INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Castro M, Vida R, Galeano J, Cuesta JA. 2025 Scarce data, noisy inferences and overfitting: the hidden flaws in ecological dynamics modelling. *J. R. Soc. Interface* **22**: 20250183.
https://doi.org/10.1098/rsif.2025.0183

Received: 5 March 2025
Accepted: 8 August 2025

Subject Category:
Life Sciences–Physics interface

Subject Areas:
biocomplexity, ecosystem, biomathematics

Keywords:
ecological modeling, Bayesian inference, statistical mechanics, parameter identifiability, generalized Lotka–Volterra model

Author for correspondence:
Mario Castro
e-mail: marioc@comillas.edu

Supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8044581.

THE ROYAL SOCIETY PUBLISHING

Scarce data, noisy inferences and overfitting: the hidden flaws in ecological dynamics modelling

Mario Castro[1], Rafael Vida[2], Javier Galeano[3] and Jose A. Cuesta[2,4]

[1]Institute for Research in Technology (IIT) and Grupo Interdisciplinar de Sistemas Complejos (GISC), Universidad Pontificia Comillas, Madrid, Madrid, Spain
[2]Universidad Carlos III de Madrid, Departamento de Matematicas and Grupo Interdisciplinar de Sistemas Complejos (GISC), Leganes, Madrid, Spain
[3]Complex Systems Group, E.T.S.I.A.A.B., Universidad Politécnica de Madrid, Madrid, Madrid, Spain
[4]Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, Zaragoza, Spain

MC, 0000-0003-3288-6144; JAC, 0000-0001-9890-9367

Metagenomic data has significantly advanced microbiome research by employing ecological models, particularly in personalized medicine. The generalized Lotka–Volterra (gLV) model is commonly used to understand microbial interactions and predict ecosystem dynamics. However, gLV models often fail to capture complex interactions, especially when data are limited or noisy. This study critically assesses the effectiveness of gLV and similar models using Bayesian inference and a model reduction method based on information theory. We found that ecological data often leads to non-interpretability and overfitting due to limited information, noisy data and parameter sloppiness. Our results highlight the need for simpler models that align with the available data and propose a distribution-based approach to better capture ecosystem diversity, stability and competition. These findings challenge current bottom-up ecological modelling practices and aim to shift the focus towards a statistical mechanics view of ecology based on distributions of parameters.

1. Introduction

Recent decades have witnessed the central role of the human gut microbiome in health and disease, catalysed by the explosion in metagenomic data [1,2] and led by initiatives such as the Human Gut Project [3]. These data have triggered activity to quantify and even predict the role of gut composition in driving so-called personalized medicine: a custom-made intervention based on the patient's specific gut ecosystem and the connection between dysbiosis and disease.

Mirroring the years succeeding the Human Genome Project [4], it has been suggested that metagenomic data is not a clear-cut *book of life*, and data should be supplemented by bottom-up approximations [5]. Modern metagenomic methods sequence all RNA in a sample, allowing for both taxonomic identification (down to species or strain level) and functional analysis (genes, pathways, metabolic potential). In principle, the collected data allows us to estimate the abundances of different species and, through correlations, provide insights into microbial relationships and the interaction between a certain environment and its role in health and disease.

In parallel with data-based analysis, researchers have increasingly turned to mathematical and computational frameworks rooted in ecology and dynamical systems theory to address this limitation. These frameworks aim to bridge the gap between descriptive data and mechanistic understanding by providing a way to model the interactions within microbial communities. Such models allow for exploring ecosystem stability and resilience and offer predictive power for how interventions or environmental changes may reshape

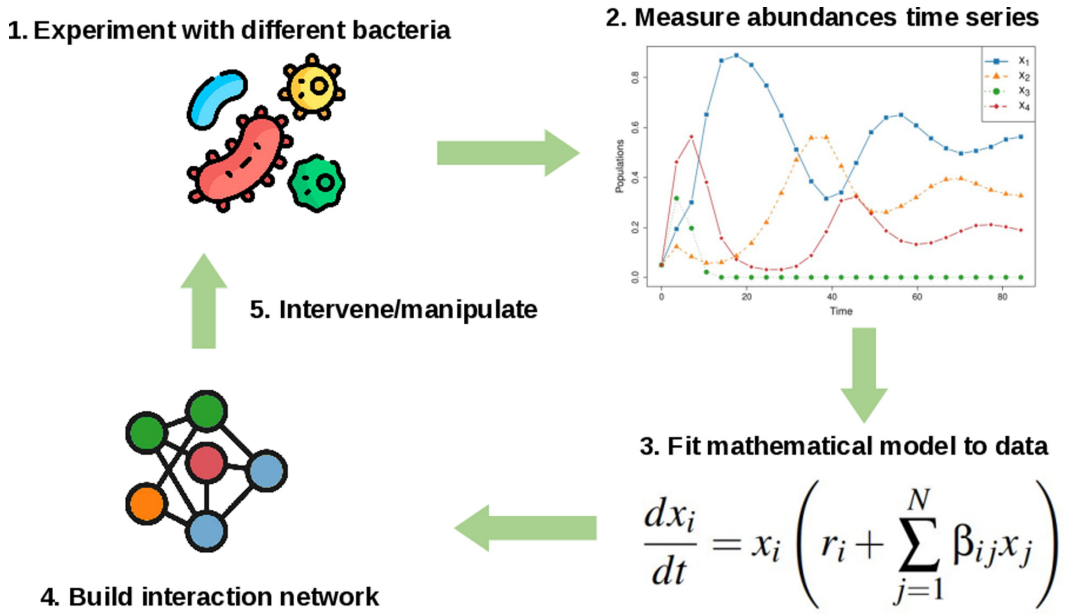

Figure 1. Sketch illustrating how experiments and models can work synergistically to personalize an individual microbiome.

microbial dynamics. These approaches seek to complement metagenomic insights with explanatory and predictive tools by focusing on the principles governing microbial coexistence and interaction.

In particular, the generalized Lotka–Volterra (gLV) model [6] and its variants (see equation (2.1) for its most basic version) borrowed from theoretical ecology may offer a link between experimental data on bacterial coexistence and microscopic interactions. Hence, the research programme proposed by several authors is straightforward: gather data, fit the equations, infer the interaction matrix and exploit pair–interaction networks [7–12] to predict the outcome of custom-made therapies aimed to re-establish a healthy microbiome [13]. This process is sketched in figure 1.

In the last few years, there has been evidence about the virtues of the gLV model (mainly in its stochastic variant) to capture universal properties of the microbiome [7,9,10,14–20]. However, despite their widespread application, recent literature has begun to cast shadows on the efficacy of these models. Critics argue that the limitations of these models are not merely a matter of practicality concerning model fitting or data availability. Instead, they point to a more profound issue, asserting that these models are inherently flawed in their ability to capture the intricate dynamics of the microbiome [21–28]. Notwithstanding, some of these concerns regarding technical limitations, particularly those related to model fitting, may be mitigated by fitting relative populations, as suggested by Remien *et al.* [29]. Other authors argue that the problem lies in the extreme simplicity of Lotka–Volterra equations, pointing to more sophisticated versions including higher-order interactions [30], or even to include the dynamics of resources that are at the centre of the interaction among species.

In this article, we use arguments from Bayesian inference and information theory to argue against the viability of this whole bottom-up enterprise. In particular, we show a mismatch between the type of data and the amount of actual information entailed in those data, on the one hand, and the level of description implicit in the gLV model and its variants to capture that information, on the other hand.

Our line of argumentation has three ingredients—that we further develop in the following sections—namely: (i) ecological data capture only a few timescales and individual steady states; (ii) the very structure of the models combined with data acquisition noise provides non-consistent interpretations of the interaction parameters; and finally (iii) gLV-type models supplemented with this sort of data are sloppy—in the sense introduced by Transtrum *et al.* [31]—so the parameters overfit the amount of existing information.

Our main message is that, as traditionally recognized in statistical mechanics and complexity science, excessive mathematical detail can provide inconsistent explanations of microscopic ecological problems when presented too confidently. We illustrate this point by combining synthetic data, where we have complete knowledge, and analysing the effect of adding noise, which would emulate how experimental data is typically obtained. Overall, we conclude that bottom-up descriptions, such as the gLV model and its variants, are often over-parametrized. As a result, the interpretation of the fitted coefficients does not align with the actual information contained in the data, claiming a radically different approach to the dynamics and function of the microbiome. Our methodological approach relies on lengthy numerical computations that cannot be automated for many species, so our argumentation is based on counter-examples rather than on mathematically rigorous proofs.

We wrap up this article arguing that the way out of this flawed enterprise is to adopt the viewpoint of statistical mechanics, where the central challenge is to discover appropriate macroscopic state variables that capture the essence of ecological dynamics and formulate parameter distributions compatible with those state variables, akin to the ensembles of statistical mechanics. The relevant macroscopic dynamics of microbiomes should follow from those ensembles in the same way that thermodynamics follows from Boltzmann's distribution. In this novel approach, finding the particulars of microbial interactions is no more interesting than determining the specific positions and momenta of particles in a gas.

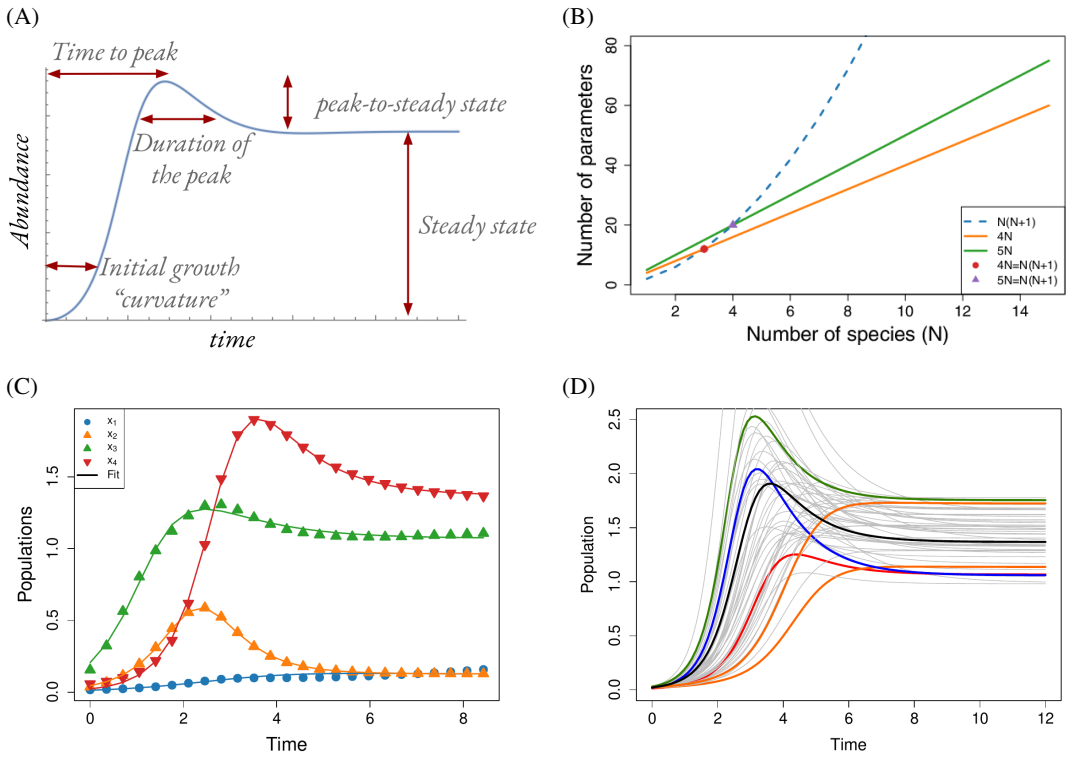

Figure 2. (A) A schematic illustration of how to estimate relevant pieces of information in a time series. As shown, a few parameters can explain this stereotypical shape even if the data has infinite precision (a continuous curve). (B) Number of parameters for the N-species gLV model (dashed blue line) and estimated number of parameters, $4N$–$5N$ (according to the rule of thumb in panel (A)). The symbols show the size above which the number of parameters, while potentially identifiable, is not explainable or meaningful from an ecological perspective. (C) An example of the time evolution of the abundance of four species according to the gLV equations (symbols). The solid lines on top of each species are fitted (equation 2.2). (D) Illustration of how equation (2.2) accommodates different shapes with just five parameters. Colours illustrate how the steady state, peak and other features can be varied independently. For instance, orange curves (sigmoids) can also be fitted with just three parameters.

2. The three limitations of population dynamics models

We will focus our discussion on the gLV model of theoretical ecology, which describes the evolution of the abundances x_i of an ecological community of $i = 1, \ldots, N$ species through the dynamical system,

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i\left(r_i + \sum_{j=1}^{N} \beta_{ij} x_j\right), \quad i = 1 \ldots N, \tag{2.1}$$

where r_i represents the (unconstrained) exponential growth rate of species i in isolation, and the term $\beta_{ij}x_j$ quantifies the effect of species j in the growth of species i. We collectively refer to the coefficients (β_{ij}) as the interaction matrix. Its diagonal elements can be regarded as (minus) the reciprocal of the so-called carrying capacities of the species in isolation.

We are aware of variants of this model including additional equations for the resources and their consumption (consumer-resource models [32]), stochastic terms accounting for demographic or environmental fluctuations [18,19], or terms with higher-order interactions [30] that aim at capturing the combined effect of several species in the growth rate of another. However, as we will argue in §3, far from fixing the problem we are about to analyse, these extensions only worsen it because they proliferate the number of parameters.

2.1. Data of species abundances contains only a few pieces of information

As mentioned in §1, we argue that the problem with equation (2.1) and its extensions to accommodate higher-order interactions or even competition for resources relies not just on the model itself but on the type of data used to infer the parameters of those models.

To illustrate this with a back-of-the-envelope calculation, consider a typical time series, as depicted in figure 2C. The shape of the curve, dissected in figure 2A, suggests that there are at most four (five if one leaves free the initial condition) pieces of independent information per population. For N species, this accounts for $5N$ independent pieces of information.

We can make this claim more quantitative by proposing the (heuristic) equation,

$$x(t) = \frac{a_1 + a_2 e^{-a_3 t}}{1 + a_4 e^{-a_5 t}} \tag{2.2}$$

,

to fit curves like that of figure 2A. Figure 2D shows how versatile this expression is to reproduce that kind of curve and, in particular, how it describes reasonably well the data of figure 2C (solid lines). Note that the curves can have a sigmoid form, a clearly

defined maximum, and different parameters lead to different timescales and peak and steady state locations. Of course, we can fit the same data using formulae with more parameters, but this will overfit the data. Different combinations of the parameters will produce similar fits. This is the fingerprint of sloppiness—parameters can be widely varied without changing the fit appreciably.

In contrast, a model like the gLV equations for N species involves—in its simplest form (2.1) with at most pairwise interactions—at least $N + N^2$ parameters. The consequence is that for $N \geq 4$, the number of parameters in the model surpasses the actual pieces of information available from the data, as shown in figure 2B, rendering the models inherently sloppy and unreliable for the purpose they are employed, as we discuss below.

To simulate synthetic ecosystems, we randomly sampled the growth rates r_i, as well as the negative self-interactions $-\beta_{ii}$ ($K_i \equiv -\beta_{ii}^{-1} > 0$ are carrying capacities) from exponential distributions (the maximum entropy distribution for positive random variables [33]) with mean 1/2. Interspecies interactions β_{ij} ($i \neq j$) are sampled from normal distributions of mean 0 and standard deviation 1 (the maximum entropy distribution for generic variables with fixed variance). This choice for the priors does not alter the conclusions, as they set a typical timescale and the order of magnitude of the abundances at steady state. After these random numbers are generated, we compute the steady state and stability and iterate until it is feasible (has positive abundances for all species) and linearly stable [34–36].

In figure 2C, we show a simulation of four species. The dots are the simulation of the gLV equation, and the solid line is the best fit to equation (2.2). Of course, we are not claiming here that equation (2.2) models the trajectories of the gLV. In fact, if the eigenvalues associated with the stability of the steady state have a non-zero imaginary part, this function would not work—it cannot reproduce oscillatory behaviours. However, we want to rationalize the idea that, disregarding the number of points measured in a curve, the shape of the curve encodes a limited number of parameters (no more than five in this case) *regardless of the time-series resolution*.

Note that our argument here is not that the parameters in equation (2.1) are not practically identifiable [37]; with enough data and precision, they are. We argue that they are not meaningful because they are over-parametrizing the data, i.e. there are many more parameters r_i, β_{ij} in the model than overall relevant pieces of information in the series. As depicted in figure 2B, we expect this problem to be more dramatic for larger ecosystems because the number of irrelevant parameters scales as $O(N^2)$.

2.2. Noisy data provide compatible but contradictory species interactions

In this section, we explore the role of noise in parameter inference from data. As Bayesian inference is computationally costly, we illustrate the effect for three species (figure 3) and relegate analogous analyses for four and five species to the electronic supplementary material. We have tested this for many random choice parameters, and the conclusions remain the same. The time series we analyse are obtained by simulating equation (2.1) and then adding a lognormal noise with different intensities, as this distribution guarantees positivity of the abundances. It is consistent with more complex stochastic versions of the gLV model [18,38]. We use different values of the lognormal standard deviation to understand the role of noise on inference.

We perform Bayesian inference through the Stan software (see electronic supplementary materials, §S1 and §S2, for the details on simulation parameters and the code, also downloadable from Zenodo [39]. This approach allows us to compute the posterior distribution for the *observed* noisy data as

$$P(\boldsymbol{\Omega}|\mathbf{y}(t), \mathbf{x}(0)) = \frac{1}{Z} P(\mathbf{y}(t)|\mathbf{x}(0), \boldsymbol{\Omega}) P(\boldsymbol{\Omega}), \tag{2.3}$$

where boldface letters denote vectors of the corresponding quantities, $\boldsymbol{\Omega} \equiv \{\mathbf{r}, \boldsymbol{\beta}, \boldsymbol{\sigma}\}$ (statistically independent),

$$Z = \int P(\mathbf{y}(t)|\mathbf{x}(0), \boldsymbol{\Omega}) P(\boldsymbol{\Omega})\, d\boldsymbol{\Omega} \tag{2.4}$$

and $d\boldsymbol{\Omega} = d^N\mathbf{r} d^{N^2}\boldsymbol{\beta} d^N\boldsymbol{\sigma}$. The first probability factor on the right-hand side of (2.3) represents a product of lognormal distributions with width parameters $\boldsymbol{\sigma}$, centred at the trajectories $\mathbf{x}(t)$. Expressed in statistical language, the generative model behind the data is

$$ny_k(t) \sim \text{lognormal}(x_k(t), \sigma_k),$$

where the abundances $x_k(t)$ are obtained by a simple numerical integration (e.g. with a fourth-order Runge-Kutta method) of equation (2.1). The last three probability factors in (2.3) are the prior distributions of the model parameters.

Using the inferred posterior distribution of the parameters, we can sample new *synthetic* trajectories $\hat{y}_k$, using the so-called *posterior predictive* distribution,

$$P(\hat{\mathbf{y}}(t)|\mathbf{y}(t)) = \int P(\hat{\mathbf{y}}(t)|\mathbf{x}(0), \boldsymbol{\Omega}) P(\boldsymbol{\Omega}|\mathbf{y}(t), \mathbf{x}(0))\, d\boldsymbol{\Omega}.$$

Figure 3 shows samples from the posterior predictive distribution for six noise levels. Note that the posterior predictive trajectories capture higher noise levels by increasing the variability for larger abundances.

To assess the results in electronic supplementary material, §S3, we collect the marginal posterior distributions for the model parameters (see electronic supplementary materials, figures S1–S6, for three species; figures S7–S12 for four species; and figures S13–S18 for five species). For the rates r_i, we compute the quantile of the actual value used in the simulations. We consider good agreement between the inferred and the actual values if, for instance, that quantile is in the range 5–95%. For the interaction parameters β_{ij}, we also show the quantiles and the probability that the predicted sign of each interaction coefficient is mistaken. This

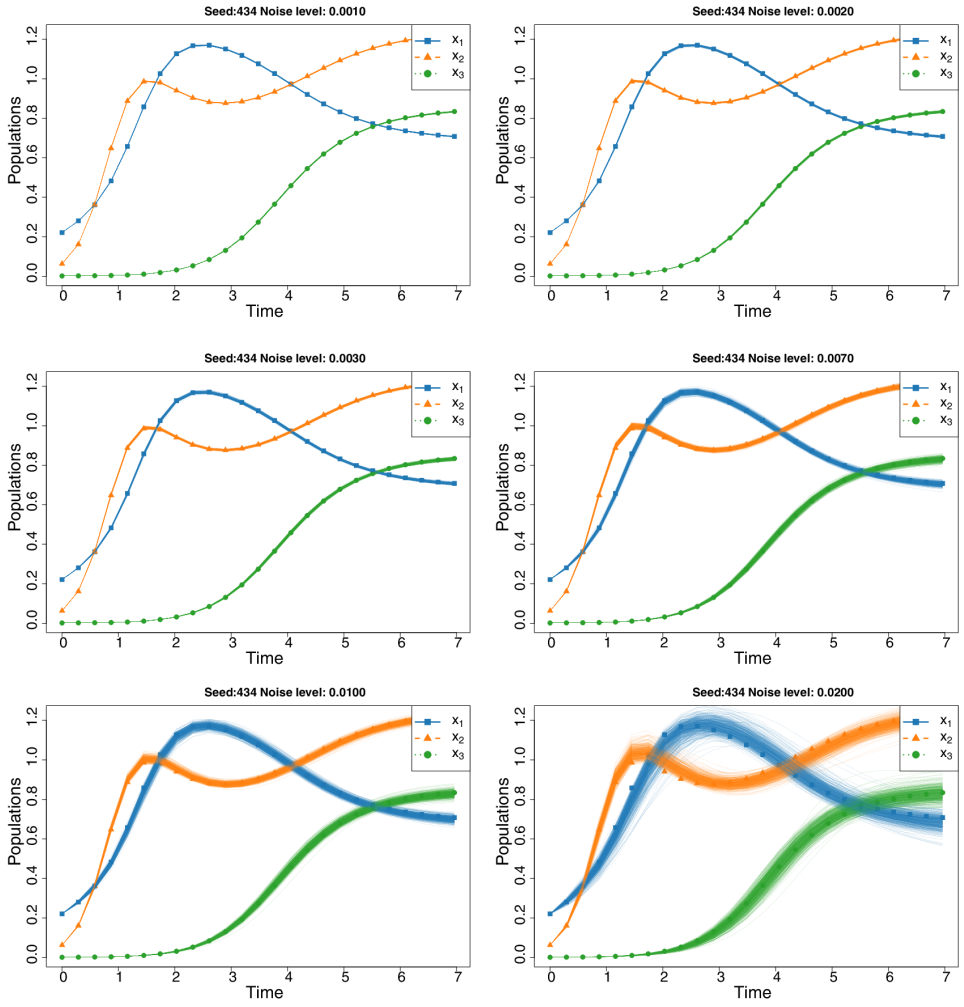

Figure 3. Posterior predictive trajectories simulated by sampling the posterior distributions (summarized in electronic supplementary material, §S3) for six different lognormal noise levels (title of each panel) added to the deterministic curves in figure 2C. The posterior predictive trajectories capture higher (lognormal) noise levels by increasing the variability for larger abundances.

means that while the posterior-predictive trajectory accurately captures the data, it is compatible with a set of parameters that wrongly infer the sign.

To summarize these results, we define three key observables. First, the number of *outliers*, which are parameters lying outside the 5th–95th percentile range. Second, the expected number of wrong signs (denoted by $\mp$ for simplicity), obtained as

$$\mathbb{E}[\mp] = \sum_{k=1}^{N^2} P_k(\mp),$$

where $P_k(\mp)$ represents the probability that the kth element of the interaction matrix (β_{ij}) is estimated with the wrong sign. This probability is obtained by integrating the posterior for that parameter over the incorrect sign interval, namely, the area of the posterior probability for that parameter corresponding to a prediction having an incorrect sign. Finally, we estimate the overall probability that the model captures at least one wrong sign as

$$P(\mp) = 1 - \prod_{k=1}^{N^2} [1 - P_k(\mp)]. \tag{2.5}$$

In table 1, we show both metrics for the parameters corresponding to figure 2C and additional examples for three and five species. Note how increasing the dimensionality produces more outliers for lower noise levels and increases the expected number of incorrect signs. This is not a failure of the method because Bayesian inference is the optimal way to quantify uncertainty in model predictions [33]. The errors, quantified by the number of outliers and equation (2.5), are not evident by inspection of the posterior predictive trajectories in figure 3 and electronic supplementary materials, figures S19 and S20. Another conclusion is that there is an ensemble of parameters that explain the data but provide opposite interpretations, as the signs of β_{ij} and β_{ji} for every pair (i, j) represent the type of interaction between species i and j.

Finally, it is worth mentioning that the posterior distributions display strong correlations among parameters, although, as explained above, all growth rates, r_i, and interaction coefficients, β_{ij}, are randomly sampled. For instance, electronic supplementary materials, figures S22 and S23, we show a pairs-plot of the marginal posterior distributions. Note how correlations occur among

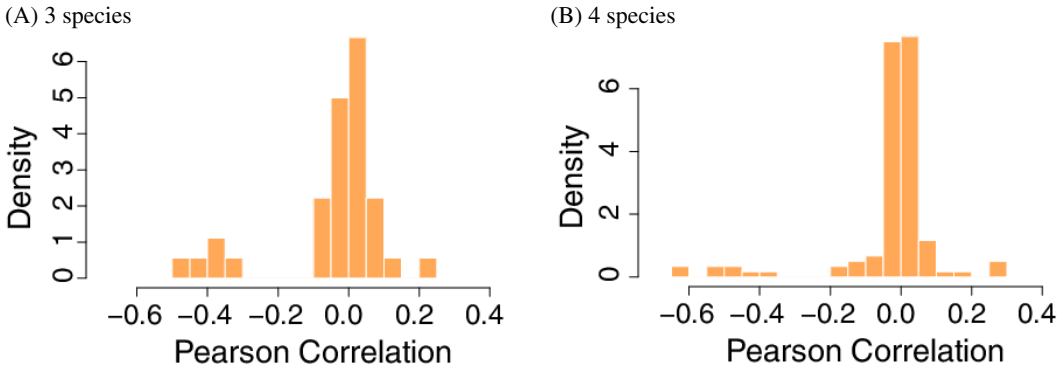

Figure 4. Normalized distributions of posterior marginal correlations between parameters for (A) the three species case (corresponding to electronic supplementary material, figure S22) and (B) the four species case (electronic supplementary material, figure S23). Note that, despite the original parameters being randomly sampled, inference mimics existing significant correlations (although most correlations are around 0). This is because there are sets of parameters that can be combined to reproduce the same reconstructed trajectory.

Table 1. Metrics quantifying how the posterior probabilities for the model parameters contain plausible sets of parameters that provide wrong predictions: $\mathbb{E}[\mp]$, the expected value of the number of wrong signs; $P(\mp)$, the probability of predicting interactions with at least one wrong sign (relative to the true underlying value in the deterministic simulation); and the number of outliers—parameters whose true value is outside the 5–95% interval of the marginal posterior probability.

noise	3 species			4 species			5 species		
	$\mathbb{E}[\mp]$	$P(\mp)$	outliers	$\mathbb{E}[\mp]$	$P(\mp)$	outliers	$\mathbb{E}[\mp]$	$P(\mp)$	outliers
0.001	0.00	0.000	1	0.05	0.050	3	5.52	1	14
0.002	0.00	0.000	1	0.37	0.341	3	6.25	1	14
0.003	0.00	0.000	0	0.73	0.624	4	6.72	1	14
0.007	0.18	0.180	2	1.33	0.882	5	7.58	1	13
0.010	0.50	0.500	3	1.82	0.938	5	7.83	1	13
0.020	1.04	0.938	7	2.54	0.981	6	8.47	1	12
0.030	1.47	0.984	7	2.76	0.986	7	9.00	1	14
0.050	1.93	0.984	9	3.01	0.990	8	9.76	1	14
0.070	2.56	0.996	9	3.39	0.992	9	10.26	1	14

different rows and columns of the interaction matrix. This suggests that aside from the concerns raised above—and summarized in table 1—there is a strong redundancy in the model.

To illustrate this redundancy, figure 4 exhibits a normalized histogram of the pairwise correlations between the posterior marginals of the parameters, which correspond to electronic supplementary materials, figures S22 and S23. Although the original parameters were sampled randomly, the inference reveals notable correlations that mimic existing significant relationships, even though most correlations cluster around zero. This occurs because the stability and feasibility conditions limit the distributions of the parameters compatible with the data, leading to effective correlations. In statistical language, the observed ecosystems represent a subsample of all the potential models *conditioned* to produce feasible stable abundances.

2.3. The generalized Lotka–Volterra model is intrinsically sloppy

In the previous two sections, we have argued that typical population dynamics curves do not seem to contain enough information to estimate all the $N(N+1)$ parameters of the N-species gLV model. So far, we have shown that the posteriors entail the optimal use of information available in the data. Still, we have not provided a quantitative measure of what we mean by *enough information* or how this can be exploited systematically. In this section, we apply the manifold boundary approximation method (MBAM) introduced in [40]. This method uses information geometry [41] to characterize the $N(N+1)$-dimensional statistical manifold defined by the model parameters. The main idea is to introduce Fisher's information matrix (FIM) as a metric tensor on this manifold and exploit its geometry to reduce the number of parameters with minimal information loss.

Generically, if our system is described by a log-likelihood $L(\mathbf{x}|\boldsymbol{\Omega}) = \log P(\mathbf{x}|\boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = (\omega_i)$ represents the parameters of the model, and if vector $\mathbf{X}$ represents the data to be fitted by the model, the elements of FIM are given by

$$I_{ij}(\boldsymbol{\Omega}) = \mathbb{E}\left[\frac{\partial}{\partial\omega_i}L(\mathbf{X}|\boldsymbol{\Omega})\frac{\partial L}{\partial\omega_j}L(\mathbf{X}|\boldsymbol{\Omega})\right]. \tag{2.6}$$

The eigenvalues of this matrix quantify the change in the amount of information produced by a variation of the model parameters along the direction of the corresponding eigenvector. Thus, small eigenvalues characterize sloppy directions along which

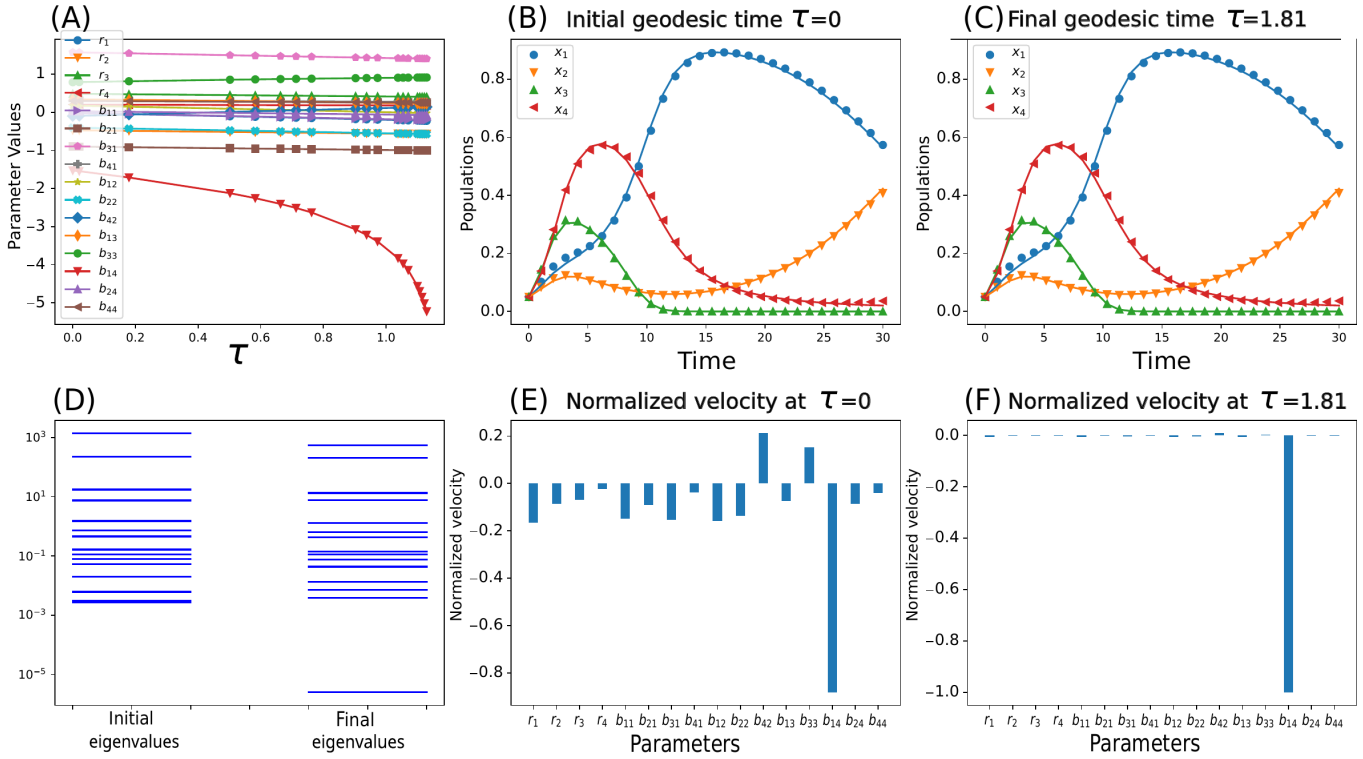

Figure 5. A combined plot to understand the MBAM. (A) Trajectories of the parameters along the geodesics (τ being the intrinsic geodesic time, i.e. the parameter in which the geodesic curve is described). The other two panels in the first row show the data (symbols) and the model (solid lines) for the values of the parameters at (B) the beginning and (C) the end of the numerical integration interval. (D) Spectrum of FIM for the parameters at the beginning (left) and end (right) of the integration interval. The smallest eigenvalue corresponds to the sloppiest combination of parameters. Note that how the sloppiest eigenvalue decreases upon reaching the manifold boundary through the geodesic. As parameters get removed, the spectrum shrinks within a few orders of magnitude. This is the signature of *stiff* (non-sloppy) models. The criteria for determining when a manifold boundary is found are based on the rate of change (velocity) of the parameters along the geodesics. Panels (E) and (F) represent the velocity of the parameters at the beginning (E) and the end (F) of the geodesic. The fact that just one parameter has a negative velocity at the end, while the rest practically do not change means that the logarithm of this parameter tends to $-\infty$, and thus it can be dropped.

the model is still accurate enough, given the empirical data, but it can be driven to simplified sub-models containing as much information and preserving the ability to make accurate predictions with fewer effective parameters [42]. The idea of MBAM is to move along the geodesics defined by the metric tensor I_{ij} towards the boundary of the manifold Iij. The rationale for this is that geodesics are the straight lines of a curved geometry; hence, they represent the shortest paths to the boundary. For further details, see [40].

We apply the MBAM systematically to another set of parameters of the gLV model, using a deterministic simulation with no added noise. This represents the (highly) ideal situation in which we possess perfect model knowledge. In figure 5, we illustrate one (arbitrary) step of the MBAM. In that figure, we show (A) the values of the logarithm of each parameter along the geodesic as a function of the geodesic time τ (geodesics are parametric curves with parameter τ); the model fitted (B) at the beginning ($\tau = 0$) and (C) at the end ($\tau > 0$) of the integration interval for the geodesic; (D) the eigenvalues of the FIM at the beginning and end of that integration; and the velocity (rate of change with τ) of each parameter along the geodesic (E) at the beginning and (F) at the end of the integration interval. In the particular case illustrated by the figure, $\log b_{14} \to -\infty$ as $\tau \to \infty$, meaning that this parameter is effectively 0.

We computed 11 steps of MBAM. In the first six of them, we can eliminate six parameters safely (β_{32}, β_{23}, β_{34}, β_{43}, β_{14} and β_{42}). In the next three steps, we can eliminate three more parameters (β_{22}, β_{33} and β_{21}) if we are willing to sacrifice the goodness of fit of species x_2. Finally, the last two steps illustrate how the fitting errors rise dramatically in a much too simplified model. Figure 6A shows the evolution of a measure of the fitting error (the sum of squared residuals) after each MBAM step. Note in figure 6B how, after removing the seventh parameter, the trajectory of x_2 is no longer well captured. The inset in figure 6A shows how the FIM spectrum shrinks by reducing the complexity of the model, corresponding to a less sloppy model after each step. Note also that the spectrum of the FIM for the final reduced model (after six steps) now spans just a few orders of magnitude, meaning that the model is no longer sloppy.

Another suggestive conclusion from this systematic reduction is that MBAM has eliminated all the parameters by making them 0. Hence, the interaction matrix of the reduced model is, after reduction, sparser. This sparsity has been recently pointed out as a signature of microbial ecosystems [18]. Note that removing parameters by making them 0 is not the only outcome of the MBAM (which also accommodates simultaneously pushing parameters to infinity). Still, we speculate that it might be a consequence of the mathematical structure of the gLV model.

In electronic supplementary material, figure S24, we show an example resembling the typical cycles in the original predator–prey formulation of the Lotka–Volterra model. As in the previous example, all the MBAM steps eliminate interaction coefficients.

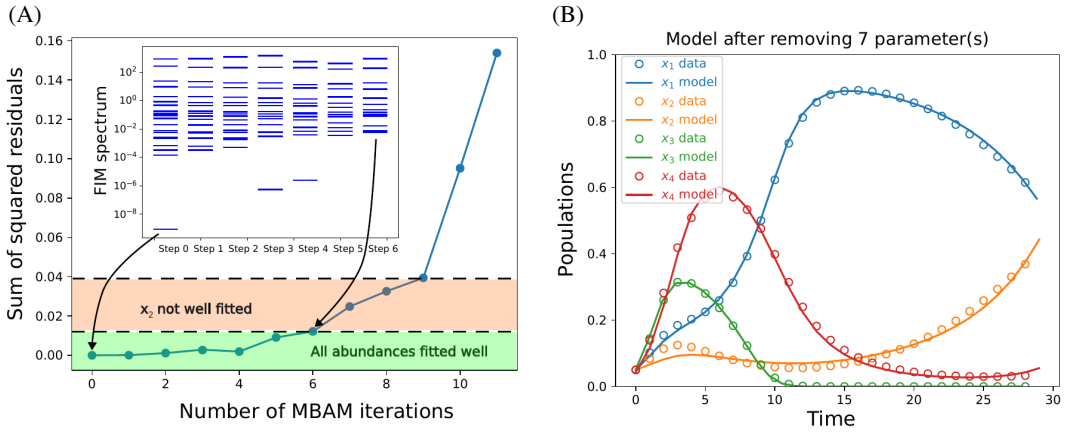

Figure 6. (A) Sum of squared residuals (difference between original data and model prediction) for model parameters in figure 5. Inset: MBAM parameter elimination shrinks the FIM spectrum, making the resulting model less sloppy according to [31]. Note that how the error jumps at seven reduced parameters corresponding to the inability of the reduced model to explain x_2, as shown in (B).

Note that when using MBAM [40], the simplifications we achieve are contingent on the data we are fitting. Having more *relevant* information makes further model parameters relevant, and the oversimplified model here obtained may be too simple. By relevant information, we do not mean necessarily longer times (as shown in electronic supplementary material, §S8) but truly new information not present in the previous data (typically, new experiments performed with different conditions or measuring additional observables).

By the same principle, it is worth emphasizing that all these numerical experiments are made with noiseless, synthetically generated data so that they would correspond to experimental data measured with extreme accuracy. Combined with the discussion about practical identifiability emerging from the Bayesian analysis in §2.2, we expect the reduction to be even more dramatic in a real scenario. Moreover, taking into account that we have more sampling points in our time series than typically available in experiments, this expectation is reinforced.

3. Discussion and conclusions

Our overarching argument throughout this article is simple but has deep implications: population dynamics time series do not contain enough information to estimate and confidently interpret the interactions between species in an ecosystem. Typically, the time evolution is governed by a few timescales and typical abundance sizes (what we summarize in figure 2A as *pieces of information*). This hand-waving assessment—which we made more quantitative through equation (2.2)—has two implications: one regarding the extent to which we can infer interactions from the data (using Bayesian inference, the optimal way to extract information from data), and another one using the principles of information geometry (which provides a systematic way to remove non-informative parameters).

Regarding the first, we showed (through the marginal posterior distributions for the parameters) that, although the problem is in principle structurally identifiable [29], in practice, there is a significant probability that not only the value but even the sign of a given interaction parameter β_{ij} is wrong—unless the priors are so close to the original parameters that the posteriors are biased beforehand (electronic supplementary material, §S7). This has important biological implications because we can incorrectly predict the ecological nature of species interactions (competition, commensalism, mutualism, parasitism, …). Also, the posterior densities show that the real parameter used in the simulation cannot be practically identified due to noise. Furthermore, the larger the population is, the more uncertain determining the underlying interactions (even their sign) becomes.

On the other hand, MBAM illustrates how the gLV model is over-parametrized, thus substantiating our graphical argument in figure 2A. The first conclusion is that the removed parameters are, mainly, coefficients of the matrix β_{ij} set to 0. This aligns with the idea, recently put forward in a study that infers interaction matrices of microbial communities [18], that the interaction matrices β_{ij} of these systems tend to be sparse. Of course, all our analyses have been performed on relatively smooth time series. Whether the argument still carries on to more general cases (with many more species) involving complex oscillations or chaotic behaviour needs further study.

Another implication of our results is that, as gLV (or more sophisticated models) are over-parametrized, they are expected to be more prone to overfitting and to generalize poorly to future events, with this effect being larger for a larger number of species. We illustrate this in electronic supplementary material, §S8. Interestingly, extending the model reduction up to $t = 6$ improves the fitting, but the inferred (and eliminated) parameters differ, as shown in electronic supplementary material, figure S27.

One should bear in mind that the problems revealed in inferring the interactions of the gLV model are not exclusive to this system but apply to any sufficiently complex dynamics [31,43]. Thus, our discussion here is completely general and pertains to the science of complex systems at large.

Any critical reader might suggest potential solutions to this pessimistic scenario. In the problem under discussion, one possible solution would be to extend the gLV model to include additional mechanisms—such as competition/production for metabolites

or other resources [44]. This might indeed improve our understanding of microbial interactions, but those models have a complexity cost that should be supplemented with relevant empirical data. Otherwise, our main argument about the relevant *pieces of information* in the data will imply that the new parameters related to the species-resource interplay will suffer from the same problems as those described in this article.

Alternatively, we can employ stochastic versions of the gLV model [18,38]. These generalizations offer additional insights by accommodating variability in the data as random fluctuations. This helps reduce overfitting and suggests focusing on the distributions of interactions. However, implementing this method is not straightforward, as the stability and observability of an ecosystem already influence the underlying correlations within the data (as illustrated in figure 4).

Instead of approaching the problem by trying to amend the model, we advocate for a paradigm shift in ecological research, moving the focus from traditional dynamical systems to a framework grounded in the principles of statistical mechanics. The central challenge lies in defining the appropriate state variables that capture the essence of ecological dynamics and facilitate specific strategies for ecosystem manipulation towards desired outcomes. We introduce an ensemble approach to the analysis of ecological data, positing that the joint posterior distribution, $P(\boldsymbol{\Omega}|\mathbf{y}(t), \mathbf{x}(0))$, extends beyond mere statistical inference. Specifically, instead of referring to a parameter as the unique numerical value that better fits the data, this approach advocates for assigning a whole ensemble of parameters compatible with the data, so the interpretation of the model parameters is more robust, on one hand; and contains additional information about the uncertainty associated to that parameter, on the other hand.

In conclusion, this framework encompasses a multitude of potential scenarios that can explain observed data, thereby maximizing the informational yield from experimental findings. Accordingly, it is essential to develop robust macroscopic state variables that encapsulate critical ecological phenomena such as diversity, stability, symbiosis and interspecies competition, all of which can be derived from these joint distributions. As we look to the future, we recognize that this presents an exciting and necessary research agenda that promises to deepen our understanding of ecological systems and enhance our ability to manage and sustain them effectively.

Ethics. This article does not present research with ethical considerations.

Data accessibility. Data are generated automatically from the code available at [39].

Supplementary material is available online [45].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. M.C.: conceptualization, formal analysis, funding acquisition, investigation, methodology, software, visualization, writing—original draft, writing—review and editing; R.V.: conceptualization, formal analysis, investigation, methodology, writing—review and editing; J.G.: conceptualization, formal analysis, investigation, methodology, writing—review and editing; J.A.C.: conceptualization, formal analysis, investigation, methodology, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by grants PID2021-122711NB-C21, PID2022-141802NB-I00(BASIC), PID2021-128966NB-I00 and PID2022-140217NB-I00, funded by MICIN/AEI/10.13039/501100011033 and 'ERDF/EU A way of making Europe'.

Acknowledgements. We thank Susanna Manrubia and Saúl Ares for fruitful discussions about minimal models and parameter interpretation in sloppy systems, and Aniello Lampo for critically reading the manuscript and useful suggestions.

References

1. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. 1998 Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol*. **5**, R245–R249. (doi:10.1016/s1074-5521(98)90108-9)
2. Steele HL, Streit WR. 2005 Metagenomics: advances in ecology and biotechnology. *FEMS Microbiol. Lett*. **247**, 105–111. (doi:10.1016/j.femsle.2005.05.011)
3. The Human Microbiome Project Consortium. 2012 A framework for human microbiome research. *Nature* **486**, 215–221. (doi:10.1038/nature11209)
4. Venter C *et al*. 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921. (doi:10.1038/35057062)
5. Ball P. 2023 *How life works: a user's guide to the new biology*. Chicago, IL, USA: University of Chicago Press.
6. Venturelli OS, Carr AV, Fisher G, Hsu RH, Lau R, Bowen BP, Hromada S, Northen T, Arkin AP. 2018 Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol*. **14**, e8157. (doi:10.15252/msb.20178157)
7. Faust K, Raes J. 2012 Microbial interactions: from networks to models. *Nat. Rev. Microbiol*. **10**, 538–550. (doi:10.1038/nrmicro2832)
8. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, Sander C, Xavier JB. 2013 Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol*. **9**, 31–36. (doi:10.1371/journal.pcbi.1003388)
9. Bucci V, Xavier JB. 2014 Towards predictive models of the human gut microbiome. *J. Mol. Biol*. **426**, 3907–3916. (doi:10.1016/j.jmb.2014.03.017)
10. Jones EW, Carlson JM. 2019 Steady-state reduction of generalized Lotka-Volterra systems in the microbiome. *Phys. Rev. E* **99**, 032403. (doi:10.1103/physreve.99.032403)
11. Matchado MS, Lauber M, Reitmeier S, Kacprowski T, Baumbach J, Haller D, List M. 2021 Network analysis methods for studying microbial communities: a mini review. *Comput. Struct. Biotechnol. J*. **19**, 2687–2698. (doi:10.1016/j.csbj.2021.05.001)
12. Coyte KZ, Rao C, Rakoff-Nahoum S, Foster KR. 2021 Ecological rules for the assembly of microbiome communities. *PLoS Biol*. **19**, e3001116. (doi:10.1371/journal.pbio.3001116)
13. Abreu C, Ortiz Lopez A, Gore J. 2018 Pairing off: a bottom-up approach to the human gut microbiome. *Mol. Syst. Biol*. **14**, e8425. (doi:10.15252/msb.20188425)
14. Coyte KZ, Schluter J, Foster KR. 2015 The ecology of the microbiome: networks, competition, and stability. *Science* **350**, 663–666. (doi:10.1126/science.aad2602)
15. Grilli J. 2020 Macroecological laws describe variation and diversity in microbial communities. *Nat. Commun*. **11**, 4743. (doi:10.1038/s41467-020-18529-y)
16. Descheemaeker L, Grilli J, de Buyl S. 2021 Heavy-tailed abundance distributions from stochastic Lotka-Volterra models. *Phys. Rev. E* **104**, 034404. (doi:10.1103/physreve.104.034404)
17. Hu J, Amor DR, Barbier M, Bunin G, Gore J. 2022 Emergent phases of ecological diversity and dynamics mapped in microcosms. *Science* **378**, 85–89. (doi:10.1126/science.abm7841)
18. Camacho-Mateu J, Lampo A, Sireci M, Muñoz MA, Cuesta JA. 2024 Sparse species interactions reproduce abundance correlation patterns in microbial communities. *Proc. Natl Acad. Sci. USA* **121**, e2309575121. (doi:10.1073/pnas.2309575121)

19. Camacho-Mateu J, Lampo A, Ares S, Cuesta JA. 2024 Nonequilibrium microbial dynamics unveil a new macroecological pattern beyond Taylor's law. *Phys. Rev. E* **110**, 044402. (doi:10.1103/PhysRevE.110.044402)
20. Mallmin E, Traulsen A, De Monte SD. 2024 Chaotic turnover of rare and abundant species in a strongly interacting model community. *Proc. Natl Acad. Sci. USA* **121**, e2312822121. (doi:10.1073/pnas.2312822121)
21. Balsa-Canto E, Alonso-del-Real J, Querol A. 2020 Mixed growth curve data do not suffice to fully characterize the dynamics of mixed cultures. *Proc. Natl Acad. Sci. USA* **117**, 811–813. (doi:10.1073/pnas.1916774117)
22. Ji B, Herrgård MJ, Nielsen J. 2021 Microbial community dynamics revisited. *Nat. Comput. Sci*. **1**, 640–641. (doi:10.1038/s43588-021-00144-6)
23. Momeni B, Xie L, Shou W. 2017 Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *eLife* **6**, e25051. (doi:10.7554/elife.25051)
24. Ram Y, Obolski U, Feldman MW, Berman J, Hadany L. 2020 Reply to Balsa-Canto et al.: growth models are applicable to growth data, not to stationary-phase data. *Proc. Natl Acad. Sci. USA* **117**, 814–815. (doi:10.1073/pnas.1917758117)
25. Pinto S, Benincà E, van Nes EH, Scheffer M, Bogaards JA. 2022 Species abundance correlations carry limited information about microbial network interactions. *PLoS Comput. Biol*. **18**, e1010491. (doi:10.1371/journal.pcbi.1010491)
26. Zapién-Campos R, Bansept F, Traulsen A. 2023 Inferring interactions from microbiome data. *bioRxiv*. (doi:10.1101/2023.03.30.534939)
27. Picot A, Shibasaki S, Meacock OJ, Mitri S. 2023 Microbial interactions in theory and practice: when are measurements compatible with models? *Curr. Opin. Microbiol*. **75**, 102354. (doi:10.1016/j.mib.2023.102354)
28. Lubiana Botelho L, Jeynes-Smith C, Vollert SA, Bode M. 2025 Calibrated ecosystem models cannot predict the consequences of conservation management decisions. *Ecol. Lett*. **28**, e70034. (doi:10.1111/ele.70034)
29. Remien CH, Eckwright MJ, Ridenhour BJ. 2021 Structural identifiability of the generalized Lotka–Volterra model for microbiome studies. *R. Soc. Open Sci*. **8**, 201378. (doi:10.1098/rsos.201378)
30. Bairey E, Kelsic ED, Kishony R. 2016 High-order species interactions shape ecosystem diversity. *Nat. Commun*. **7**, 12285. (doi:10.1038/ncomms12285)
31. Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. 2015 Perspective: sloppiness and emergent theories in physics, biology, and beyond. *J. Chem. Phys*. **143**, 010901. (doi:10.1063/1.4923066)
32. MacArthur R. 1970 Species packing and competitive equilibrium for many species. *Theor. Popul. Biol*. **1**, 1–11. (doi:10.1016/0040-5809(70)90039-0)
33. Jaynes ET. 2003 *Probability theory: the logic of science*. New York, NY: Cambridge University Press.
34. May RM. 1972 Will a large complex system be stable? *Nature* **238**, 413–414. (doi:10.1038/238413a0)
35. Allesina S, Tang S. 2012 Stability criteria for complex ecosystems. *Nature* **483**, 205–208. (doi:10.1038/nature10832)
36. Liu X, Constable GWA, Pitchford JW. 2023 Feasibility and stability in large Lotka Volterra systems with interaction structure. *Phys. Rev. E* **107**, 054301. (doi:10.1103/physreve.107.054301)
37. Castro M, de Boer RJ. 2020 Testing structural identifiability by a simple scaling method. *PLoS Comput. Biol*. **16**, e1008248. (doi:10.1371/journal.pcbi.1008248)
38. Descheemaeker L, de Buyl S. 2020 Stochastic logistic models reproduce experimental time series of microbial communities. *eLife* **9**, e55650. (doi:10.7554/elife.55650)
39. Universidad Pontificia Comillas. 2025 Code for manuscript: 'Scarce data, noisy inferences, and overfitting: the hidden flaws in ecological dynamics modelling'. Zenodo. (doi:10.5281/zenodo.16747311)
40. Transtrum MK, Qiu P. 2014 Model reduction by manifold boundaries. *Phys. Rev. Lett*. **113**, 098701. (doi:10.1103/physrevlett.113.098701)
41. Amari S. 2021 Information geometry. *Int. Stat. Rev*. **89**, 250–273. (doi:10.1111/insr.12464)
42. Brown KS, Sethna JP. 2003 Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E* **68**, 021904. (doi:10.1103/physreve.68.021904)
43. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. 2007 Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol*. **3**, e189. (doi:10.1371/journal.pcbi.0030189)
44. Skwara A, Gowda K, Yousef M, Diaz-Colunga J, Raman AS, Sanchez A, Tikhonov M, Kuehn S. 2023 Statistically learning the functional landscape of microbial communities. *Nat. Ecol. Evol*. **7**, 1823–1833. (doi:10.1038/s41559-023-02197-4)
45. Castro M, Vida R, Galeano J, Cuesta JA. 2025 Supplementary material from: Scarce data, noisy inferences and overfitting: the hidden flaws in ecological dynamics modelling. Figshare. (doi:10.6084/m9.figshare.c.8044581)

Journal of the Royal Society Interface

Supplementary Information:
Scarce Data, Noisy Inferences, and Overfitting: The Hidden Flaws in Ecological Dynamics Modelling

Mario Castro[1,2], Rafael Vida[1,2,4], Javier Galeano[2,4], José A. Cuesta[2,3,5]

[1]Institute for Research in Technology (IIT), Universidad Pontificia Comillas, Madrid, Spain
[2]Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain
[3]Universidad Carlos III de Madrid, Departamento de Matemáticas, Leganés, Spain
[4]Complex System Group, Universidad Politécnica de Madrid, Madrid, Spain
[5]Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, Zaragoza, Spain

S1. Parameters used in the simulation

In Table S1, we collect the parameters used to simulate Eq. (S1):

$$\frac{dx_i}{dt} = x_i \left(r_i + \sum_{j=1}^{N} \beta_{ij} x_j \right), \quad i = 1 \dots N, \tag{S1}$$

and the main figure corresponding to those parameters. In the case of the simulations used to apply Bayesian inference, the code is freely available at https://zenodo.org/records/16747311, so we indicate the random seed used in the simulations (for reproducibility). Recall that

- x_i represents the population size of species i,
- r_i is the intrinsic growth rate of species i,
- β_{ij} denotes the interaction coefficient between species i and j.

Table S1. Parameters used in this work. The first column points to the main figure related to those parameters. The seed used in the simulation code is shown in the case of those used to make Bayesian inference. The numbers have been rounded to the second significant figure for compactness.

Main figure	$x_i(0)$	r_i	β_{ij}	Reference
3 (main text)	$\begin{bmatrix} 0.22 \\ 0.063 \\ 0.0014 \end{bmatrix}$	$\begin{bmatrix} 1.1 \\ 3.9 \\ 0.78 \end{bmatrix}$	$\begin{bmatrix} -1.8 & 1.2 & -1.7 \\ -1.5 & -2.6 & 0.35 \\ -0.12 & 1.3 & -2.7 \end{bmatrix}$	This work (seed 434)
5 (main text)	$\begin{bmatrix} 10 \\ 10 \\ 10 \\ 10 \end{bmatrix}$	$\begin{bmatrix} -0.16 \\ -0.63 \\ -0.82 \\ 1.9 \end{bmatrix}$	$\begin{bmatrix} 0.29 & -1.8 & -0.21 & 1.5 \\ 0.92 & -2.0 & -0.22 & 0.48 \\ 1.8 & 0.65 & -1.5 & -1.3 \\ 1.4 & -1.3 & -0.32 & -1.4 \end{bmatrix}$	This work
S19	$\begin{bmatrix} 0.017 \\ 0.037 \\ 0.16 \\ 0.058 \end{bmatrix}$	$\begin{bmatrix} 0.47 \\ 1.8 \\ 2.6 \\ 0.64 \end{bmatrix}$	$\begin{bmatrix} -0.14 & 0.38 & 0.22 & -0.43 \\ -1.4 & -1.2 & 0.023 & -1.1 \\ 2.2 & 1.7 & -2.8 & -0.044 \\ -0.76 & 2.0 & 0.24 & -0.78 \end{bmatrix}$	This work (seed 64977)
S20	$\begin{bmatrix} 0.049 \\ 0.025 \\ 0.04 \\ 0.011 \\ 0.11 \end{bmatrix}$	$\begin{bmatrix} 2.7 \\ 6.1 \\ 0.60 \\ 2.2 \\ 0.70 \end{bmatrix}$	$\begin{bmatrix} -8.2 & 0.74 & 1.7 & 0.09 & -0.37 \\ 0.15 & -1.9 & 0.37 & 0.56 & -0.42 \\ -0.49 & 0.31 & -4.3 & 0.33 & 0.15 \\ -1.1 & -0.45 & 1.4 & -4.4 & 2.7 \\ -0.95 & 0.67 & -0.52 & -0.80 & -0.95 \end{bmatrix}$	This work (seed 74435)
S24	$\begin{bmatrix} 11 \\ 14 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.25 \\ -0.50 \\ -0.50 \end{bmatrix}$	$\begin{bmatrix} -0.0010 & -0.040 & -0.040 \\ 0.040 & -0.0020 & -0.020 \\ 0.020 & 0.040 & -0.0030 \end{bmatrix}$	This work
S25	$\begin{bmatrix} 10 \\ 14 \\ 4 \end{bmatrix}$	$\begin{bmatrix} 6 \\ 4 \\ 2 \end{bmatrix}$	$\begin{bmatrix} -0.05 & 0.15 & -0.20 \\ -0.01 & -0.027 & 0.050 \\ 0.10 & -0.10 & -0.015 \end{bmatrix}$	[1]
S26	$\begin{bmatrix} 0.04 \\ 0.04 \\ 0.12 \\ 0.04 \\ 0.03 \\ 0.07 \end{bmatrix}$	$\begin{bmatrix} 2.4 \\ 0.005 \\ 0.4 \\ 4.3 \\ 0.5 \\ 1.7 \end{bmatrix}$	$\begin{bmatrix} -0.48 & -0.09 & 1.6 & -1.5 & -1.8 & 1.7 \\ 1.3 & -0.1 & -1.1 & 1.2 & -0.72 & -1.6 \\ -0.5 & 1.1 & -1 & 2.9 & -1.3 & 0.67 \\ 0.42 & 0.13 & -0.9 & -0.42 & -1.1 & 1.1 \\ 0.18 & 1.6 & 1.7 & -0.09 & -3.5 & 0.005 \\ -0.11 & 0.55 & -1.4 & 1.9 & -0.65 & -3.5 \end{bmatrix}$	This work (seed 72088)

S2. Bayesian inference

We consider an N-species generalised Lotka-Volterra model where the population dynamics are governed by the system of differential equations in Eqs. (S1).

Here, we introduce a novel approach to model this using an approximate numerical solution combined with a log-normally distributed noise (as discussed in the main text). Thus, we proceed in two steps: (a) we apply a classical fourth-order Runge-Kutta (RK4) integrator to generate prospect abundances for each species, and (b) we generate the abundances at the next time step through a Bayesian hierarchical model. We detail these two steps in what follows.

(a) Fourth-order Runge-Kutta integrator

Given the abundances $x_{n-1,i}$ at time step $n-1$, we compute

$$k_{1,i} = x_{n-1,i}\left(r_i + \sum_{j=1}^{N}\beta_{ij}x_{n-1,j}\right), \tag{S1}$$

$$k_{2,i} = \left(x_{n-1,i} + \frac{dt}{2}k_{1,i}\right)\left(r_i + \sum_{j=1}^{N}\beta_{ij}x^{(1)}_{n-1,i}\right), \tag{S2}$$

$$k_{3,i} = \left(x_{n-1,i} + \frac{dt}{2}k_{2,i}\right)\left(r_i + \sum_{j=1}^{N}\beta_{ij}x^{(2)}_{n-1,i}\right), \tag{S3}$$

$$k_{4,i} = \left(x_{n-1,i} + dtk_{3,i}\right)\left(r_i + \sum_{j=1}^{N}\beta_{ij}x^{(3)}_{n-1,i}\right). \tag{S4}$$

where

$$x^{(r)}_{n-1,j} = x_{n-1,j} + \frac{dt}{2}k_{r,j}, \quad r = 1, 2, 3, \tag{S5}$$

and produce

$$x^{*}_{n,i} = x_{n-1,i} + \frac{dt}{6}(k_{1,i} + 2k_{2,i} + 2k_{3,i} + k_{4,i}). \tag{S6}$$

(b) Bayesian hierarchical model

The parameters of the model follow prior distributions:

$$\sigma_i \sim \text{Exponential}(1), \tag{S7}$$

$$r_i \sim \text{Exponential}(0.1), \tag{S8}$$

$$\beta_{ij} \sim \mathcal{N}(0, 2). \tag{S9}$$

Given the population values $x_{n,i}$ obtained via RK4, the likelihood function assumes:

$$x_{n,i} \sim \text{Lognormal}(\log x^{*}_{n,i}, \sigma_i), \tag{S10}$$

with the $x^{*}_{n,i}$ obtained in (S6)

S3. Marginal posterior distributions

(a) 3 species

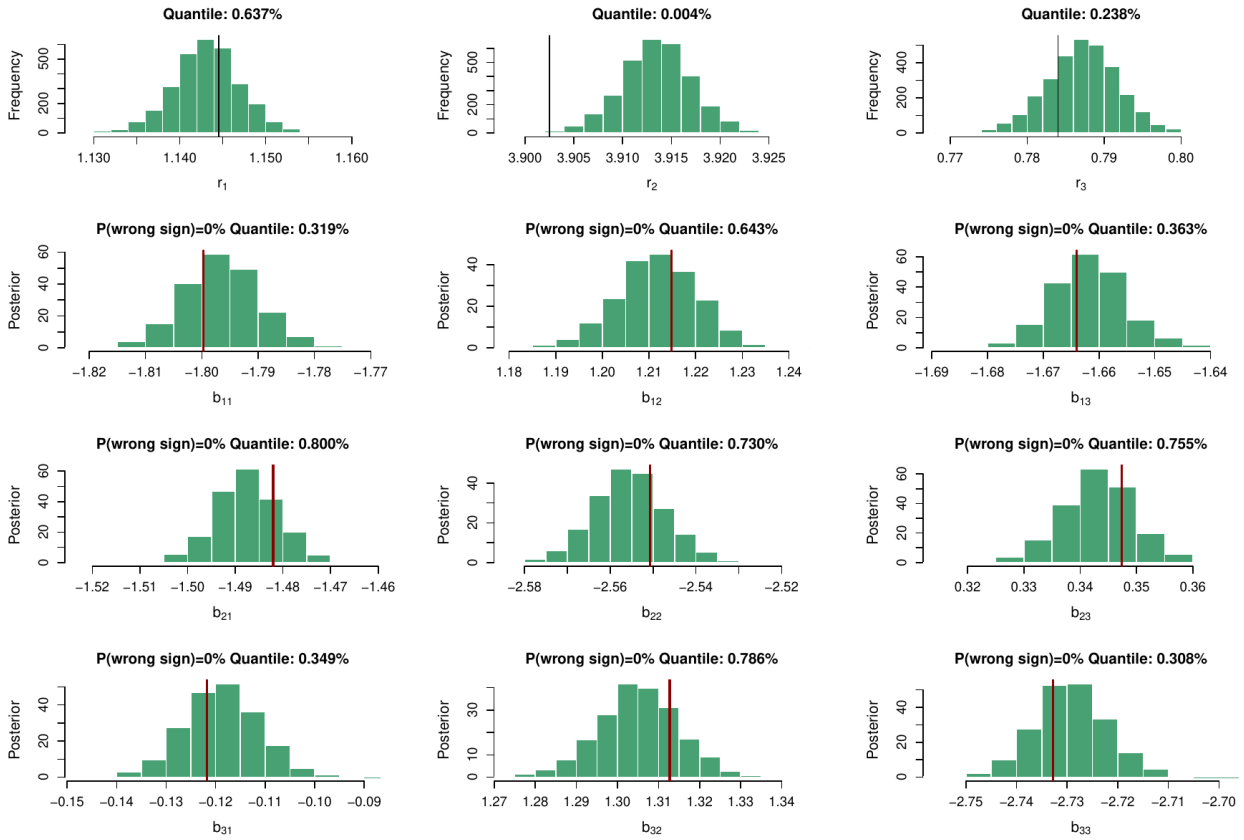

Figure S1. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.001. The vertical red line is the original parameter value used in the simulation.

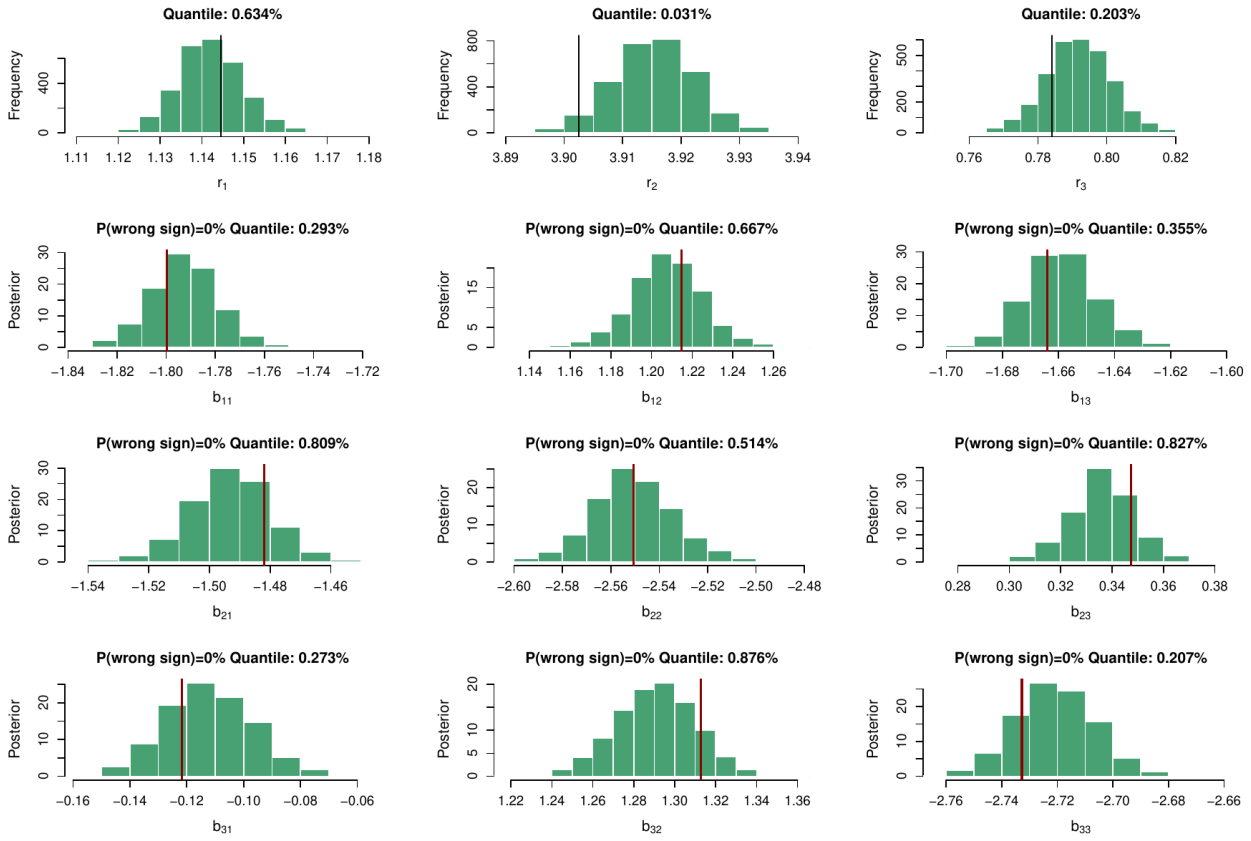

Figure S2. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.002. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

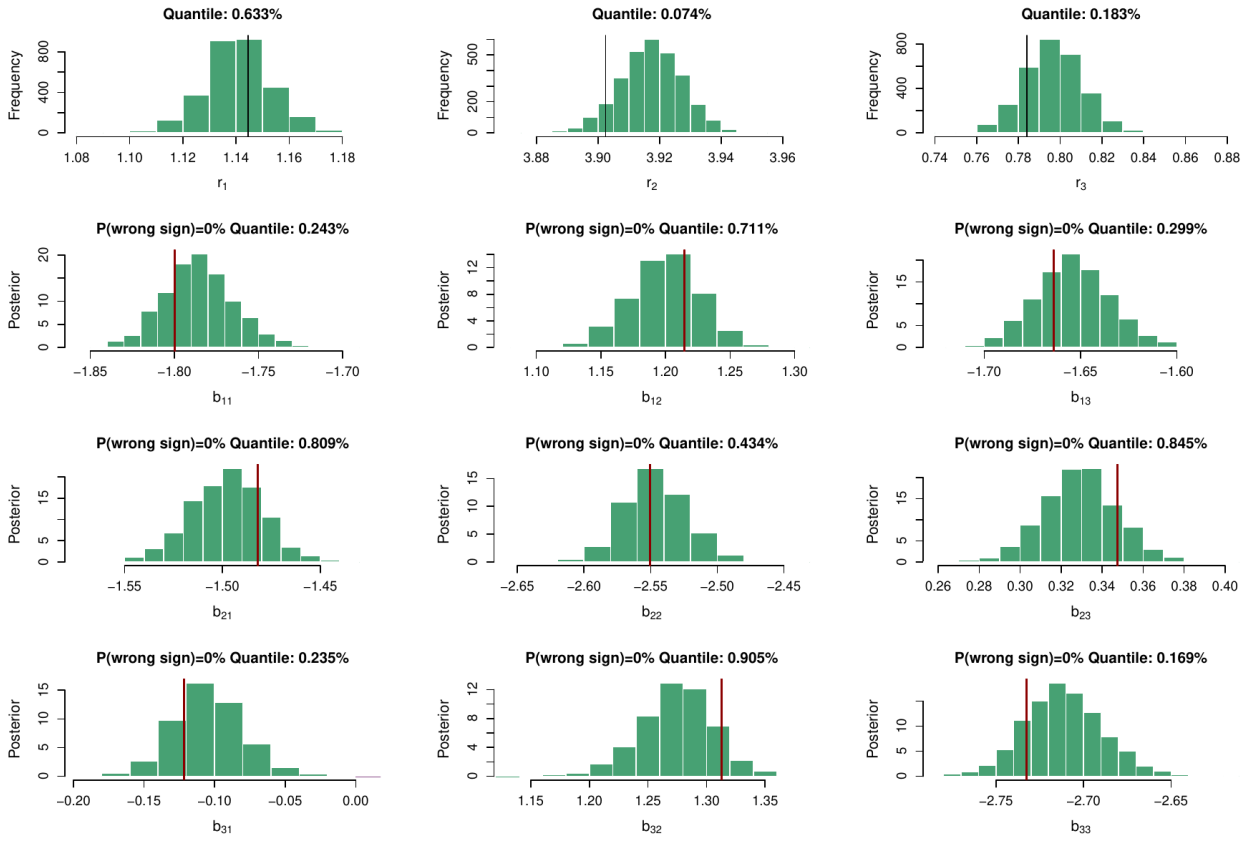

Figure S3. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.003. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

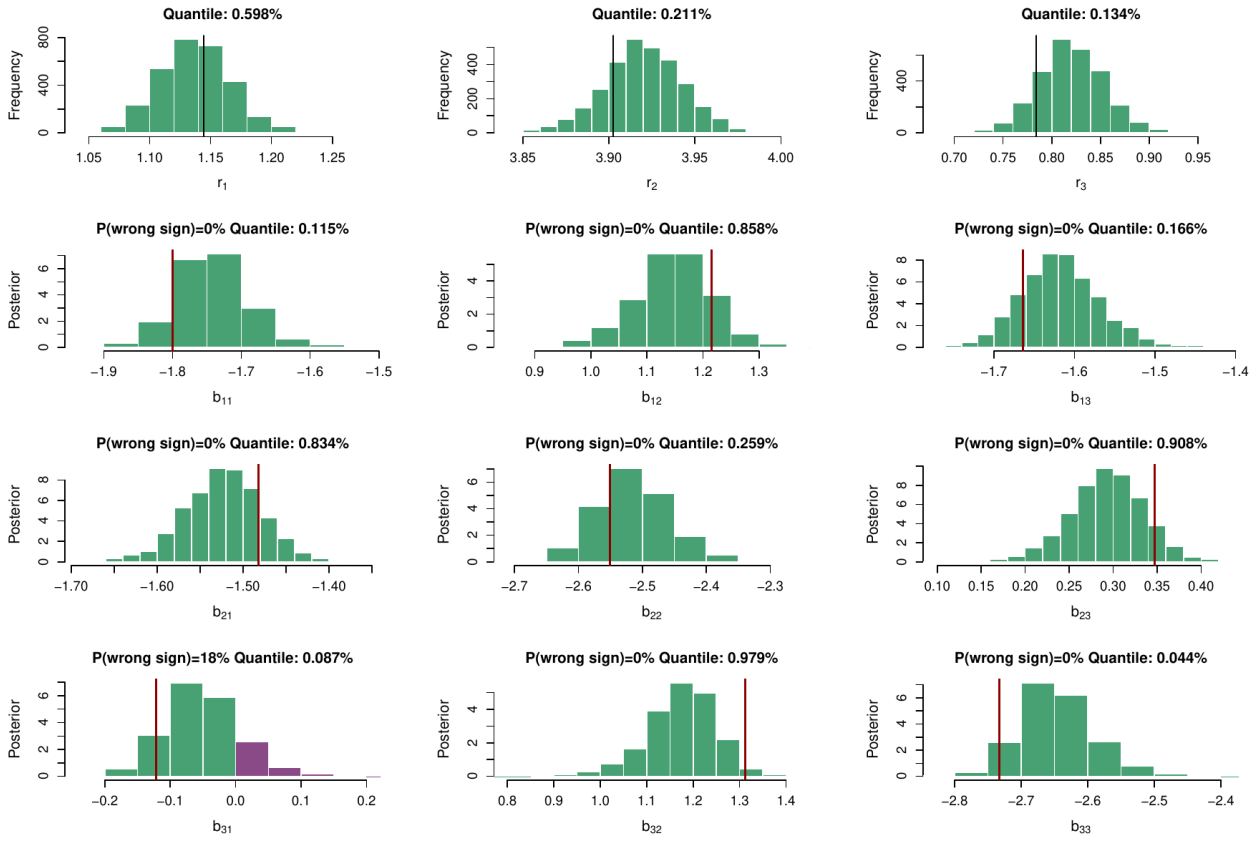

Figure S4. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.007. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

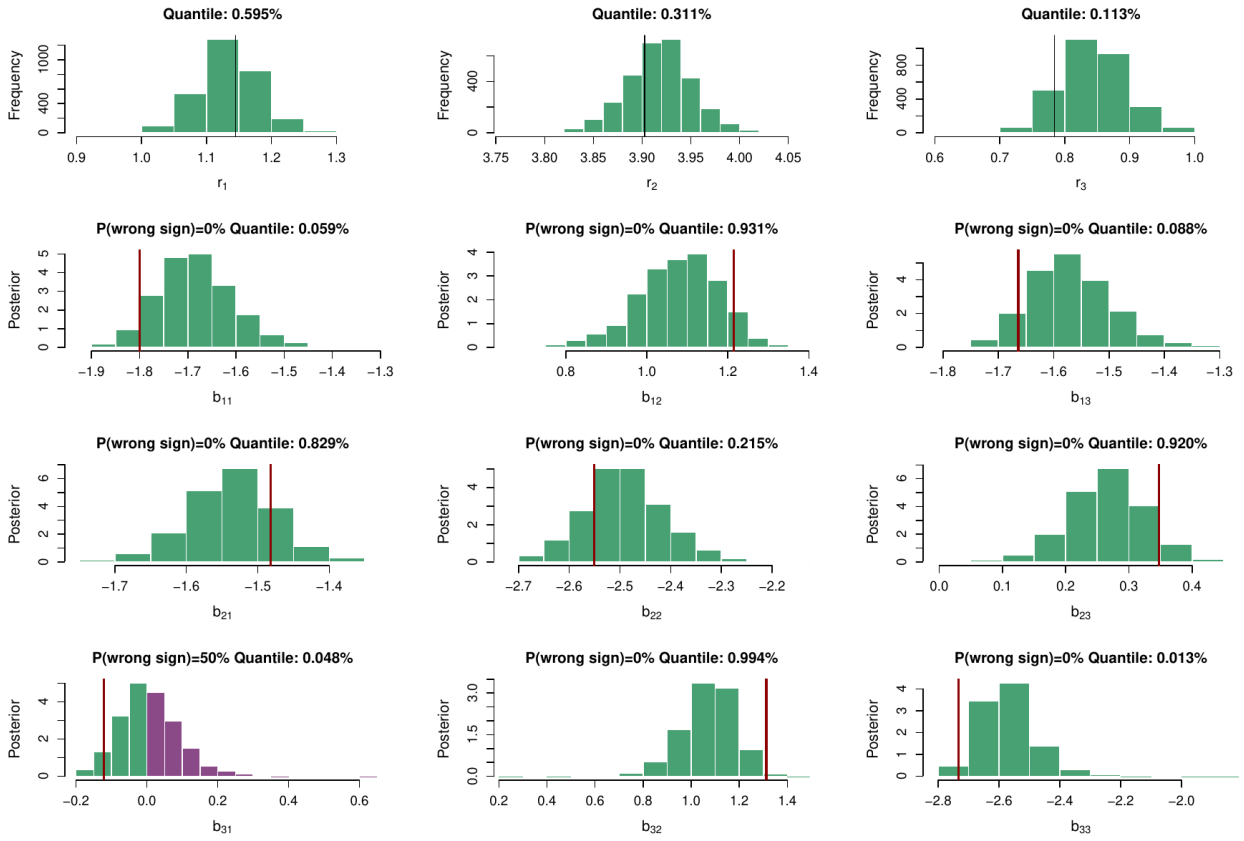

Figure S5. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.01. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

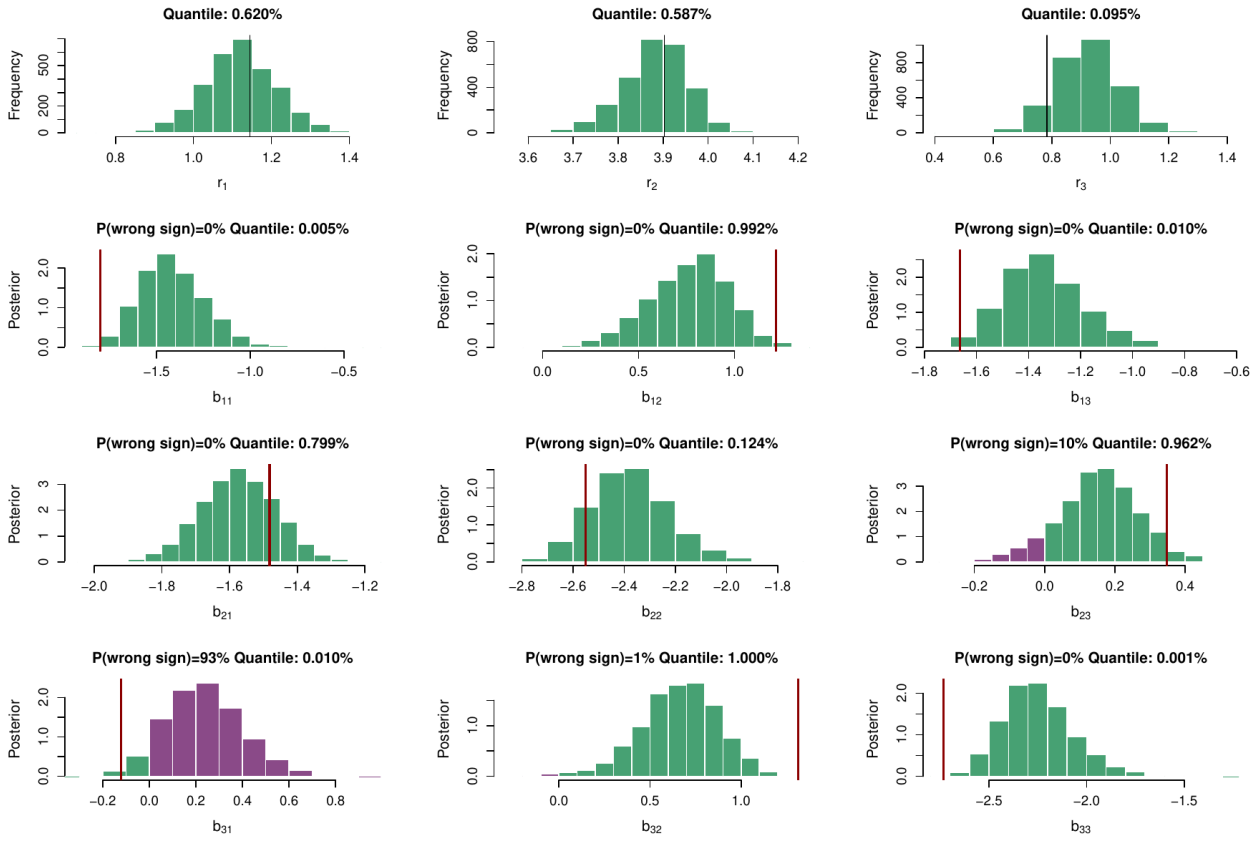

Figure S6. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.02. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

(b) 4 species

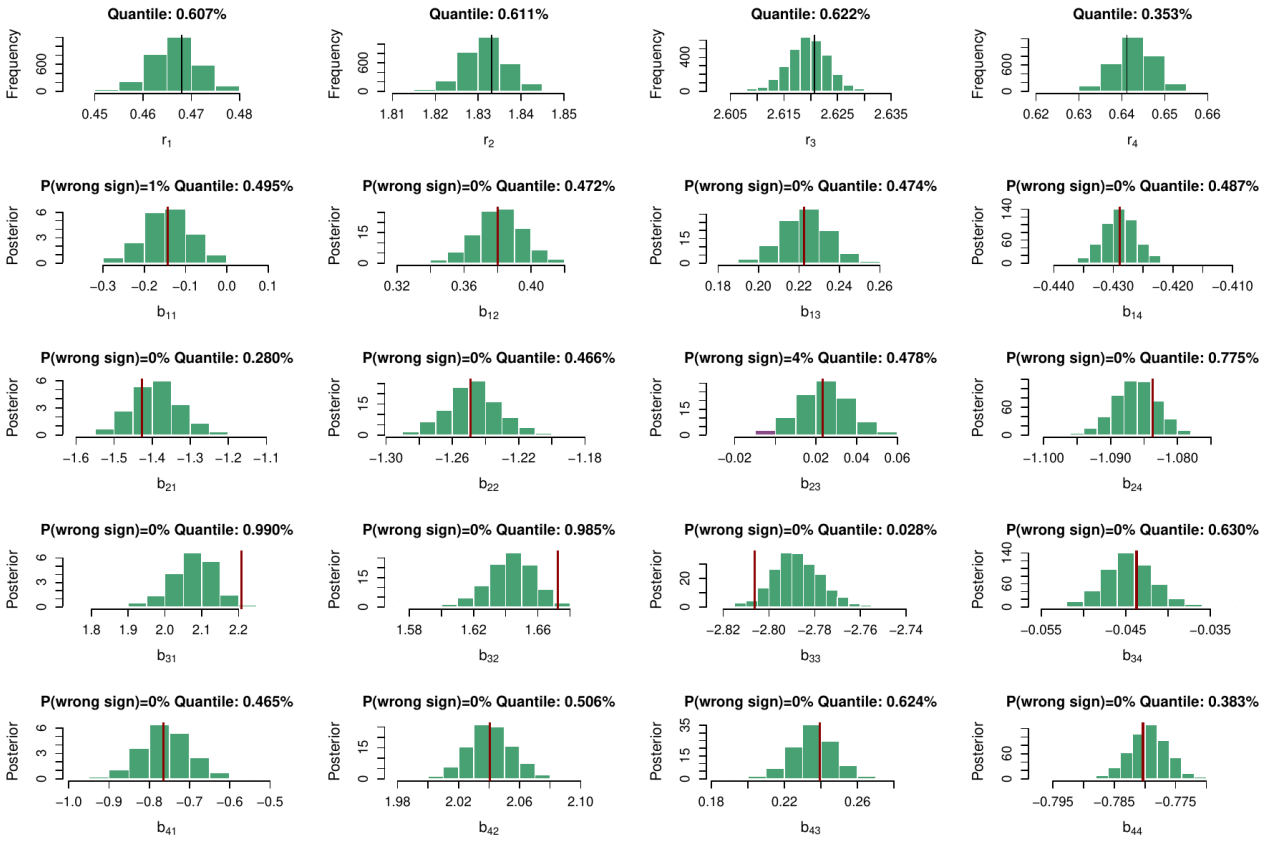

Figure S7. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.001. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

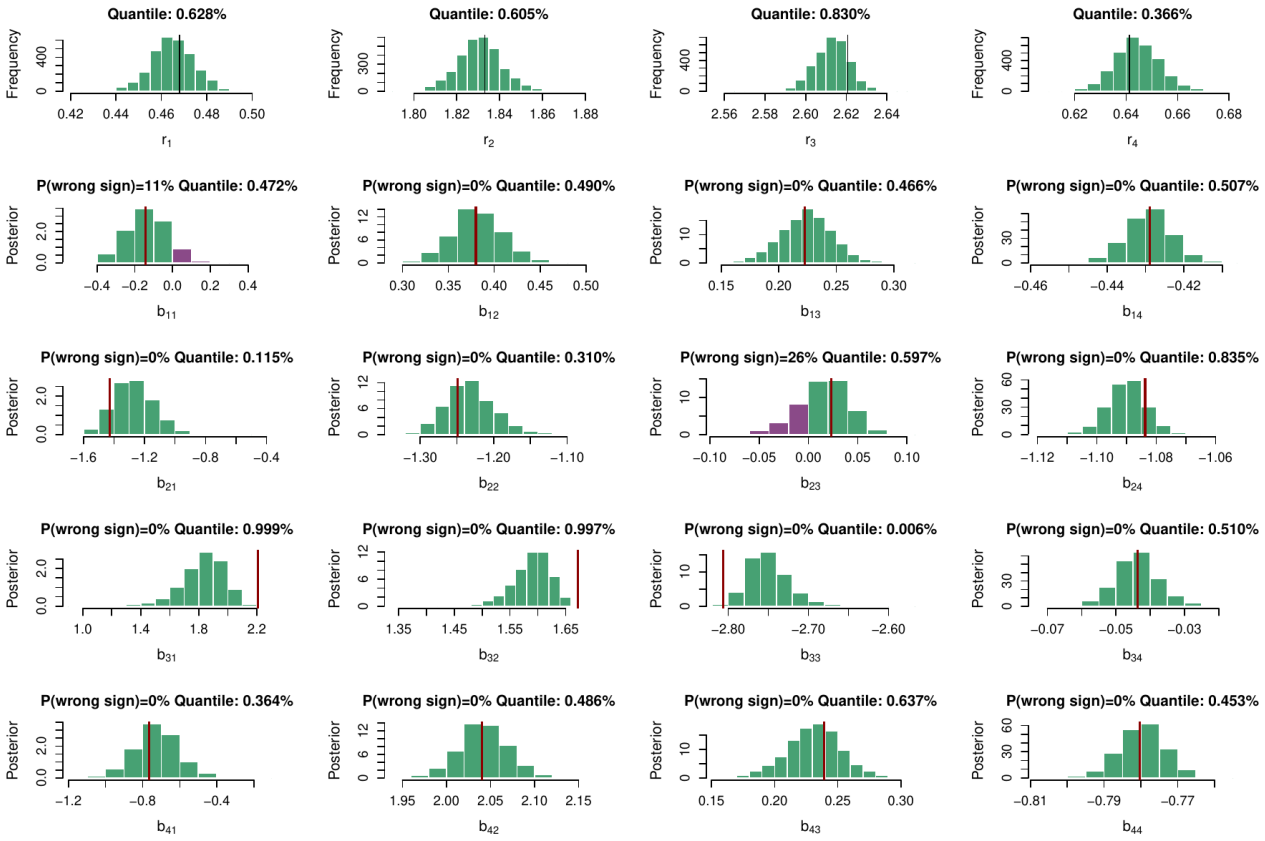

Figure S8. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.002. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

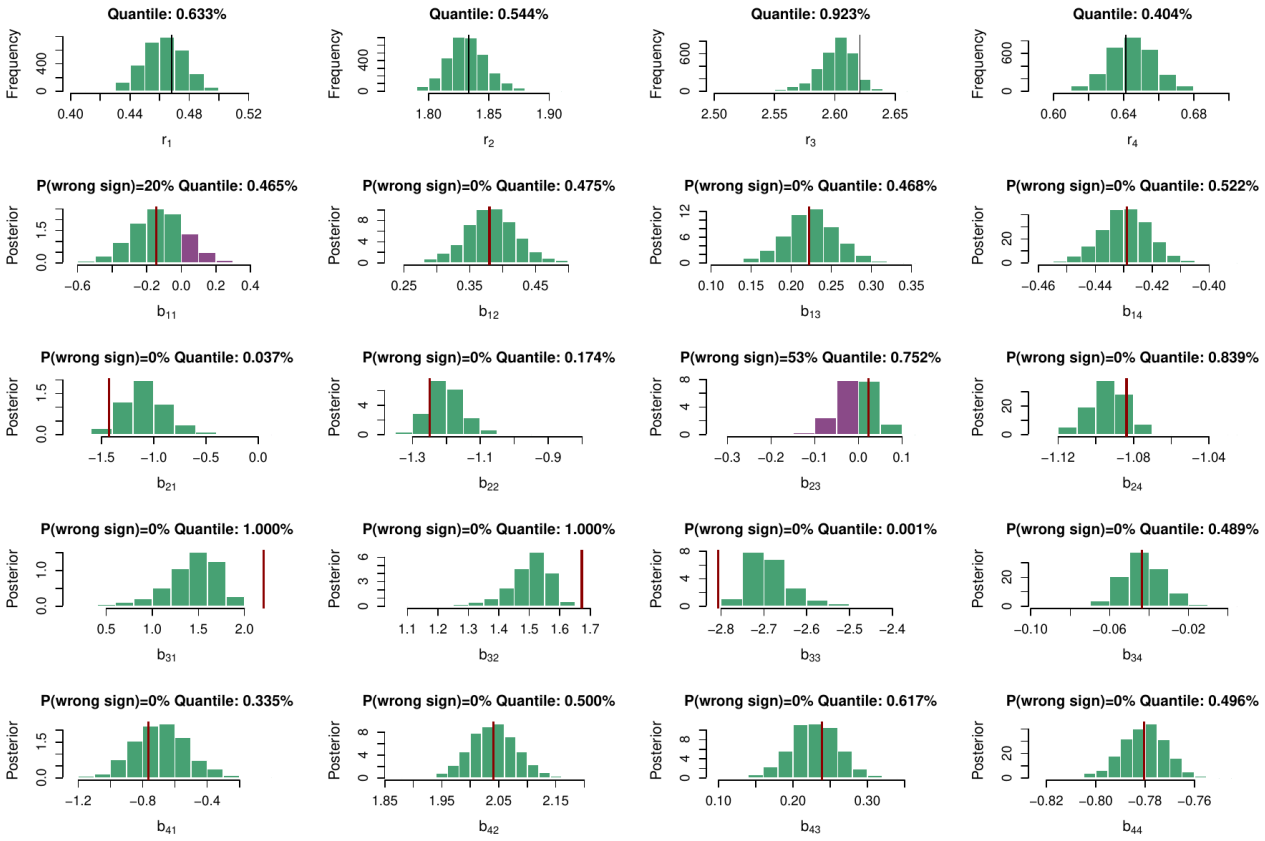

Figure S9. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.003. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

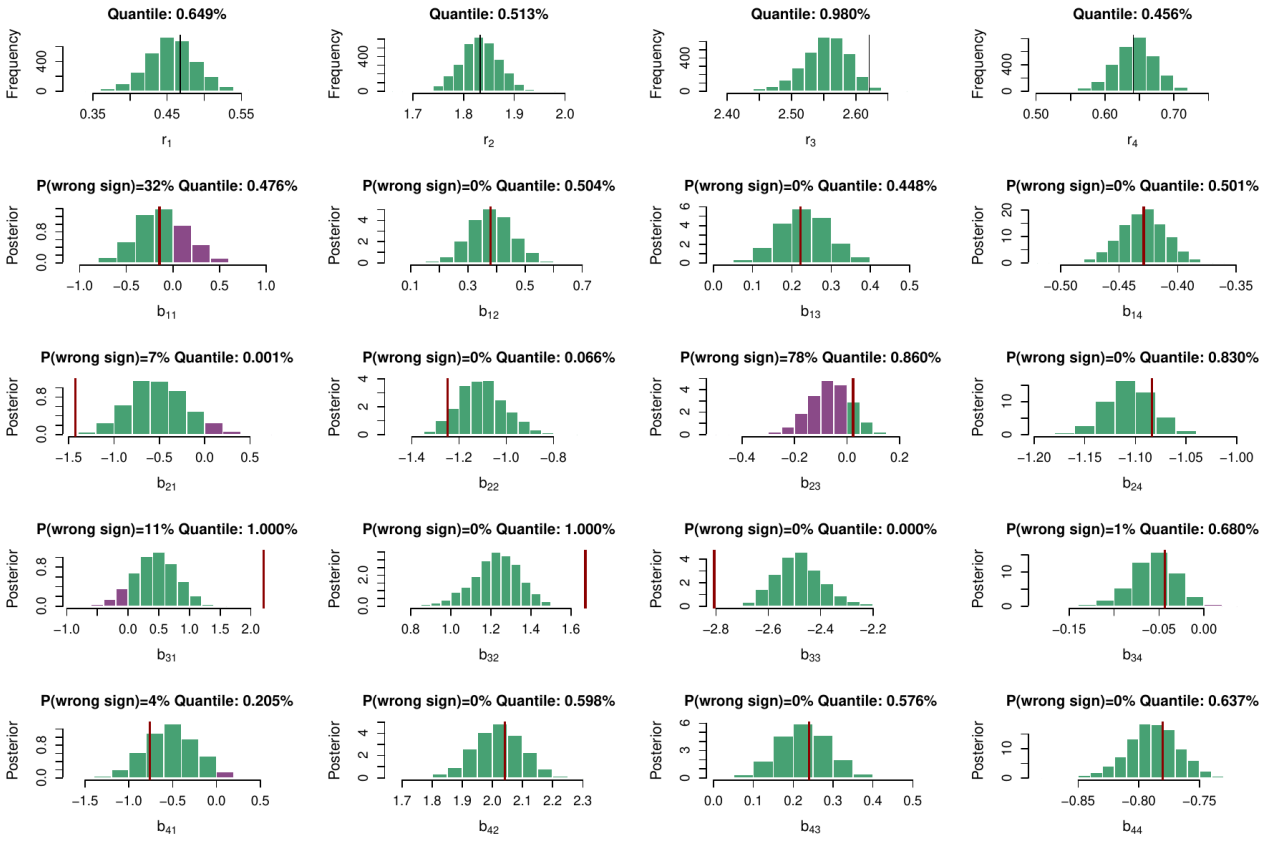

Figure S10. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.007. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

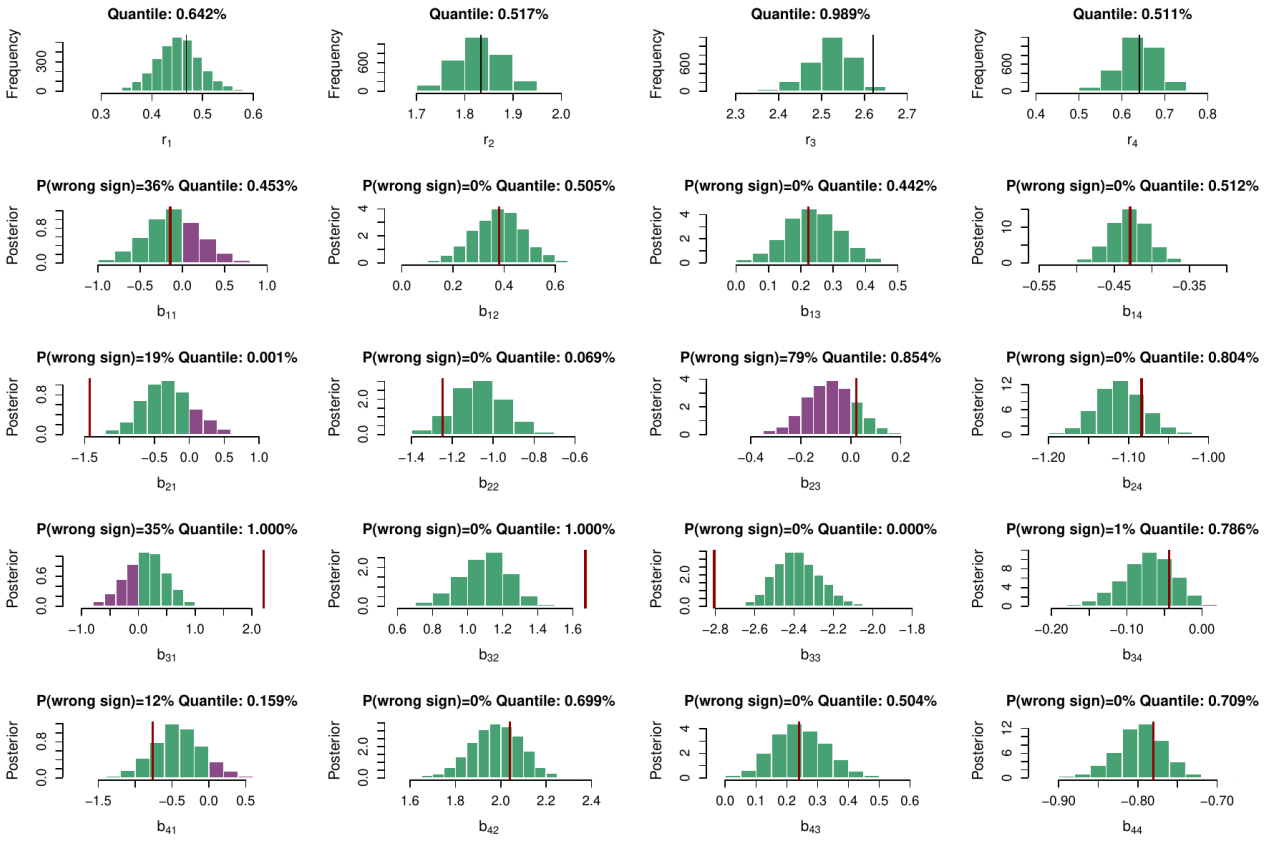

Figure S11. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.01. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

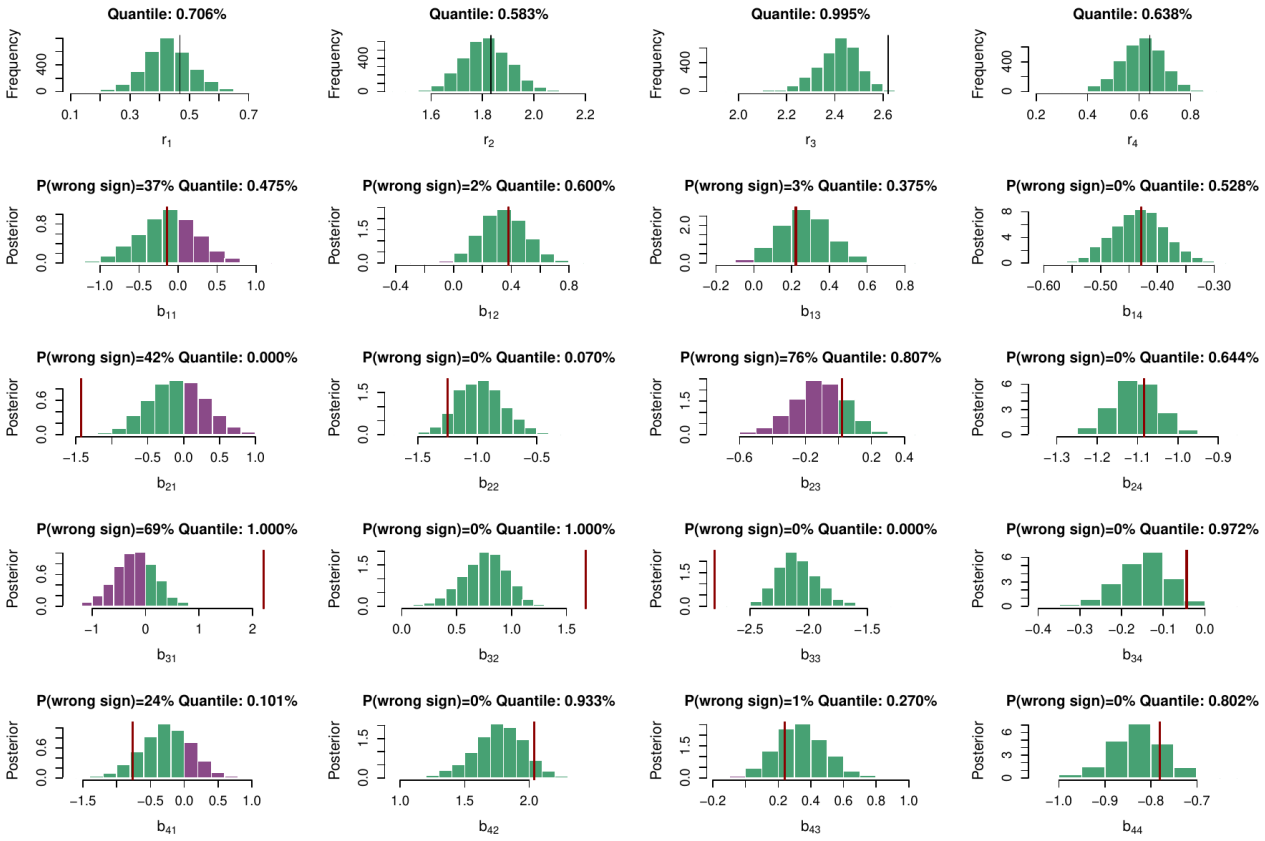

Figure S12. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.02. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

(c) 5 species

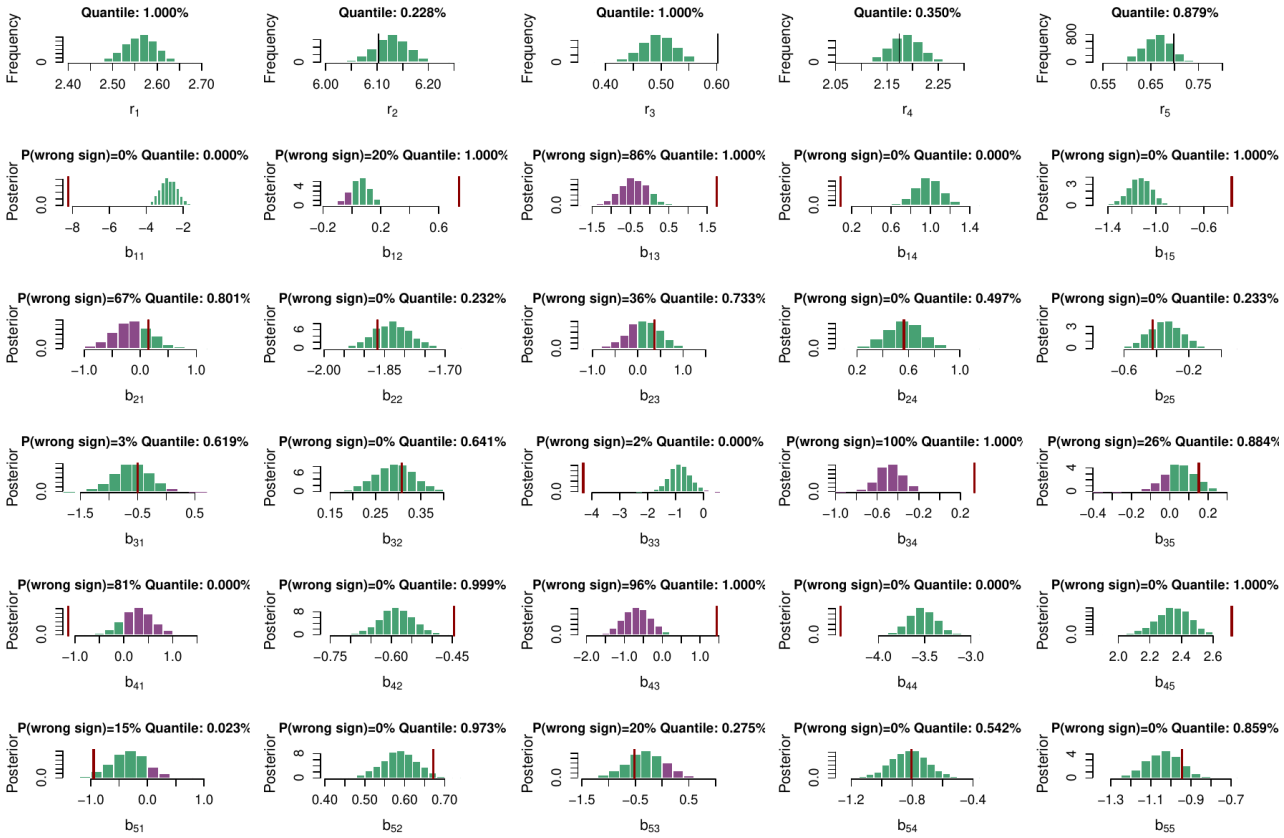

Figure S13. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.001. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

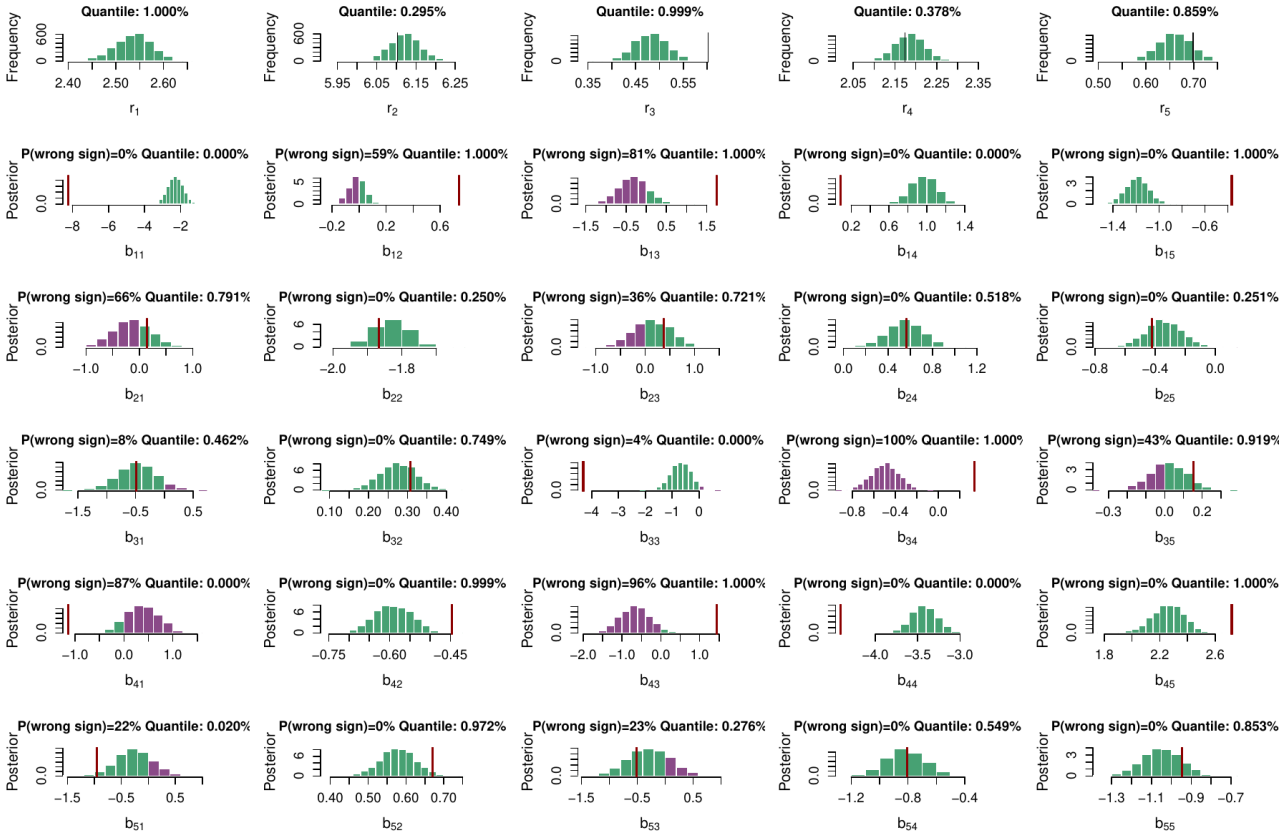

Figure S14. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.002. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

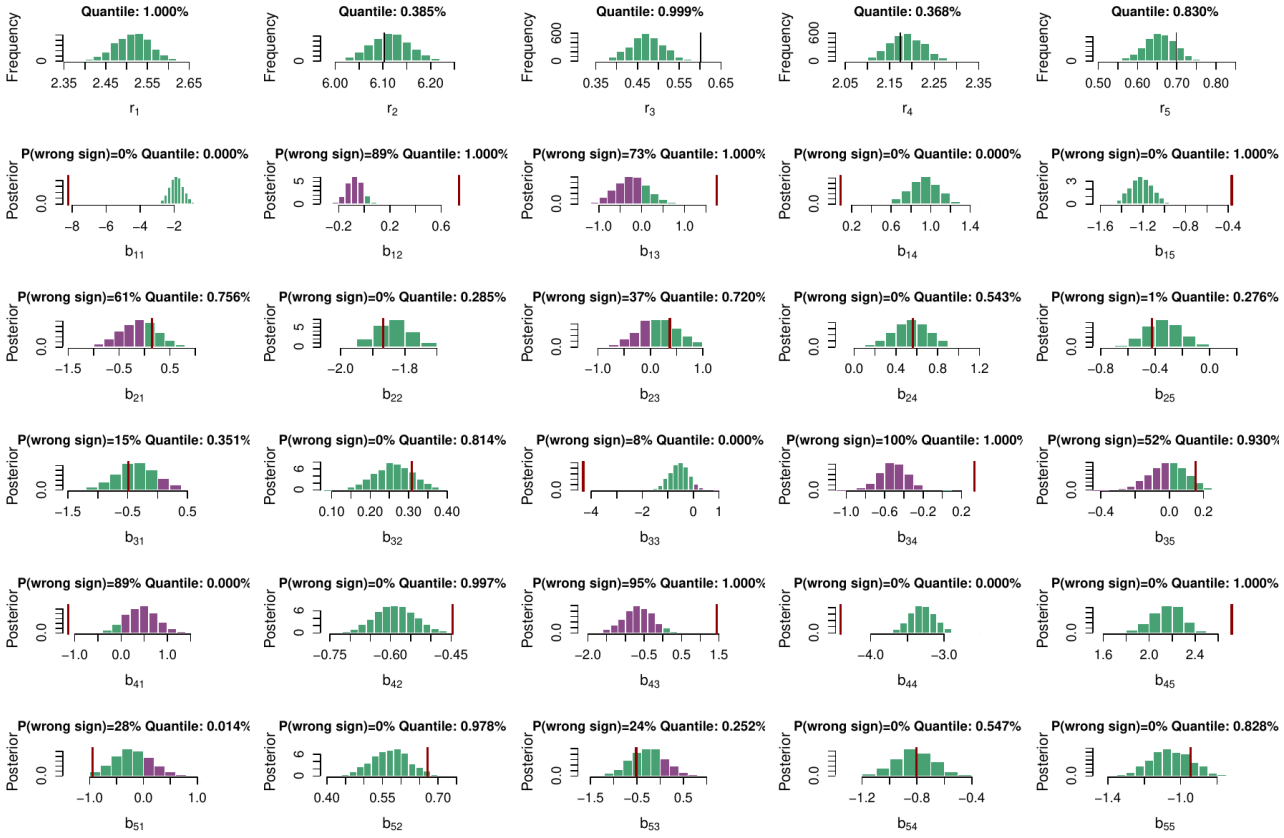

Figure S15. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.003. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

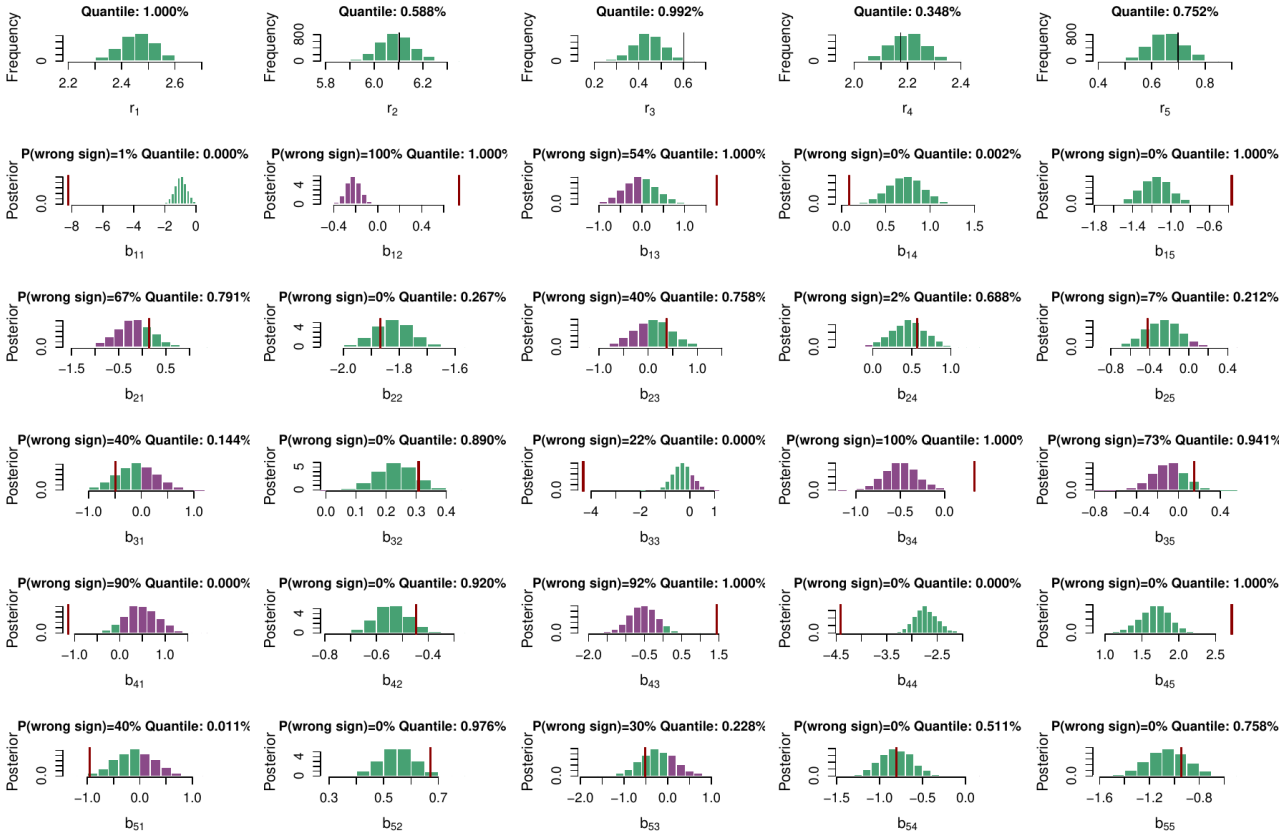

Figure S16. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.007. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

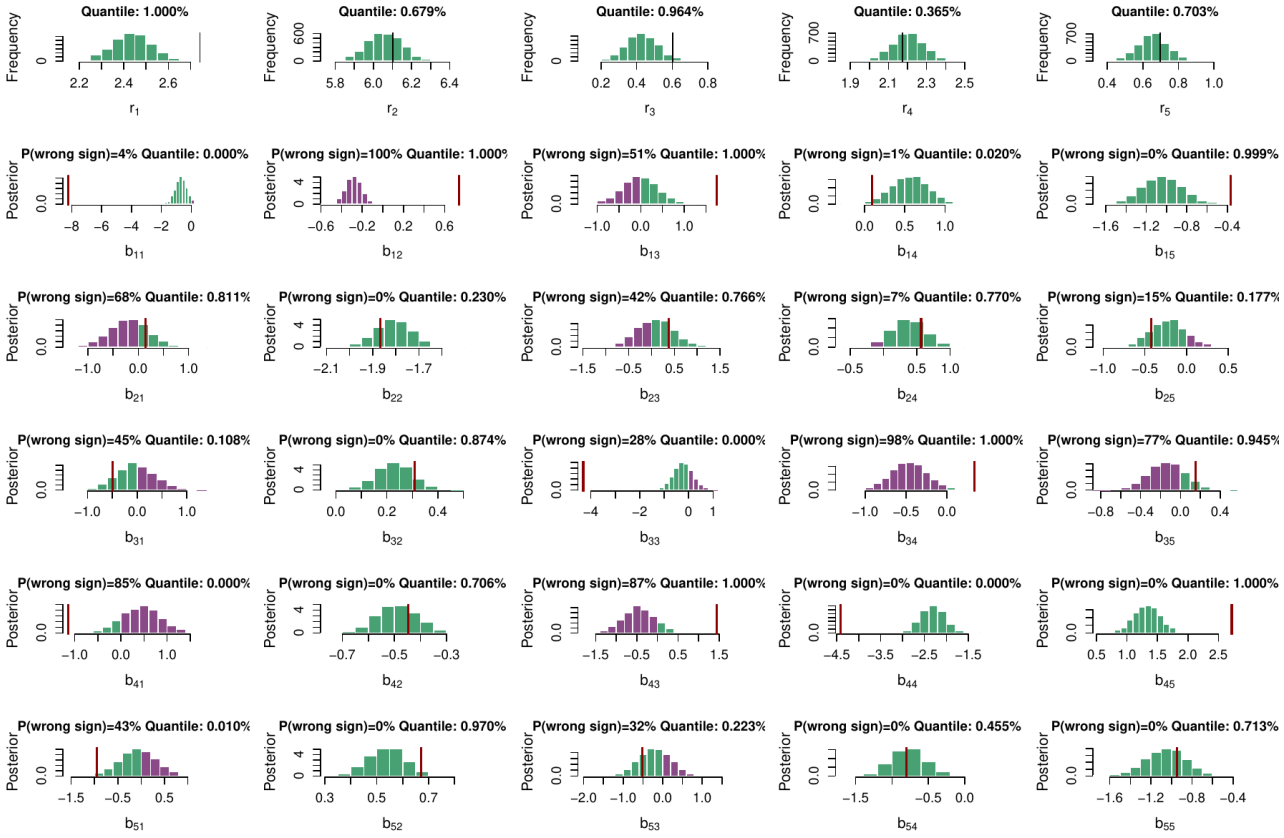

Figure S17. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.01. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

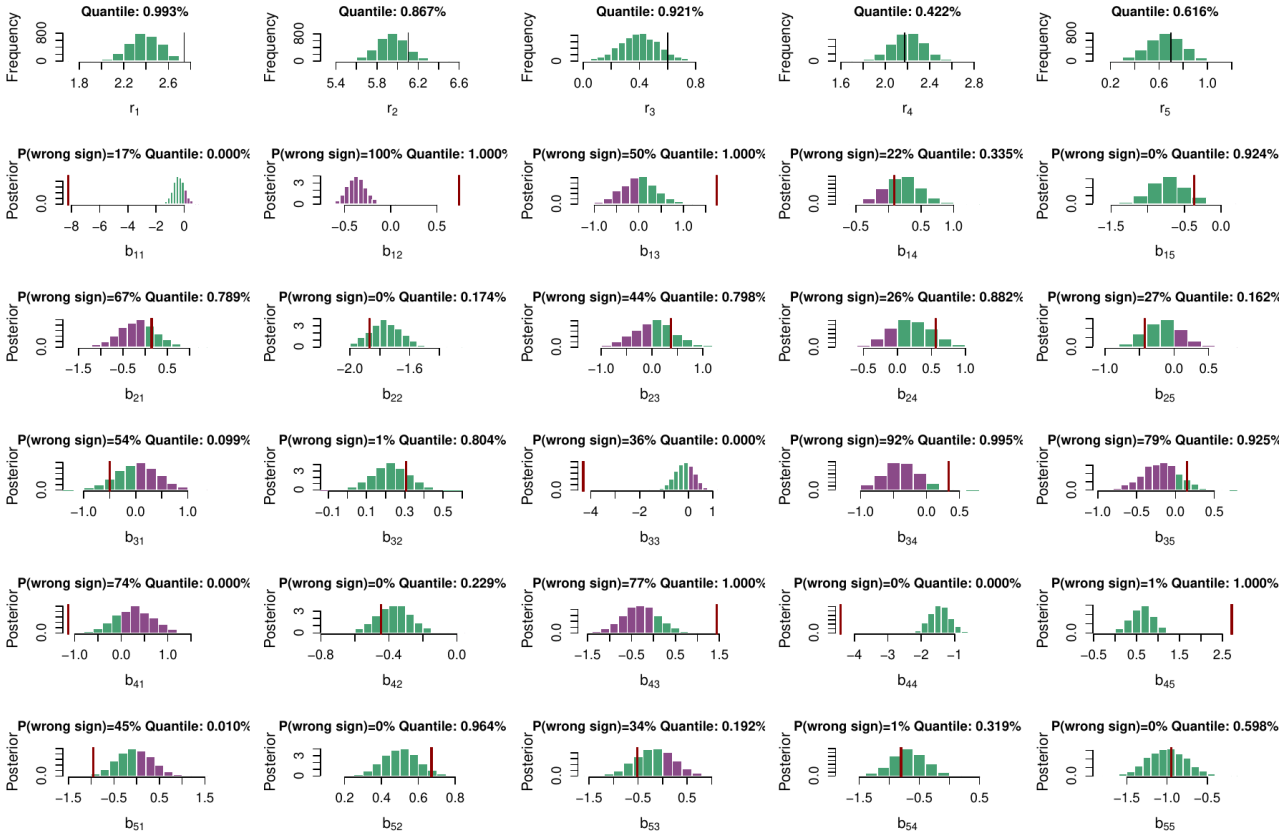

Figure S18. Marginals of the posterior distributions for the parameters of the gLV adding log-normal noise with standard deviation 0.02. The vertical red line is the original parameter value used in the simulation. In purple, we show those ranges of parameters that have a change of sign.

S4. Posterior predictive checks

(a) 4 species

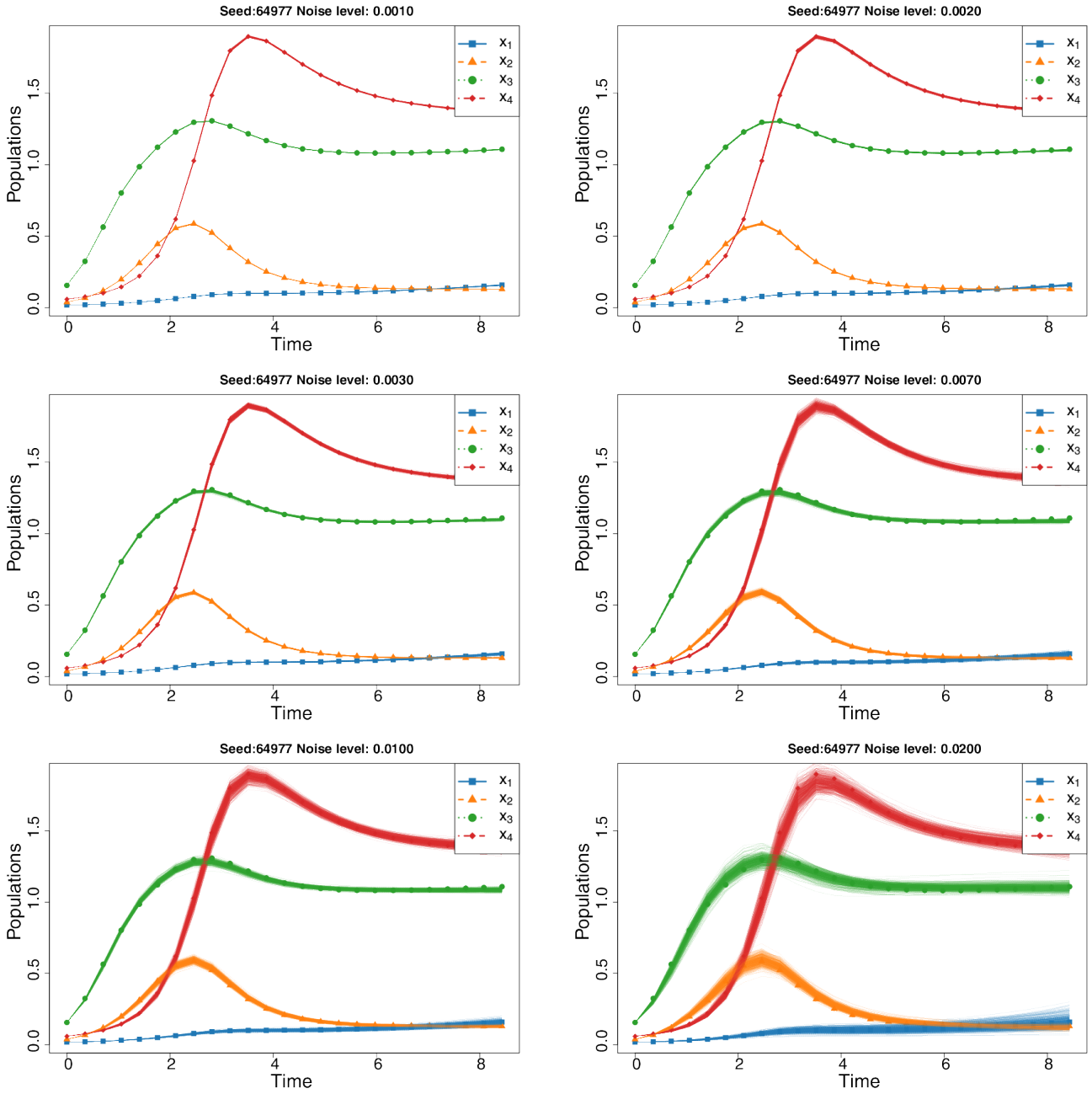

Figure S19. Posterior predictive trajectories simulated by sampling the posterior distributions (summarized in Sec. S3) for 6 different log-normally noise levels (title of each panel) added to the deterministic curves. The posterior predictive trajectories capture higher (log-normal) noise levels by increasing the variability for larger abundances.

(b) 5 species

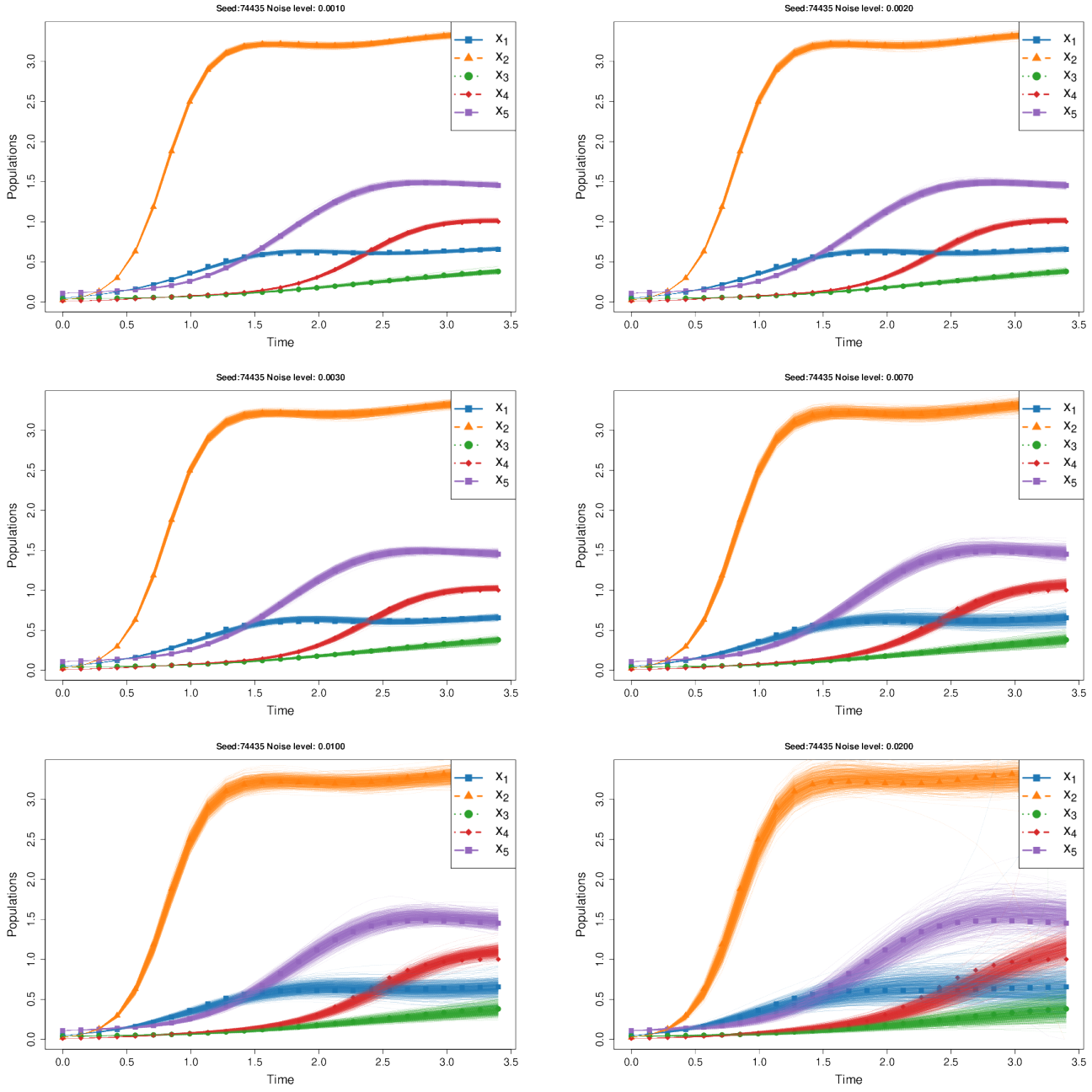

Figure S20. Posterior predictive trajectories simulated by sampling the posterior distributions (summarised in Sec. S3) for 6 different log-normally noise levels (title of each panel) added to the deterministic curves. The posterior predictive trajectories capture higher (log-normal) noise levels by increasing the variability for larger abundances.

(c) 6 species

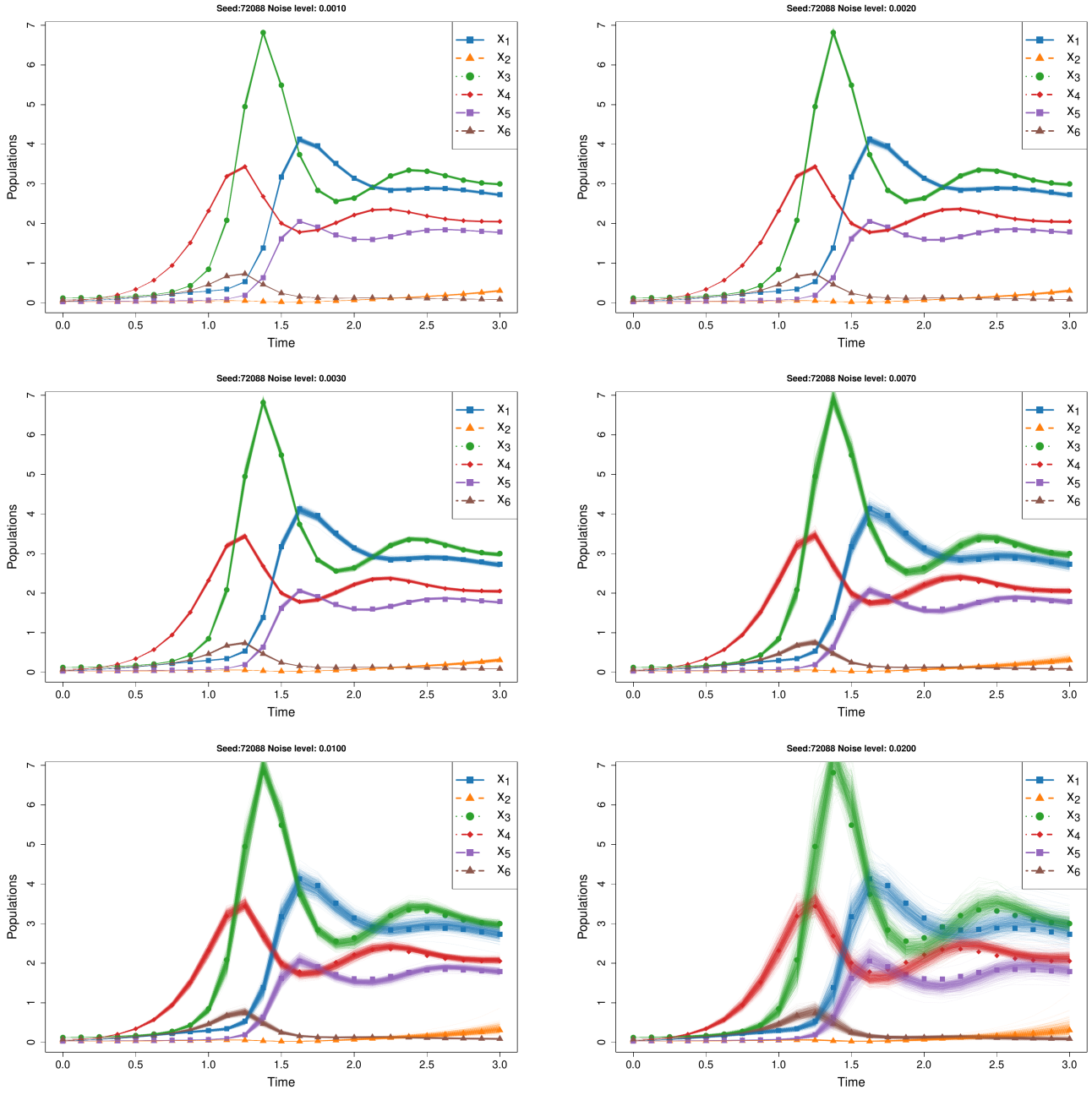

Figure S21. Posterior predictive trajectories simulated by sampling the posterior distributions (summarized in Sec. S3) for 6 different log-normally noise levels (title of each panel) added to the deterministic curves. The posterior predictive trajectories capture higher (log-normal) noise levels by increasing the variability for larger abundances.

S5. Correlations between posterior marginals for 4 species

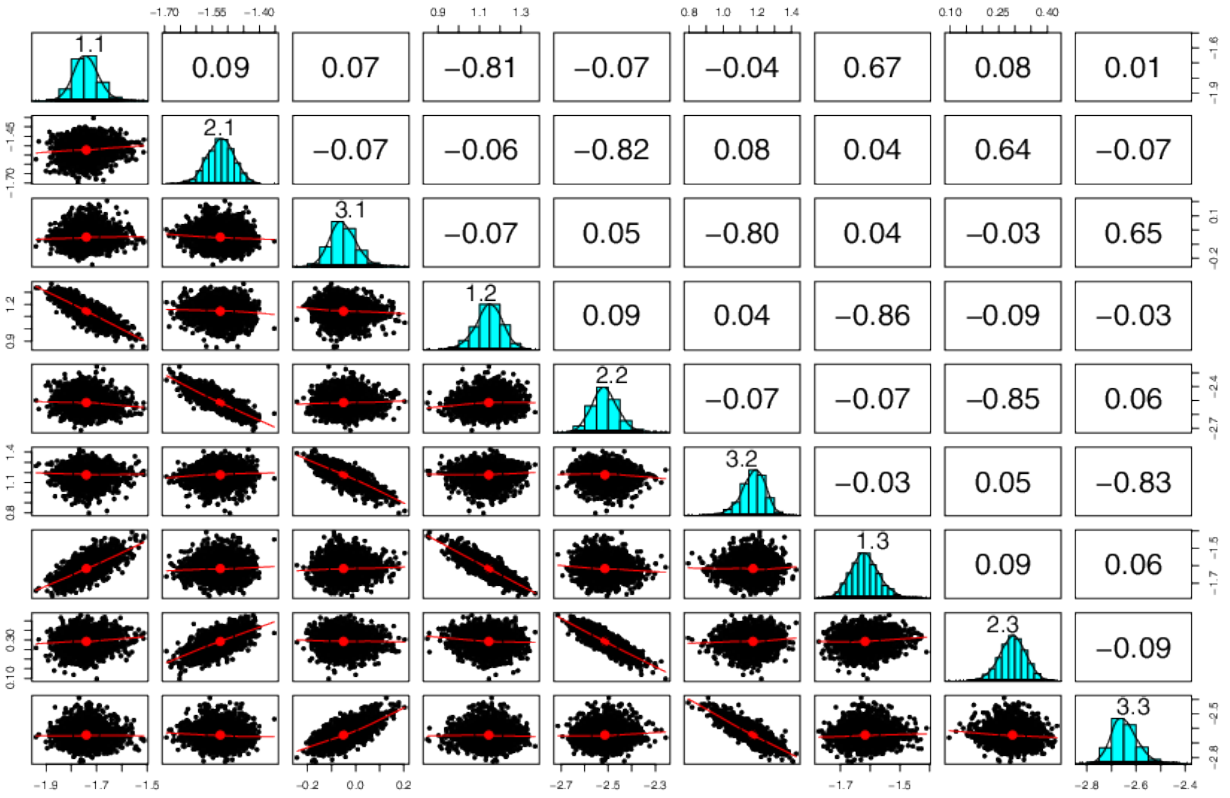

Figure S22. Pairs-plot of the marginal posterior distributions for the elements of the interaction matrix, β_{ij} for 3 species, and noise level 0.007. The label on the diagonal of the plot shows the index of the coefficient of β_{ij}. In the upper triangular part of the plot, we show the Pearson correlation between distributions. The numbers above each histogram indicate the indices of the interaction matrix. Note how the condition imposed on the parameters in order to provide positive, stable, steady-state abundances induces spurious correlations between parameters.

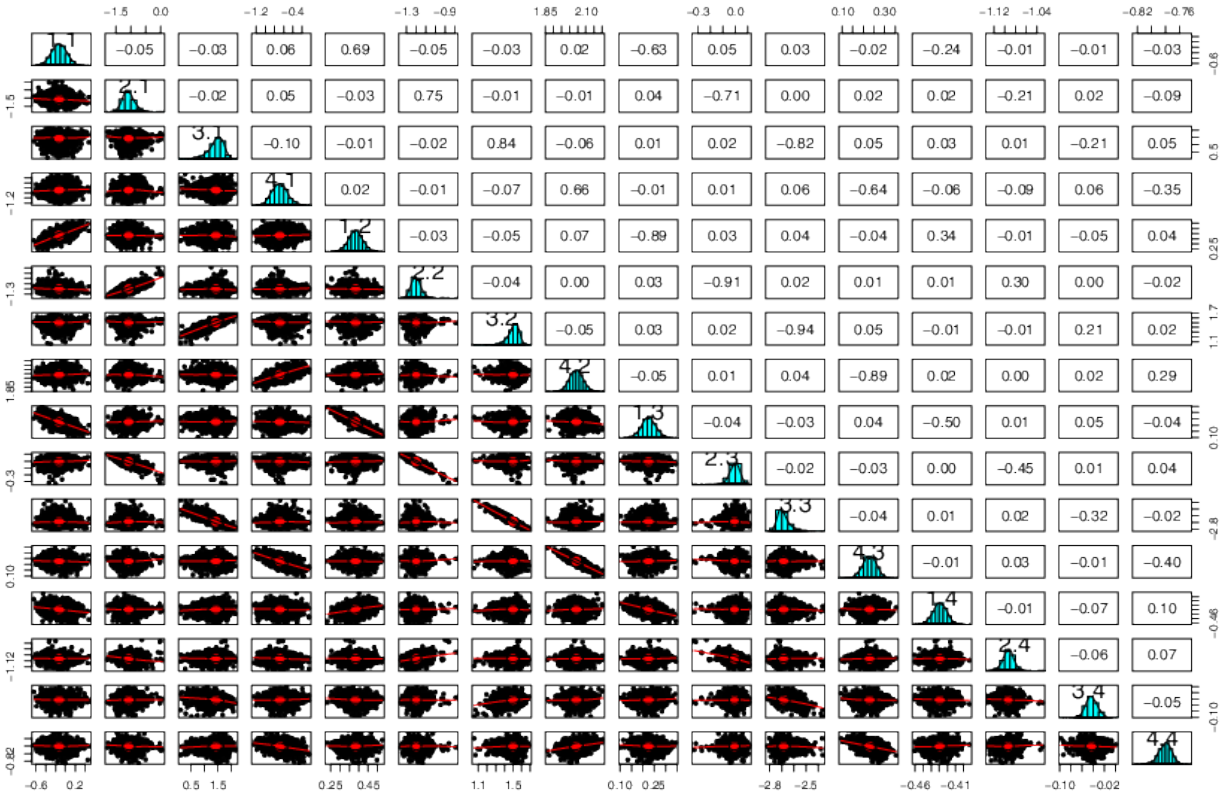

Figure S23. Pairs-plot of the marginal posterior distributions for the elements of the interaction matrix, β_{ij} for 4 species, and noise level 0.003. The label on the diagonal of the plot shows the index of the coefficient of β_{ij}. In the upper triangular part of the plot, we show the Pearson correlation between distributions. The numbers above each histogram indicate the indices of the interaction matrix. Note how the condition imposed on the parameters in order to provide positive, stable, steady-state abundances induces spurious correlations between parameters.

S6. An example of MBAM for parameters giving periodic solutions

In Fig. S24, we show the initial and final model after applying the MBAM algorithm. The sequence provided by MBAM is as follows:

$$\beta_{33} \to 0, \beta_{22} \to 0, \beta_{23} \to 0, \text{and } \beta_{31} \to 0.$$

Further elimination provides inaccurate fitting of x_3 as shown in the figure.

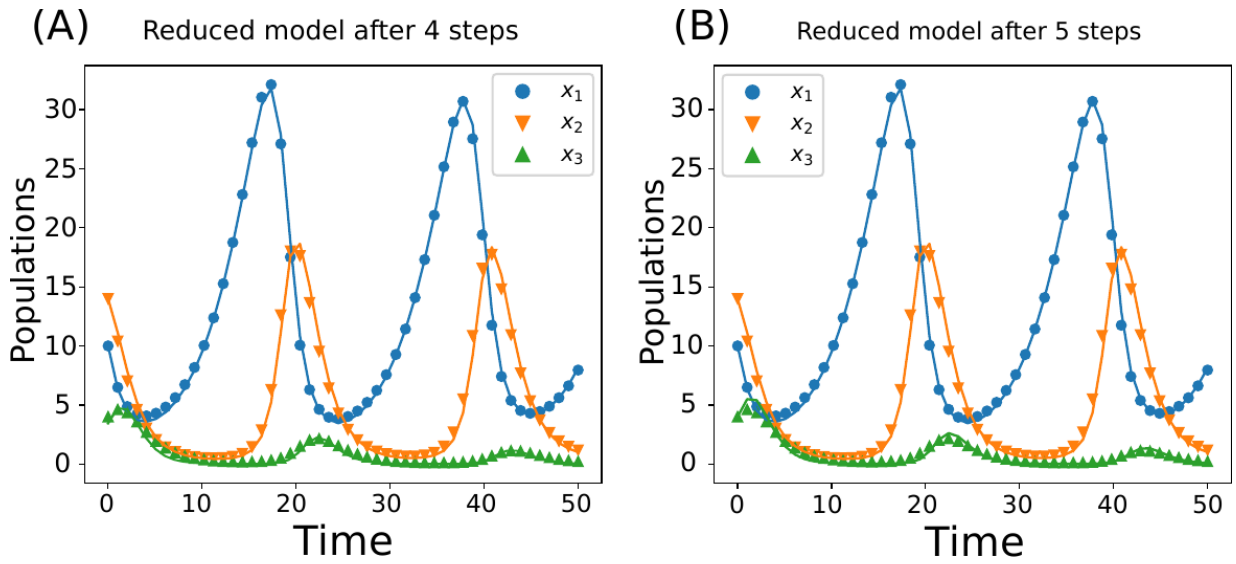

Figure S24. Comparison of a system entailing periodic oscillations (see parameters in Table S1) after (A) 4 MBAM and (B) 5 MBAM steps (solid lines). The symbols represent the original deterministic trajectories. Note how removing the 5th parameter fails to describe x_3 accurately.

S7. Identifiability vs Sloppiness

As shown in Ref. [2], sloppiness and identifiability (practical or structural) are different concepts. The former relates to parameter meaning constrained to the information in the data, and the latter to the ability to assign a unique value to each model parameter.

In Ref. [1], the authors proved structural identifiability and illustrated that fitting relative abundances also provided practical identifiability. However, there is a caveat in their analysis: the priors used to make the Bayesian inference in that work were constrained to be between uniformly distributed 0.6 and 1.4 times the original ones, so the change in sign has probability 0.

To illustrate this, we applied MBAM to the same data parameters as in [1], which allowed us to eliminate 4 parameters while the fitting of the model is still accurate, as shown in Fig. S25. In Table S1 we collect the parameters used in the simulation. The sequence of parameters eliminated by MBAM is:

$$\beta_{33} \to 0, \beta_{21} \to 0, r_3 \to 0, \text{and } \beta_{23} \to 0,$$

but removing an additional one ($\beta_{22} \to 0$) fails to fit the model. This is insightful because it manifests that priors must be uninformative enough and that identifiability does not guarantee a lack of sloppiness and the ability to explain the data with a simpler model.

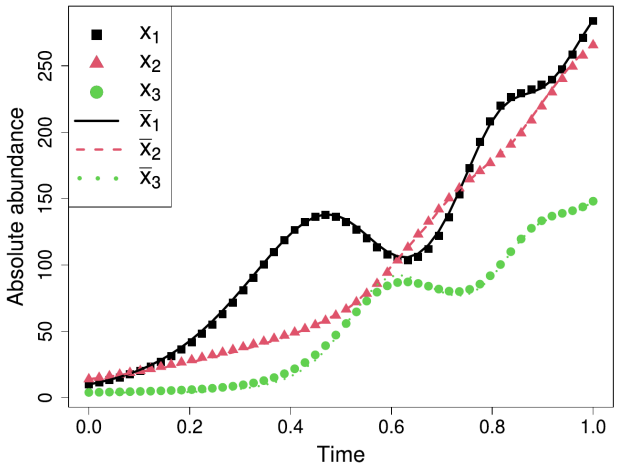

Figure S25. Comparison of exact simulation (symbols) of gLV model for 3 species using the same parameters as in Ref. [1] and the reduced model after 4 MBAM steps (lines). Note how, even though the model is identifiable, MBAM still purges 4 parameters while preserving the structure and idiosyncrasies of the data.

S8. The problem of forecasting with overparametrized models

In this section, we show that fitting data for different times does not necessarily solve the problem of interpretability and sloppiness of gLV-like models. Specifically, longer time series can only help if they add more effective information than that already implicit in the shorter series. It is only through more relevant information that we can determine more parameters. But if extending the time series ends in a steady state, for instance, then it doesn't matter how many new points we add to the time series—it doesn't improve prediction.

To illustrate this, in Fig. S26, we show an example with $N = 6$ species. In Fig. S26(A), we show how the full gLV (with 42) parameters and a reduced version where we have eliminated 12 fit the data accurately for times $t <= 3$ where they have been fitted, but both predict poorly and even the full model extrapolates worse for longer times, as shown in Fig. S26(B). So, not only is the interpretability of the interactions not attainable, but more complex models also perform worse for out-of-sample data.

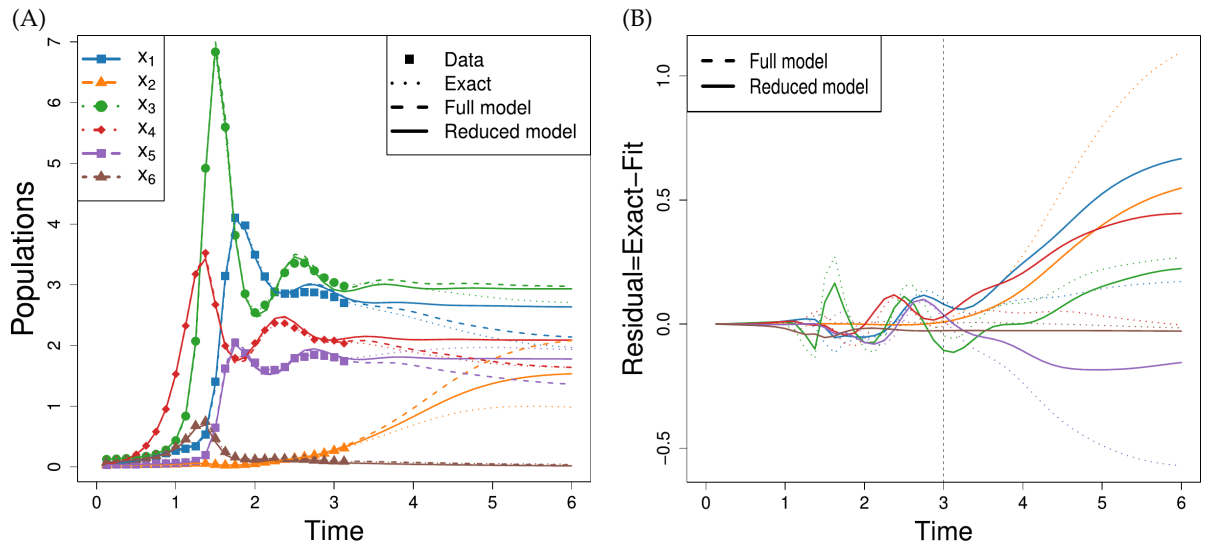

Figure S26. Time series obtained from a gLV model with $N = 6$ species. (A) Abundances of the species as obtained from the exact model (dotted lines), the fit to the bullet points (obtained from the exact data by adding a small amount of noise) of the full model with 42 parameters (dashed lines), and the fit to a reduced model with 30 (out of 42) non-zero parameters (full lines). (B) Difference between the abundances predicted by both fitted models (full and reduced) and the exact abundances. The vertical dotted line marks the end of the fitted data. Beyond this point, the curves are forecasting. This figure illustrates the impossibility of long-term forecasting these time series. As time passes beyond the available data, both models start forecasting poorly—the worse the longer the time.

To test this, in Fig. S27, we demonstrate why we claim that inferring parameters is doomed. The main message is that we can remove the same number of parameters, but the removed parameters in both models are different (shown in red). In particular, for shorter times, the removed parameters are

$$r_2, \beta_{12}, \beta_{22}, \beta_{25}, \beta_{32}, \beta_{42}, \beta_{52}, \beta_{54}, \beta_{61}, \beta_{62}, \beta_{65}, \beta_{66},$$

and for longer,

$$r_2, \beta_{16}, \beta_{25}, \beta_{26}, \beta_{36}, \beta_{46}, \beta_{52}, \beta_{54}, \beta_{56}, \beta_{62}, \beta_{64}, \beta_{66},$$

where those in red are missing in the other scenario. This removal procedure is computationally expensive, so for $N = 6$ we have stopped at 12 removed parameters, but extending the series does not have any effect in this case. As said above, if the longer times contain additional or richer features, we agree that there will be more relevant parameters, but this is precisely our overarching point: With poor data, you have poor estimates and ecological data is typically not rich enough (except, as suggested by Reviewer 1, the dynamics is chaotic and the whole series is contingent).

Finally, note that the time series we use are almost noiseless, so using noisier data will provide much more different estimations of the parameters, so their interpretation in microscopic terms is meaningless (as the higher experimental errors will accommodate higher variability in the set of parameters that provide curves that are close to the data).

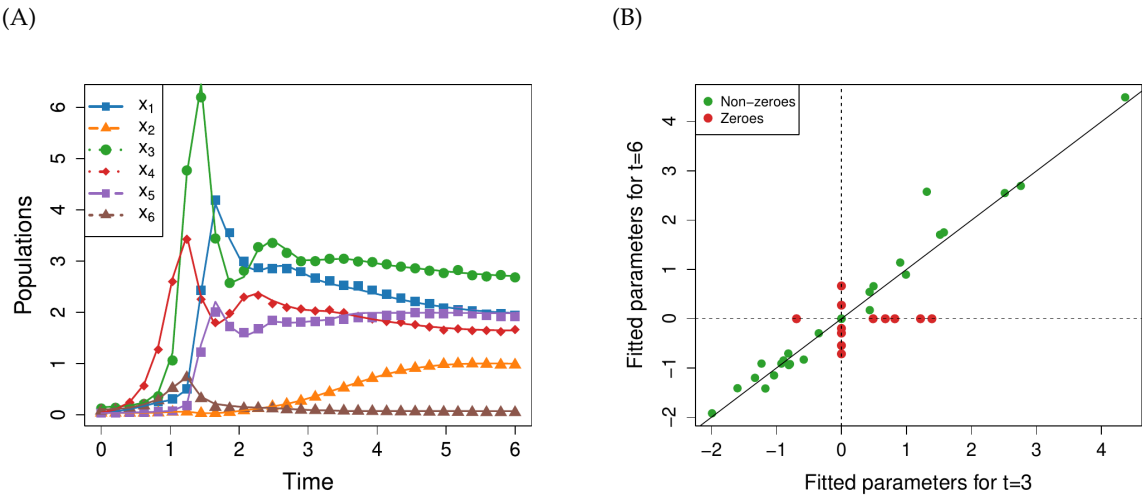

Figure S27. Extending the fitting time changes the interpretation of the inferred parameters. (A) Data generated (symbols) with the same (42) parameters as in Fig. S26 but extending the data until $t = 6$ and comparison with the fitting to a reduced model with just 30 parameters (solid lines). The agreement is excellent, considering the reduced model has 12 fewer parameters. (B) We compare the parameters of both reduced models (the ones fitted up to $t = 3$ and $t = 6$, respectively). The solid line is a reference line with slope 1. The dashed lines are a guide to the eye to identify those parameters that are, indeed, 0 in the reduced model obtained by fitting to times up to $t = 3$ and $t = 6$. The removed parameters in both models are different (shown in red). The fact that they lie on the lines and not on the origin shows that both sets of zero parameters are not identical.

References

[1]C. H. Remien, M. J. Eckwright, and B. J. Ridenhour, "Structural Identifiability of the Generalized Lotka–Volterra Model for Microbiome Studies", R. Soc. Open Sci. **8**, 201378 (2021).

[2]O.-T. Chis, A. F. Villaverde, J. R. Banga, and E. Balsa-Canto, "On the relationship between sloppiness and identifiability", Mathematical biosciences **282**, 147–161 (2016).