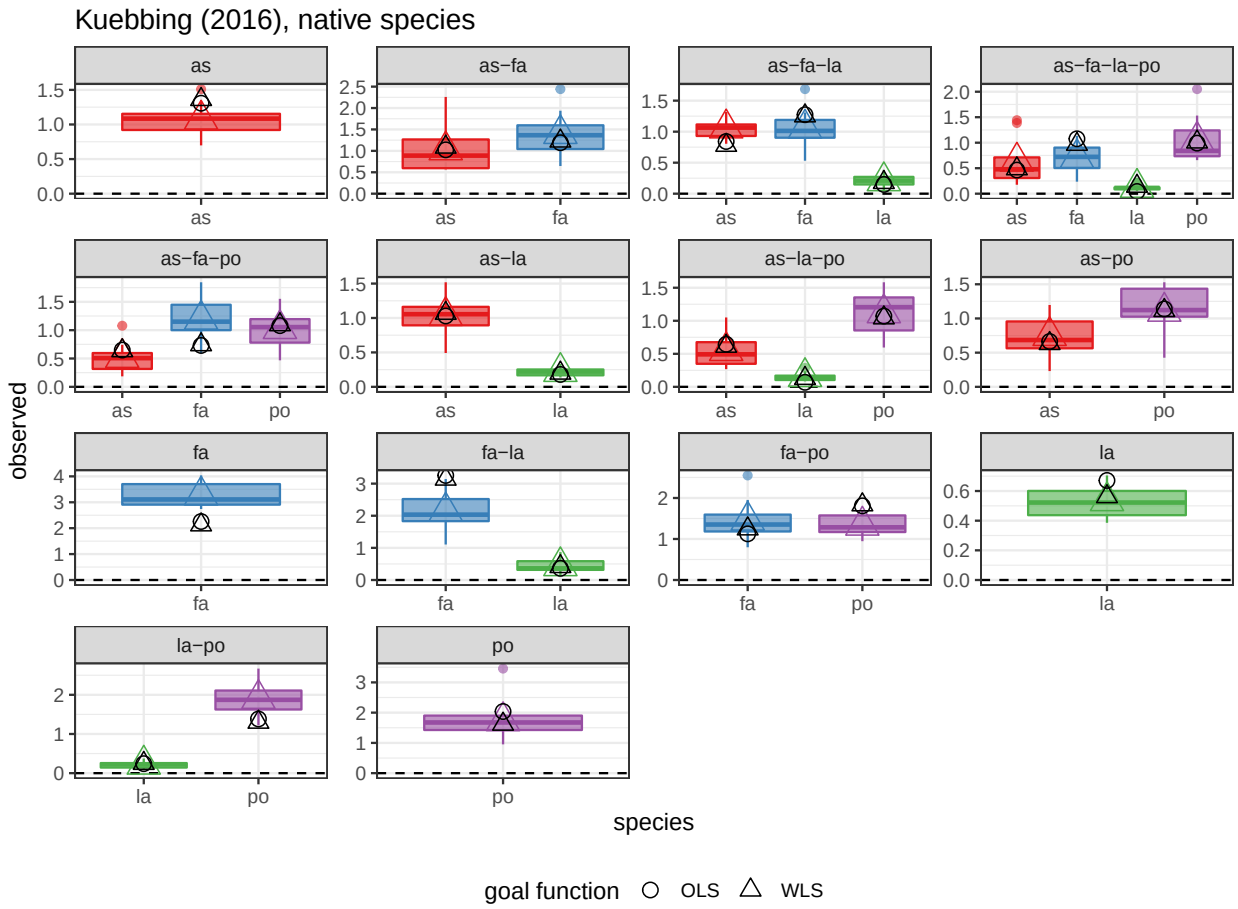

Figure S13: As Figure 6, but for the data of Kuebbing et al. (2015), native plants.

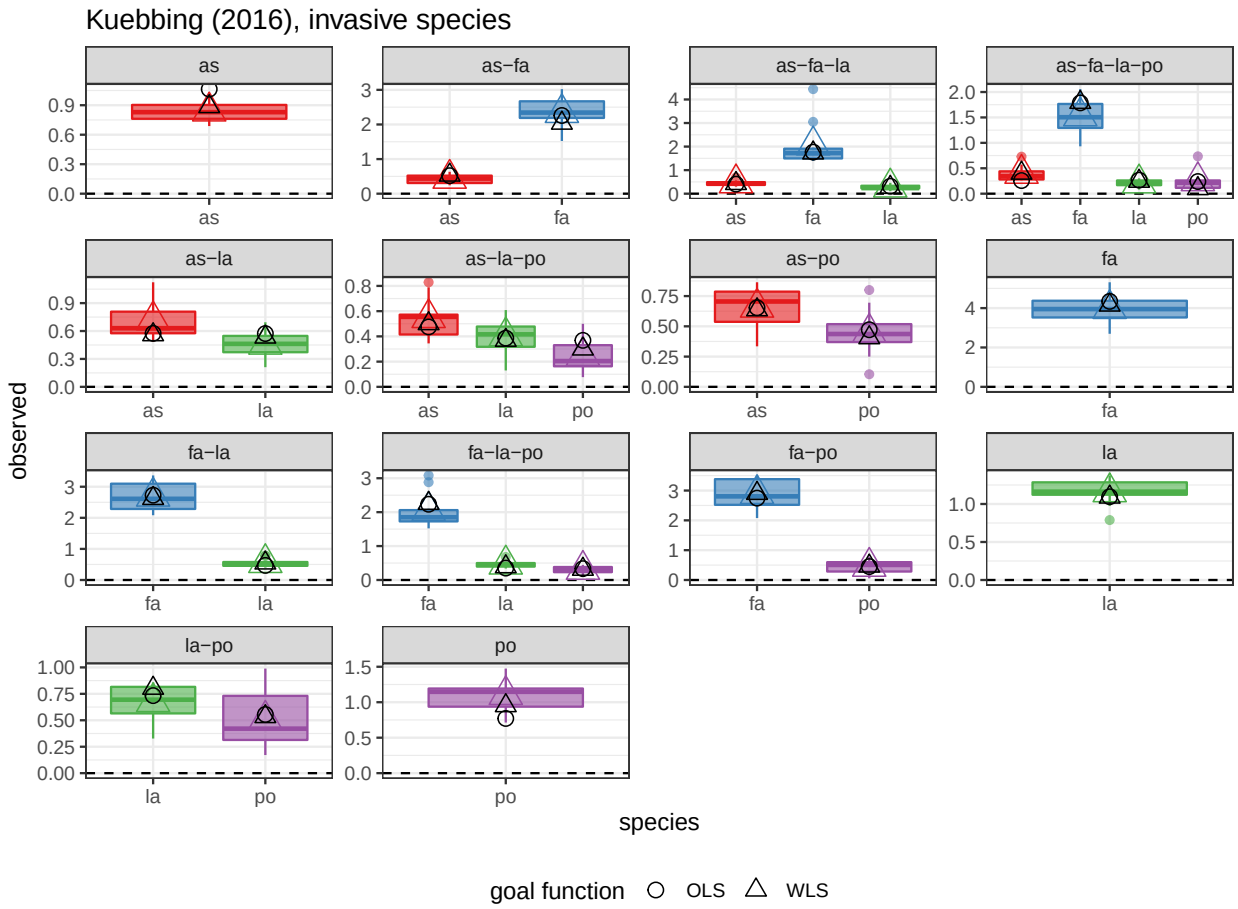

Figure S14: As Figure 6, but for the data of Kuebbing et al. (2015), non-native plants.

References

Ansari, A. F., Reddy, Y., Raut, J., & Dixit, N. M. (2021). An efficient and scalable top-down method for predicting structures of microbial communities. *Nature Computational Science*, *1*(9), 619–628.

Carpenter, S., Cottingham, K., & Stow, C. (1994). Fitting predator-prey models to time series with observation errors. *Ecology*, *75*(5), 1254–1264.

Carrara, F., Giometto, A., Seymour, M., Rinaldo, A., & Altermatt, F. (2015). Inferring species interactions in ecological communities: A comparison of methods at different levels of complexity. *Methods in Ecology and Evolution*, *6*(8), 895–906.

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.

Clark, J. S., Carpenter, S. R., Barber, M., Collins, S., Dobson, A., Foley, J. A., Lodge, D. M., Pascual, M., Pielke R., J., Pizer, W., Pringle, C., Reid, W. V., Rose, K. A., Sala, O., Schlesinger, W. H., Wall, D. H., & Wear, D. (2001). Ecological forecasts: An emerging imperative. *Science*, *293*(5530), 657–660.

De Valpine, P., & Hastings, A. (2002). Fitting population models incorporating process noise and observation error. *Ecological Monographs*, *72*(1), 57–76.

Dormann, C. F., & Roxburgh, S. H. (2005). Experimental evidence rejects pairwise modelling approach to coexistence in plant communities. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1569), 1279–1285.

Downing, A. L., Jackson, C., Plunkett, C., Ackerman Lockhart, J., Schlater, S. M., & Leibold, M. A. (2020). Temporal stability vs. Community matrix measures of stability and the role of weak interactions. *Ecology Letters*, *23*(10), 1468–1478.

Du, X., King, A. A., Woods, R. J., & Pascual, M. (2017). Evolution-informed forecasting of seasonal influenza a (H3N2). *Science Translational Medicine*, *9*(413).

Ellner, S. P., Seifu, Y., & Smith, R. H. (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology*, *83*(8), 2256–2270.

Fort, H. (2018). On predicting species yields in multispecies communities: Quantifying the accuracy of the linear Lotka-Volterra generalized model. *Ecological Modelling*, *387*, 154–162.

Friedman, J., Higgins, L. M., & Gore, J. (2017). Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology & Evolution*, *1*(5), 1–7.

Gaston, K. J., & McArdle, B. H. (1994). The temporal variability of animal abundances: Measures, methods and patterns. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *345*(1314), 335–358.

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

Gilpin, M. E. (1973). Do hares eat lynx? *The American Naturalist*, *107*(957), 727–730.

Grilli, J. (2020). Macroecological laws describe variation and diversity in microbial communities. *Nature Communications*, *11*(1), 1–11.

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

Ives, A. R., Dennis, B., Cottingham, K., & Carpenter, S. (2003). Estimating community stability and ecological interactions from time-series data. *Ecological Monographs*, *73*(2), 301–330.

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.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, *13*(6), 47–60.

Perry, J. N., Smith, R. H., Woiwod, I. P., & Morse, D. R. (2012). *Chaos in real data: The analysis of non-linear dynamics from short ecological time series* (Vol. 27). Springer Science & Business Media.

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, *4*(5), 1–17.

Powell, T. M., & Steele, J. H. (2012). *Ecological time series*. Springer Science & Business Media.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, *12*(1), 145–151.

Routledge, R. D., & Swartz, T. B. (1991). Taylor's power law re-examined. *Oikos*, 107–112.

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.

Stouffer, D. B., Godoy, O., Dalla Riva, G. V., & Mayfield, M. M. (2021). The dimensionality of plant–plant competition. *bioRxiv*. https://doi.org/10.1101/2021.11.10.467010

Sugihara, G., & May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, *344*(6268), 734–741.

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*(6106), 496–500.

Taylor, L. R. (1961). Aggregation, variance and the mean. *Nature*, *189*(4766), 732–735.

Vandermeer, J. H. (1969). The competitive structure of communities: An experimental approach with protozoa. *Ecology*, *50*(3), 362–371.

Venturelli, O. S., Carr, A. V., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., Hromada, S., Northen, T., & Arkin, A. P. (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular Systems Biology*, *14*(6), e8157.

Verhulst, P.-F. (1838). Notice sur la loi que la population poursuit dans son accroissement. *Correspondence Mathematique Et Physique*, *10*, 113–121.

Voit, E., Davis, J., & Olivenca, D. (2021). *Inference and validation of the structure of Lotka-Volterra models*.

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

Weiss-Lehman, C. P., Werner, C. M., Bowler, C. H., Hallett, L., Mayfield, M. M., Godoy, O., Aoyama, L., Barabás, G., Chu, C., Ladouceur, E., Larios, L., & Shoemaker, L. (2022). Disentangling key species interactions in diverse and heterogeneous communities: A Bayesian sparse modeling approach. *Ecology Letters*, *25*(5), 1263–1276.

Xiao, Y., Angulo, M. T., Friedman, J., Waldor, M. K., Weiss, S. T., & Liu, Y.-Y. (2017). Mapping the ecological networks of microbial communities. *Nature Communications*, *8*(1), 1–12.