The Impact of Noise in Estimating Interactions of Microbial Communities

A Thesis

Presented in Partial Fulfilment of the Requirements for the

Degree of M.S.

with a

Major in Bioinformatics and Computational Biology

in the

College of Science

University of Idaho

by

Mariah Eckwright

Major Professor: Christopher Remien, Ph.D.

Committee Members: Brian Dennis, Ph.D.; James Foster, Ph.D.; Benjamin Ridenhour, Ph.D.

Department Administrator: David Tank, Ph. D.

May 2021

Authorization to Submit Thesis

This thesis of Mariah Eckwright, submitted for the degree of M.S. with a major in Bioinformatics and Computational Biology and titled "The Impact of Noise in Estimating Interactions of Microbial Communities," has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: Christopher Remien Date 4/26/2021
Christopher Remien, Ph.D.

Committee Members: Brian Dennis Date 4/21/2021
Brian Dennis, Ph.D.

James A Foster Date 2021-4-23
James Foster, Ph.D.

Date 23-Apr-2021
Benjamin Ridenhour, Ph.D.

Department Administrator: Date 4/28/21
David Tank, Ph.D.

Abstract

Recent advances in DNA sequencing have led to a boom in microbiome analyses to determine community composition. Fitting such community composition data to mathematical models allows for the estimation of interspecies interactions within a microbial community. In this thesis, we explore the extent to which noise inherent to time-series microbiome data interferes with the inference of interspecies interactions. We first create a small synthetic test community with structure mimicking real microbial communities based on the generalized Lotka-Volterra (gLV) model, incorporating differing levels of two types of noise, process and measurement noise. We then establish a method of parameter estimation for both the gLV and multivariate autoregressive (MAR) models, and apply the method to our synthetic dataset with varying levels of noise. We find that interspecies interactions can be well estimated even with moderate levels of process noise, but even modest amounts of measurement noise lead to poor estimates of interactions.

Acknowledgements

I would like to thank my advisor Dr. Christopher Remien for his guidance, support and patience during my degree process. I feel very fortunate to have worked with such a positive, optimistic individual, so eager to see me succeed. I would also like to thank my committee members, Drs. Benjamin Ridenhour, James Foster, Brian Dennis, and formerly Erkan Buzbas. The valuable input and feedback I received from them over the course of my study elevated my research project beyond my own expectations.

Finally, I would like to thank my close family and friends, who always encouraged me and applauded my successes, even when I felt like I was failing. Fifty times, thank you.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Recently, advances in 16s rRNA sequencing (RNA-Seq) have dramatically changed the way microbial communities are analyzed [1, 2, 3]. Reductions in cost and more efficient methods of sequencing have made it possible to take frequent microbial samples, creating long time-series datasets, for example those found in Koenig et al. [4] and David et al. [5]. These long-term datasets can provide insight into the ways microbiomes change over time [6, 7]. With a greater understanding of shifting microbial populations comes valuable understanding of how a microbiome can be influenced by outside factors, such as foreign bacteria or antibiotics. A more precise knowledge of the impact these outside factors can have on the human microbiota can inform medical decision-making and opens the door to treatment plans personalized to an individual's internal communities. Similar techniques might be applied to assessing or protecting the health of ecological microbiomes, such as those found in soil or large bodies of water [8, 9, 10].

A specific form of analysis that has become more tenable with higher resolution microbial read count data is interaction inference. Pairwise species interactions are estimated to gain an understanding of the overall community structure. A robust body of literature already exists seeking to meaningfully quantify microbial interactions in these newer, denser sets of read count data [11, 12, 13, 14].

The aim of this thesis is to quantify how noise interferes with microbial population inference. Chapter 2 begins with a brief overview of previous literature. We then define the models we will use for simulation and estimation, the generalized Lotka-Volterra (gLV) and multivariate autoregressive (MAR) models. Chapter 3 explains the methods used for parameter estimation. This chapter also examines the results of the model fitting experiments. The final chapter discusses the model fitting, and draws conclusions from the results.

CHAPTER 2

Simulation methods

2.1 Related work

The rapid advancement in high-throughput sequencing techniques has been accompanied by the development of a large number of methods to infer community interactions, some entirely novel and others adapted from related fields [15, 16, 17]. A common approach is to model time-series microbiome count data using a generalized Lotka-Volterra (gLV) model. Model parameters representing pairwise species interactions are then estimated by one of several different methods. Among the earliest of the gLV-based approaches is the Stein et al. [18] method, which incorporates Tikhonov regularization into the parameter estimation process. LIMITS (Learning Interactions from Microbial Time Series) [19] is another early gLV-based method, which estimates interaction parameters of a discrete gLV model through linear regression and bootstrapping. The algorithm favors sparse interaction matrices and is used by the authors to identify “keystone species” which drive community interactions.

MDSINE (Microbial Dynamical Systems Inference Engine) presents a packaged pipeline for modeling microbiome time series [12]. The algorithm estimates interaction parameters of a transformed linear version of the gLV model. Several different methods are explored for regularization of the estimated parameters, including MLCRR (Maximum-Likelihood Constrained Ridge Regression), BAPCS (Bayesian Adaptive Penalized Counts Splines), BAL (Bayesian Adaptive Lasso), and BVS (Bayesian Variable Selection).

MetaMIS (Metagenomic Microbial Interaction Simulator) infers interactions using a method based on partial least squares regression [20]. MetaMIS uses abundance-ranking based on average levels of OTUs to generate several interaction networks. Generated networks differ in the cutoff abundance for inclusion of rare OTUs.

SgLV-EKF (Stochastic gLV with extended Kalman filter) infers interactions for a stochastic version of the gLV model [21]. The authors argue that previous methods only address measurement noise in microbiome data, while ignoring the inherent noise (process noise) of the underlying system. Fitting a stochastic gLV model and using the extended Kalman filter in estimation are presented as a way to account for both measurement and process noise.

Gao et al. [22] present a method to estimate interaction parameters of a gLV model through forward stepwise regression. Overfitting is handled with a penalty term based on the Bayesian Information Criterion (BIC). Bootstrap aggregation is used to stabilize the stepwise regression.

IMPARO (Inferring Microbial interactions through Parameter Optimisation) finds gLV interaction terms through matrix optimisation [23]. The authors emphasize the value of non-unique solutions in microbial modeling.

The gLV model has seen prominent use in inferring microbial interactions. Outside of gLV-based methods, there have also been a range of other approaches taken. One of the first attempts at inferring interactions was SparCC (Sparse Corollations for Compositional Data) [24]. SparCC infers species correlation using estimated species variances. Correlation for a large system is found by iteratively calculating and removing the most strongly correlated species pairs.

The RMN (Rule-based Microbial Network) algorithm infers interactions through fitting a nonlinear regulatory OTU-triplet (NRO) model [25]. OTU triplets include a pivotal OTU, one of its competitors, and one of its cooperators. Reliability of OTU triplets is assessed with a lack-of-fit function.

ESABO (Entropy Shifts of Abundance vectors under Boolean Operations) detects competitive and synergistic links between pairs of microbial species using binarized abundance data [26]. The method compares shuffled and unshuffled species pairs to determine losses and gains in entropy, determining whether two species interact in a competitive or synergistic way.

MDeep (Microbiome-based Deep-learning) infers microbiome structure using a deep-learning-based method [27]. The authors use phylogenetic clustering as a prior to inform better predictions.

This body of work shows a wide range of approaches and innovations in microbial community analysis. Absent from the current literature is an understanding of how severely noise in time-series data may affect analysis results. This study aims to fill that gap by applying common methods of analysis to a synthetic system with a “known” solution.

2.2 Background

We start by defining the gLV model, which takes the form

$$\frac{dX_i}{dt} = a_i X_i \left(1 - \frac{X_i}{K_i}\right) + \sum_{j \neq i} b_{i,j} X_i X_j. \tag{2.1}$$

Here, X_i is the abundance of species i, a_i and K_i are its intrinsic growth rate and carrying capacity respectively, and $b_{i,j}$ is the effect of species j on the growth rate of species i. The fact that interspecies interactions are presented entirely by one parameter in the gLV model makes comparing the strength of different interactions very straightforward.

When estimating parameters in Chapter 3, we will largely focus on the parameters describing interactions between species, $b_{i,j}$ in Equation (2.1). In particular, we would like to assess the accuracy of estimated parameters compared to the “true” parameters describing a system. To this end, we would like to create a synthetic dataset representing a small microbial time series. Such a dataset—where we know the underlying parameters and what the time series should look like—will allow us to more easily find the conditions under which our methods of estimation succeed and fail. The method can be scaled up to examine larger systems in future studies.

The interactions of the gLV model can be represented by an SxS matrix β with entries $b_{i,j}$, where S is the number of species in the system. The 9x9 β matrix used to create our three-species synthetic dataset is seen in Table 2.1. Values were chosen through trial-and-error to specifically produce a system where each species oscillates before settling into a steady-state value. The accuracy of inferred parameters relies on observing changes in each species' abundance relative to one another, so a certain amount of change in abundance is preferable to species which remain always at a steady state. Additional parameters such as initial conditions and intrinsic growth rates may be found in Table 2.1. The population dynamics of the three species in the synthetic microbiome can be seen as both absolute and relative abundance in Figure 2.1.

In addition to the gLV model, we will also examine another model common in fitting population dynamics: the multivariate autoregressive (MAR) model [28, 29, 30]. Here we will focus on the one-step-ahead version of the MAR (MAR(1)), where terms at the current time step depend only on terms at the time step immediately preceding them. The MAR(1) can be written as

$$X_i(t+1) = \mu_i + \sum_{j=1}^{S} d_{i,j} X_j(t) \tag{2.2}$$

where $X_i(t)$ is the abundance of species i at time t, μ_i is the intrinsic growth rate of species i, and $d_{i,j}$ is the effect of species j on the growth of species i. Of particular interest is the MAR(1)'s application to stability analysis, where model parameters may be used to assess how a community responds to stimuli which push the system away from stationarity [31]. Community restabilization and response to outside stimuli are topics of great interest in microbial community studies. The potential application to community stability analysis motivates the choice of the MAR(1) model.

2.2.1 Process noise: stochastic differential equations

Equation (2.1) can be used to produce stable synthetic time-series data in which species neither go extinct (or nearly extinct) nor do they diverge towards infinity. However, the synthetic data generated up to this point are completely deterministic. Microbiome read count data is inherently noisy for numerous reasons, and we want to reflect these multiple

Parameter	Value
$X_1(0)$	500
$X_2(0)$	400
$X_3(0)$	270
a_1	6
a_2	4
a_3	2
K_1	120
K_2	150
K_3	135
$b_{1,2}$	0.15
$b_{1,3}$	-0.2
$b_{2,1}$	-0.01
$b_{2,3}$	0.05
$b_{3,1}$	0.1
$b_{3,2}$	-0.1

Table 2.1: **Table of parameters used to generate the basic three-species simulated system.**

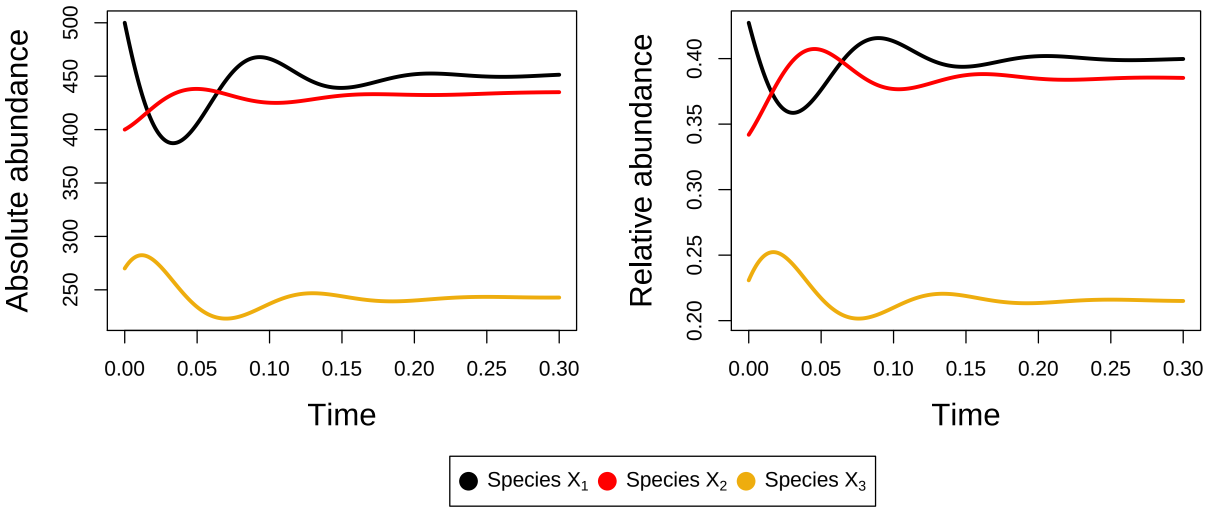

Figure 2.1: **Plots showing population counts over time for absolute (left) and relative (right) abundance.** Absolute abundance is a plot of the actual observed counts for each species, while relative abundance is the abundance of each species relative to one another at each time point.

sources of noise in our synthetic datasets. To do so, we will employ methods to include both process noise and measurement noise in our simulated datasets.

We first modify the system to include process noise, which describes how random variations in the underlying processes that influence population sizes affect observed trajectories. Stochasticity of this nature can be included in many different ways, depending on assumptions of its source (e.g., demographic, environmental). As a simple way of modifying Equation (2.1) to include environmental variability, we introduce the stochastic differential equations

$$dX_i(t) = \left(a_i X_i(t) \left(1 - \frac{X_i(t)}{K_i} \right) + \sum_{j \neq i} b_{i,j} X_i(t) X_j(t) \right) dt + \sigma_i X_i(t) dW_i(t), \tag{2.3}$$

where σ_i determines the magnitude of the process noise, and $dW_i(t)$ is standard Brownian motion $N(0, \sqrt{dt})$ [32]. We simulate sample paths of Equations (2.3) using the `pomp` and `simulate` functions in the R package `POMP` [33, 34].

2.2.2 Measurement noise: error from sampling

Measurement noise is introduced to read count data during data collection. At each time point, there is a true abundance of each species in a system. In our synthetic datasets this is found by simulating Equations (2.3). In biological datasets, a number of reads which subsets the total population is taken at each time point, which may skew the actual population values. To include measurement noise in our simulated data in a manner that is similar to the natural measurement process, we model the counts of each species as a multinomially-distributed draw from the true population abundances. For each time point, the observed number of reads of each species (X_i) is given by

$$\begin{gathered} (x_1, x_2, ..., x_S) \sim Mult\,(V; p_1, p_2, ..., p_S) \\ p_i = \frac{\hat{x}_1}{\sum_{i=1}^{n} \hat{x}_i} \end{gathered} \tag{2.4}$$

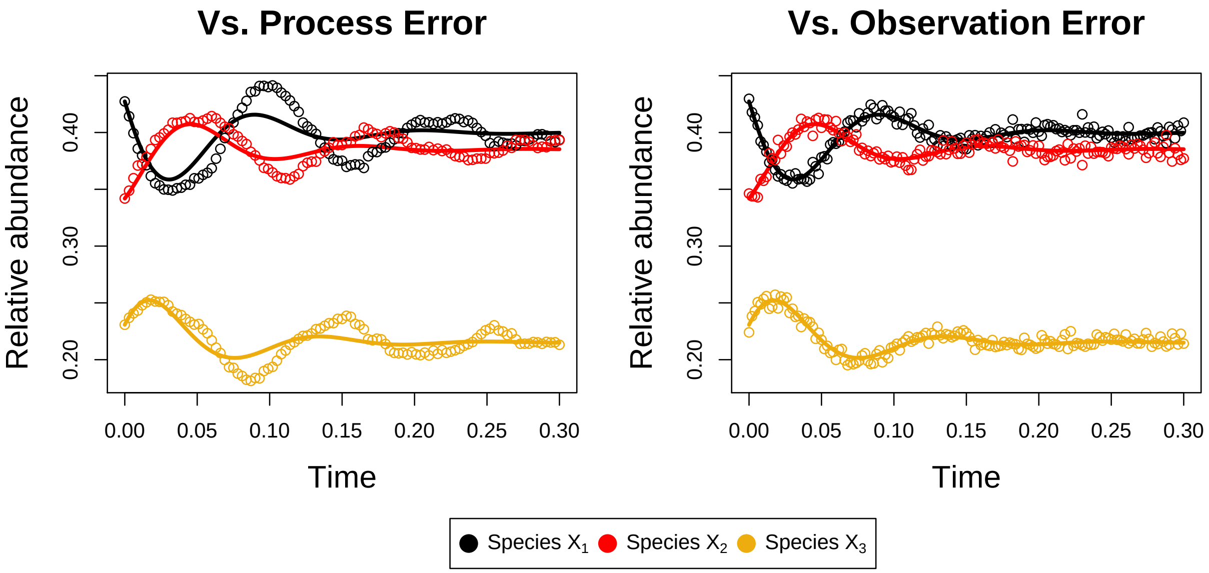

Figure 2.2: **Deterministic and stochastic series of relative abundance counts.** The plots show the same relative abundance data as in Figure 2.1, with the inclusion of process noise (left) and measurement noise (right).

where $\hat{x}_i$ is the absolute abundance of species i after the inclusion of process error, and V is the total number of reads sampled. As V increases, we expect the spread of total reads to more accurately reflect the actual distribution of species and decreasing overall noise from the measurement process.

In this section, we applied established methods of ecological modeling to create a small synthetic microbial system. Going forward, we will now apply different methods of analysis to this synthetic system under a variety of circumstances (e.g. varied noise levels) to investigate the limits of the analysis methods.

CHAPTER 3

Parameter Estimation Methods

3.1 Model comparison

3.1.1 Relating gLV and MAR parameters

Like the gLV model, information about the strength of pairwise interactions in the MAR model is limited to a single parameter. The parameters $d_{i,j}$ from Equations 2.2 are equivalent to the gLV model's $b_{i,j}$. A comparison of inferred parameters of the two models could provide valuable insight into the performance of both models. However, the models assume different functional forms, preventing a direct comparison of interaction parameters. To relate the gLV and MAR interaction parameters, we must assume the gLV is at a steady state and find a linearized, discretized version of the model at its endemic equilibrium, producing a MAR(1) model. We start the transformation by writing the gLV model in Equation (2.1) as

$$\frac{dX_i}{dt} = a_i X_i + \sum_{j=1}^{S} b_{i,j} X_i X_j \tag{3.1}$$

where $b_{i,i} = -a_i/K_i$. To linearize (3.1), we first find the Jacobian matrix (J) with entries

$$\begin{aligned} J_{i,j} &= b_{i,j} X_i \\ J_{i,i} &= a_i + 2b_{i,i} X_i + \sum_{j \neq i} b_{i,j} X_j. \end{aligned} \tag{3.2}$$

With J evaluated at the steady state $\bar{X}$, this leads to the linear system of differential equations

$$\frac{d\Delta X}{dt} = J|_{\bar{X}} \Delta X. \tag{3.3}$$

Discretizing system (3.3) with time step h gives the difference equation

$$\frac{\Delta X(t+h) - \Delta X(t)}{h} = J|_{\bar{X}} \Delta X(t) \tag{3.4}$$

which can be rearranged to the form

$$X(t+h) = X(t) + hJ|_{\bar{X}}\Delta X(t). \tag{3.5}$$

Note that Equation (3.5) is itself a MAR(1) model. Matching coefficients in Equations (3.5) to their counterparts in Equations (2.2) yields

$$\begin{aligned} \mu_i &= 0 \\ d_{i,j} &= hb_{i,j}\bar{X}_i, & i \neq j \\ d_{i,j} &= 1 + h\left(a_i + 2b_{i,j}\bar{X}_j\right), & i = j. \end{aligned} \tag{3.6}$$

Thus, at steady state (or in stationarity) Equations (3.6) relate the parameters of a gLV model describing the abundance of species i to those of a MAR model describing the change in species i over discrete time. Note that we cannot directly obtain unique gLV parameters given MAR parameters, but the converse is possible. In particular, the pairwise interaction terms $d_{i,j}$ where $i \neq j$ give the per capita force of interaction of species j on the growth of species i while the pairwise interaction terms $b_{i,j}$ give the per capita force of interaction per species i of species j on the growth of species i.

3.1.2 The relative abundance case

Following the methods in section 3.1.1, the interaction parameters $b_{i,j}$ of the gLV model can be related to the interaction parameters $d_{i,j}$ of the MAR model, but are scaled by the mean absolute abundances of each species. Repeating the same process as in section 3.1.1 for the relative abundance equation gives

$$
\begin{aligned}
\mu_i &= h\left(a_i\bar{x}_i + N\sum_{j=1}^{S} b_{i,j}\bar{x}_i\bar{x}_j\right) & \\
d_{i,j} &= Nhb_{i,j}\bar{x}_j, & i \neq j \\
d_{i,j} &= h\left(a_i + Nb_{i,j}\bar{x}_i\right) + 1, & i = j
\end{aligned}
\tag{3.7}
$$

where N is the total abundance, and $\bar{x}_i = \frac{\bar{X}_i}{N}$. Because this analysis assumes that the population is near its endemic equilibrium, N is approximately constant and only has a scaling effect on the relations. Therefore, with relative abundance data, relative interaction rates can be estimated. The pairwise interaction terms $d_{i,j}$ with relative abundance data give the per capita force of interaction of species j on species i times the total population size, which is typically an unknown parameter.

3.1.3 Background on regularization

The process of fitting a model to read count data is typically complicated by the overabundance of parameters relative to measurements. Specifically, the number of parameters will be on the order of the square of the number of species. In such an underdetermined system, the model is highly prone to overfitting. An overfit model can fail to capture the underlying structure of a system by being too greatly influenced by the single realization from which the model parameters are estimated. To counteract overfitting, we employ regularization.

Regularization reduces model overfitting by penalizing model complexity. One of the most common forms of regularization is ridge regression, which reduces unwarranted complexity by penalizing the l_2 norm of the estimated parameters [35]. In particular, the best fit parameters with ridge regression solve the equation

$$\hat{b} = argmin_b(||y - xb||_2^2 + \lambda||b||_2^2) \tag{3.8}$$

where x is a vector of predictor variables, y is a response matrix, $||.||_2^2$ denotes taking the l_2 norm, and $\lambda \in [0, \infty]$ determines the strength of regularization. Ridge regression decreases the magnitude of estimated parameters, but will not reduce any parameters to zero. An over-parameterized model will be more complex than necessary, and may over-describe the data.

To reduce the number of parameters, we consider a second type of regularization: LASSO regression. Similar to ridge regression, the Least Absolute Shrinkage and Selection Operator (LASSO) also penalizes overly complex solutions, but uses the l_1 norm $||.||_1$ as the penalty in place of the l_2 norm, to impose feature selection in estimation [36]. The interactions of different communities may be more accurately captured with ridge or LASSO regression. Instead of choosing one form of regularization, both types of regularization can be blended in elastic-net regularization.

3.1.4 Application of the elastic net

The elastic net finds the best fit parameters $\hat{b}$ as

$$\hat{b} = argmin_b(||y - xb||_2^2 + \lambda[\alpha||b||_2^2 + (1 - \alpha)||b||_1]) \tag{3.9}$$

where x is the observed data, λ controls the strength of regularization ($\lambda = 0$ being equivalent to the case with no regularization), and $\alpha \in [0, 1]$ controls the strength of ridge versus LASSO regression, with $\alpha = 1$ corresponding to pure ridge regression and $\alpha = 0$ using only the LASSO [37].

Elastic-net regression can improve the fit of a model, but only with appropriate values for α and λ. To find the optimal α and λ, we employ K-fold cross validation with 10 folds. In K-fold cross validation, the data are first split into K groups of equal size. For each subset (fold) of data, a model is trained on all the data outside the group. Each model is then tested against the remaining data, and the mean square error (MSE) is recorded. The values for α and λ that minimize the MSE are then selected.

To perform cross validation and parameter estimation for the synthetic data, we use the R package `glmnet` [38]. For each species in a system, we perform a grid search of α and λ values using `cv.glmnet`. The search returns the value of α which minimizes the mean cross-validated error of the system. This value of α, the time-series data, and a sequence of associated λ values are passed to the function `glmnet`, which returns estimates for interspecies interactions $b_{i,j}$ and intrinsic growth rates a_i.

With the default settings, the functions `cv.glmnet` and `glmnet` choose a sequence of log-spaced λ values based on the provided data. From this sequence, both functions choose an optimal value of λ based on the amount of deviance explained by a specific fit. If the deviance explained reaches a certain threshold or does not drastically change between two values of λ, `glmnet` determines that it has reached an acceptable value of λ. We found that, in relation to the three-species system described by the parameters given in Tables 2.1, the default thresholds for deviance explained often led to `glmnet` picking the smallest generated λ. In cases where the smallest λ was selected, rerunning the same scenario while supplying smaller λ values to `glmnet` resulted in smaller values of λ being selected. This suggests that the sequence of λ values generated by `glmnet` does not cover a wide enough range for our purposes. To address this, we ran `glmnet` with a self-supplied sequence of 100 log-spaced λ values, starting with the largest λ generated by `glmnet` and ending at 10^{-5}.

3.1.5 Results of Estimation with the gLV Model

We estimated interaction parameters using two types of synthetic data and two underlying models. The two types of synthetic data examined are the same system with read counts given as absolute and relative abundance. The two underlying models that the synthetic data were compared to the gLV (Equation 2.1) and the MAR(1) (Equation 2.2) models as described in Chapter 2. Time-point density, level of process error and level of measurement error were all adjusted separately between runs of the models. Process error was varied with the value of σ_i in Equation (2.3), while measurement error $\xi = \frac{1}{V}$ was varied by changing the total number of sampled reads V, as in Section 2.2.2. The values used fell in the ranges $0 < \sigma_i < 5$ and $10^3 < V < 10^7$. To show that comparable levels of noise were being contributed to the synthetic data by both process and measurement error, 30 simulations were run for each combined level of noise. MSE between the noiseless system and simulated noisy systems were then found for each species. The average MSE for all three species was recorded at each level. Figure 3.1 shows that the ranges of V and σ_i result in similar MSEs from process and measurement noise.

gLV-based Estimation with Absolute Abundance Data

We applied the gLV-based method of estimation in Section 3.1 to the simulated three-species system first described in Section 2.2. The accuracy of estimates of $b_{i,j}$ depends critically on the noise in the data. Low levels of both process and measurement error are required for accurate inference. The number of measurement samples required depends on the research question at hand and the percent error tolerated in estimates. Specifically, if parameters need to be estimated to within 25% of their true value, then our three-species community requires at least 5×10^6 measurement samples at each time point and a process error of less than 0.3. If only order of magnitude estimates are required (within 100% error), then this relaxes to 1×10^6 measurement samples, and any level of process error in the range tested.

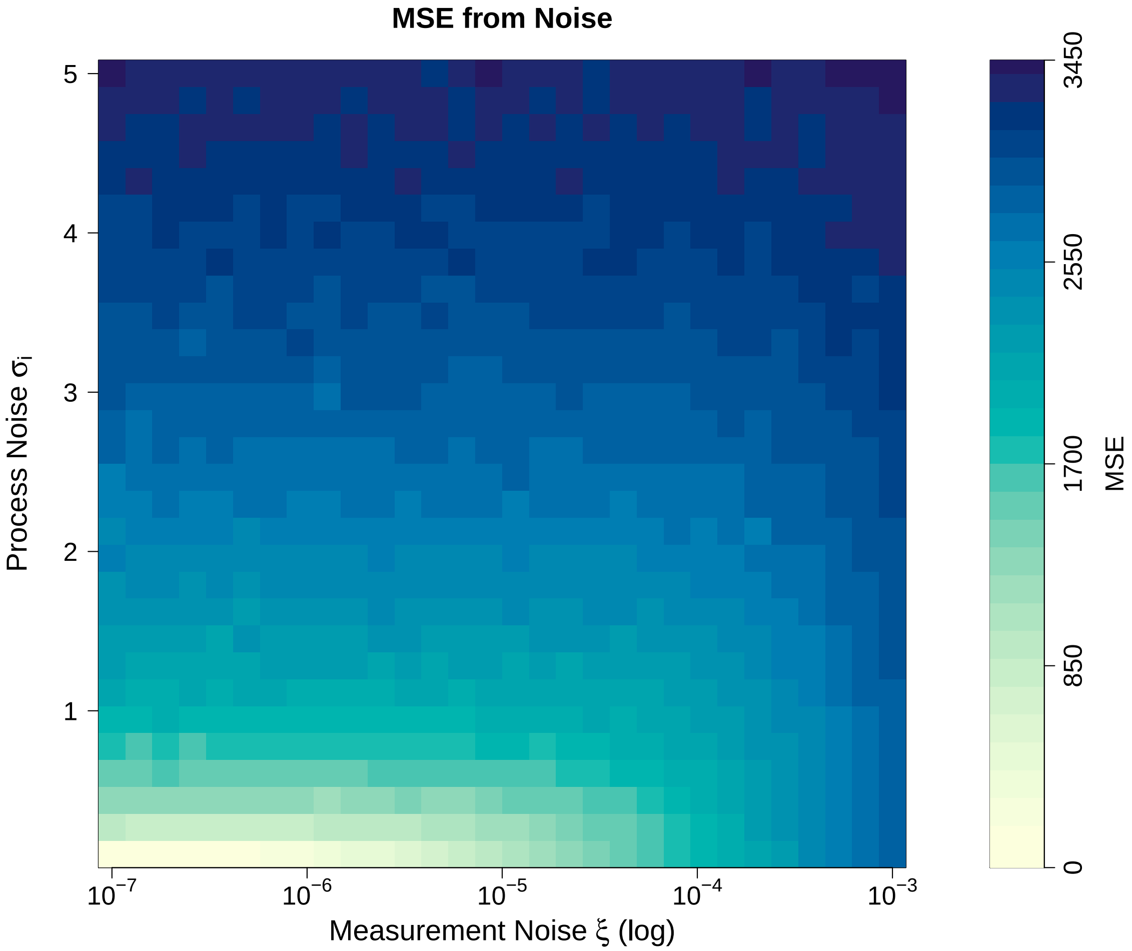

Figure 3.1: **MSE in synthetic data from differing levels of process and measurement noise.** Thirty synthetic systems were created for each combination of process and measurement noise. Average MSE of all three species in each system was found and recorded.

Process Error had a much less pronounced effect on the accuracy of estimated interaction parameters. While there was a slight increase in median percent error as process error increased, the effect was small. The estimated β matrix had the correct signs for every $b_{i,j}$ at lower error levels, but consistently had $2 - 3$ mismatched signs in interspecies interaction terms when ξ was above 10^{-4}.

To assess the accuracy of the inferred temporal dynamics, we used the estimated β matrices to reconstruct time-series of microbial abundances and compared them to the original synthetic data. Time-series were reconstructed using inferred interaction parameters in two ways. The first was to numerically solve Eqn. (2.1) using inferred parameters starting at the true initial values (long-range prediction). The second way was as one-step-ahead prediction—that is, given the abundance of species X at time $t - 1$, predict the abundance at time t using Eqn. (2.1) (short-range prediction).

Estimated time-series were compared to simulated time-series using mean squared error (MSE), calculated as

$$\frac{\sum_{i=1}^{T}(x_i - \hat{x}_i)^2}{T} \tag{3.10}$$

where x_i is the simulated abundance for species x at time i, $\hat{x}_i$ is the estimated abundance of species x at the same time point, and T is the total number of time points in the series. In cases where simulated data included only one type of error, long-range prediction showed that MSE increased with the level of error in the data (Figure 3.3). This was true for both process and measurement error. The median MSE in simulated time-series with process error increased as σ_i increased (Figure 3.3 Panel A). The opposite was true for measurement samples, where median and variance increased as the number of samples decreased (Figure 3.3 Panel B). MSE variance at high sampling rates or low process error were similar in magnitude ($\sim 10^{-2}$), but at higher levels of noise simulated data with process error had considerably more variance in MSE (10^5 with process error versus 10^4 with measurement error). Across all predictions, the MSE for reconstructions of species X_1 were higher than species X_2 and X_3, possibly owing to the overall higher species counts for species X_1.

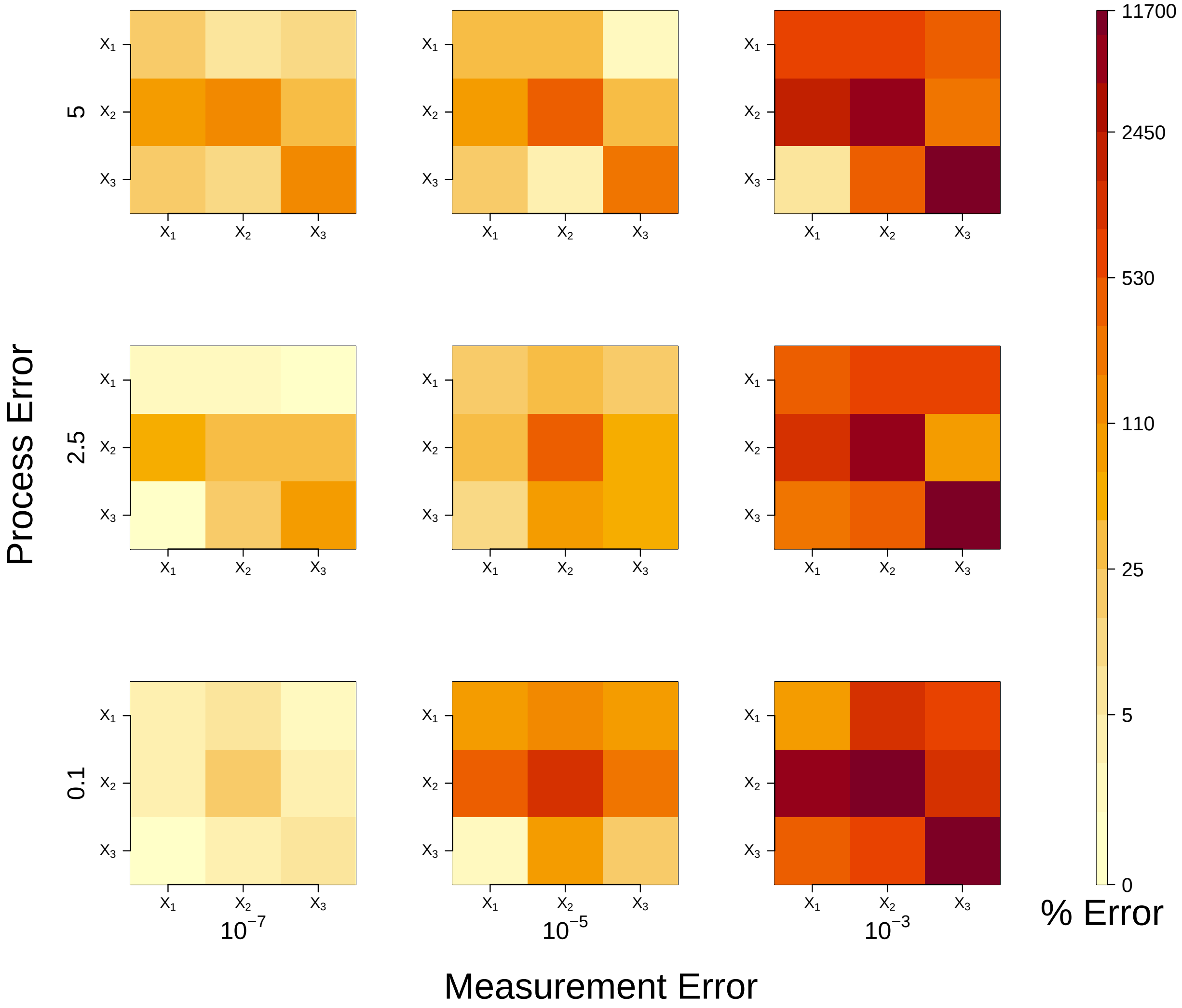

Figure 3.2: **Median error in estimated $b_{i,j}$ for systems with varying levels of noise.** The figure shows the percent error between values of $b_{i,j}$ used to simulate time-series abundance data and estimations of each $b_{i,j}$ found using the gLV model. The simulation was run at each level of process and measurement error 10 times.

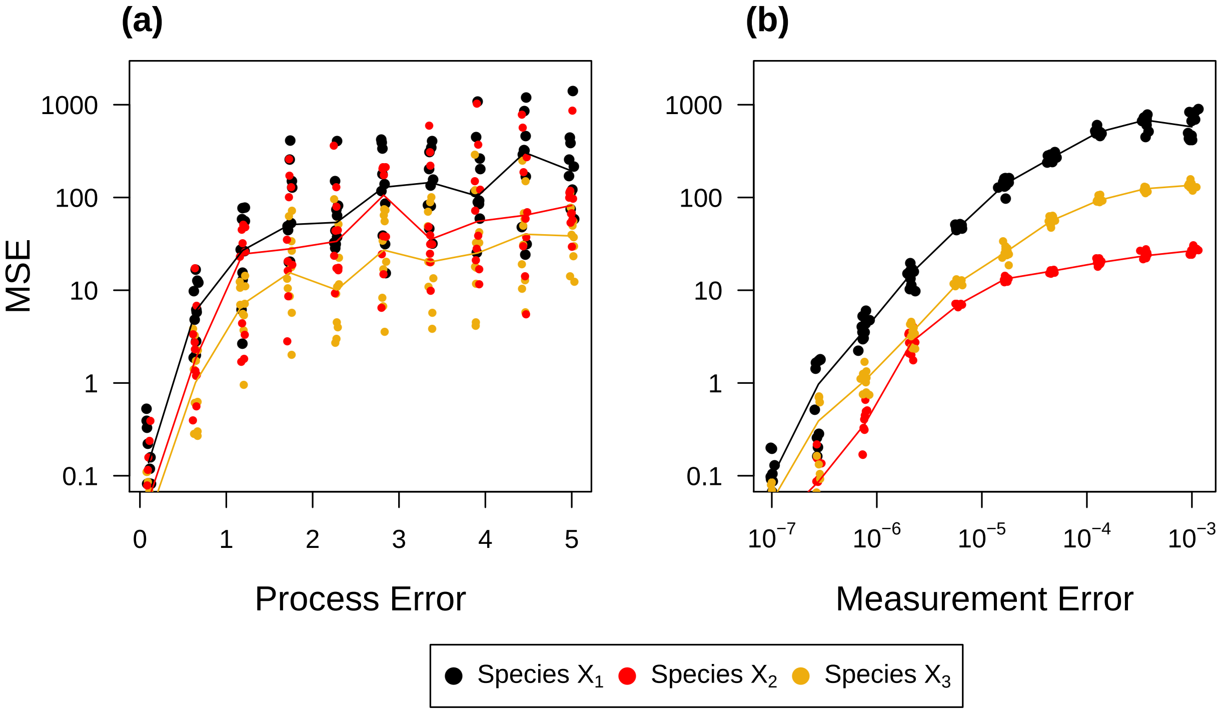

Figure 3.3: **Mean-square error in estimated time-series derived from data with only one type of error.** Parameters were estimated by using the gLV method on simulated time-series data. For this particular test, data were simulated with only process (a) or measurement (b) error.

In applying long-range prediction to simulated data with both process and measurement error, reconstructions were only accurate at low levels of both error types. For example, species X required at least 2×10^5 measurement samples at each time point and $\sigma_i < 2$ to consistently produce estimates with MSE below 40. As process error increased, the estimated time series had qualitatively similar trajectories, but tended to drift away from the true mean (Figure 3.4). High measurement error caused estimates to lose information about dynamics, but were more likely to stay close to the true mean. However, high measurement error also occasionally caused the reconstructed time series to fail completely (black points in Figure 3.5), and not converge to a steady state, thus giving MSE values in the range of 10^{20}. High levels of both process and measurement error pushed estimated time-series away from the true mean and failed to capture information about dynamics.

gLV-based Estimation with Relative Abundance Data

Interaction parameters estimated from relative abundance time series data were evaluated in the same way as absolute abundance estimates. When fitting relative abundance data with a gLV model, parameter estimates were poor (Figure 3.6), although the signs of estimated parameters were correct across all noise levels. Regardless of noise level, estimated parameters had error levels greater than $4,000\%$. In each case, three parameters were considerably more accurate in their estimates than the other parameters, with a percent error on the order of 100. These considerably-more-accurate estimated parameters may be a result of the identifiability of relative abundance gLV model, which implies that interaction parameters can only be estimated relative to each other [39]. When only one type of noise was present, time-series reconstructions from gLV-based relative abundance data were accurate at low noise levels, and lost accuracy with increased noise (Fig. 3.7). In the presence of only process error, reconstructions averaged low levels of MSE at all noise levels. High-MSE outliers were more likely in runs at higher levels of process noise, particularly for species X_1. Measurement noise had consistently accurate estimates when $\xi < 10^{-5}$. Between 10^{-5} and 10^{-4},

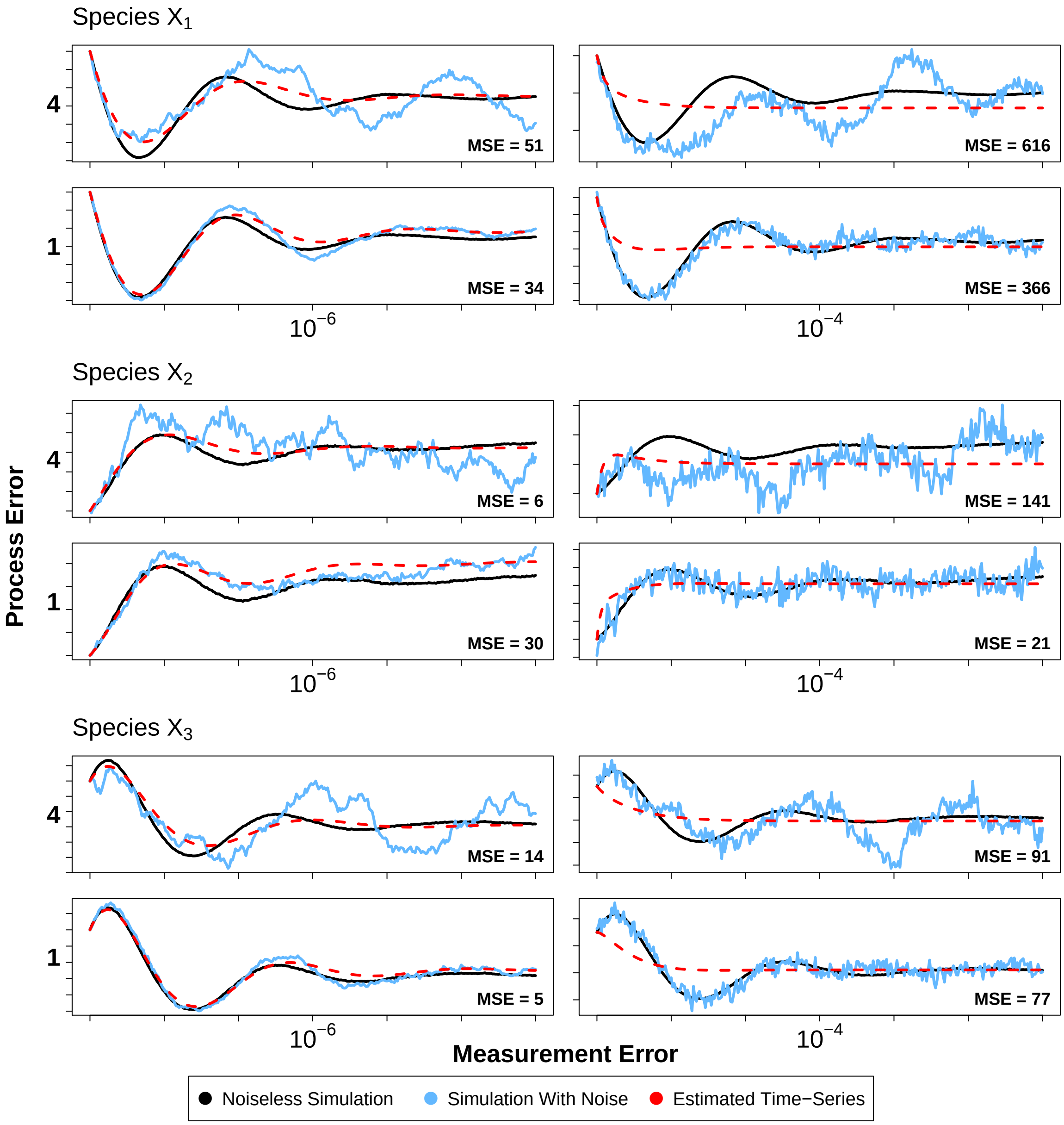

Figure 3.4: **Comparison of time-series reconstructed from gLV-estimated parameters.** For each species, four sample paths are presented with high and low levels of process and measurement error.

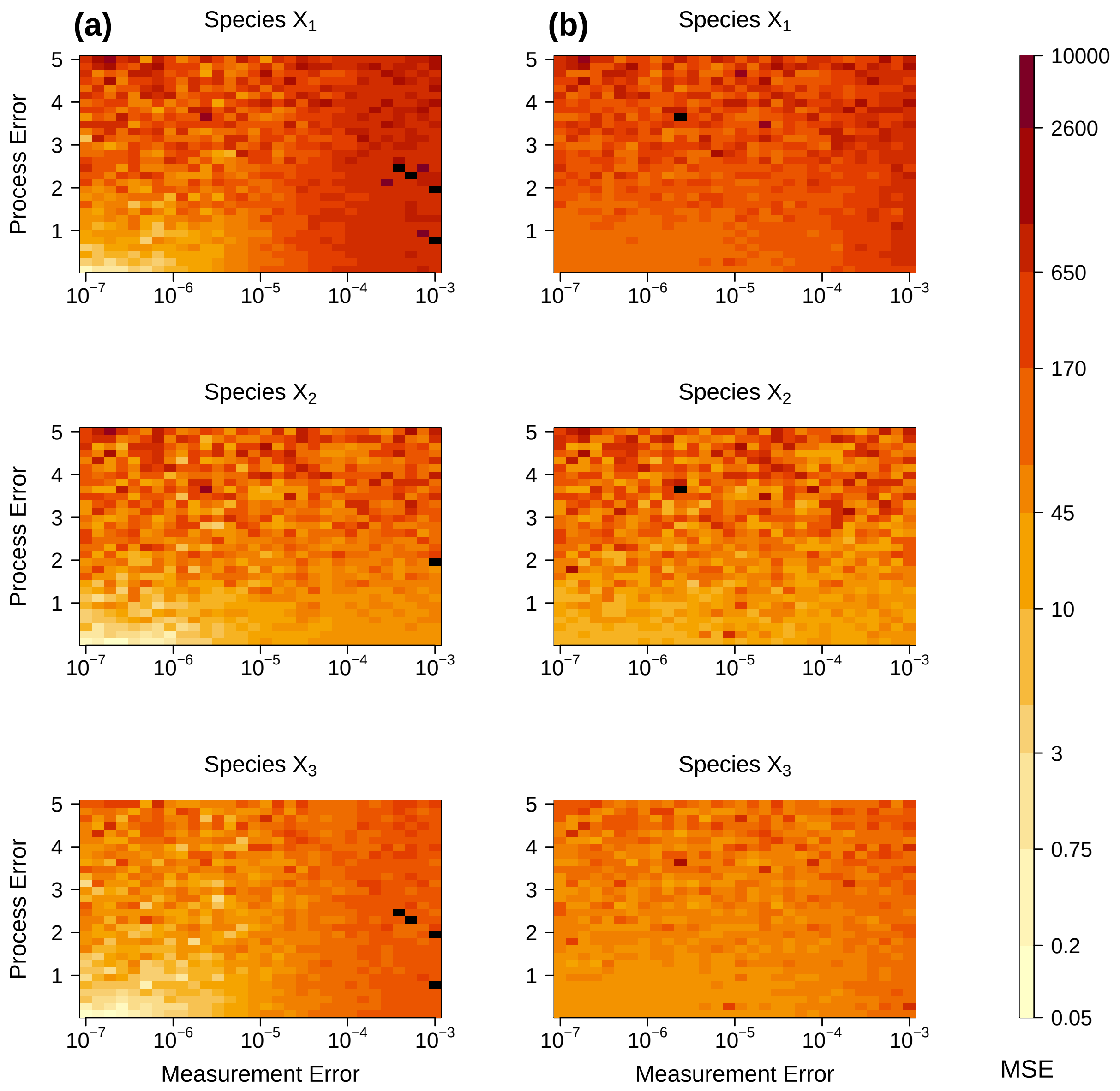

Figure 3.5: **Comparison of reconstructed time-series with differing numbers of time points; gLV model.** Each heatmap shows the MSE for reconstructions of each species at varying levels of process and measurement error. Column (a) was estimated from a time-series with observations taken every 0.001 steps, while (b) has observations every 0.01 steps.

there is a sharp spike in MSE, followed by leveling off again. This seems to be indicative of the model failing to handle larger values of ξ, which can also be seen in the reconstructed time-series themselves (Fig. 3.8). Indeed, Fig. 3.8 the reconstructed time-series show that in cases where $\xi = 10^{-4}$, reconstructed time-series level off at the mean almost immediately, accounting for the relatively low MSE. Reconstructions were much more accurate at lower values of ξ, where they were able to follow the trajectory of their respective species fairly closely regardless of process noise level. These reconstructions, however, consistently trailed off rather than finding the equilibrium.

3.1.6 Results of Estimation with the MAR Model

MAR-based Estimation with Absolute Abundance Data

Parameters estimated using the statistical MAR model followed similar trends to those estimated with the gLV model, with some notable exceptions. The MAR model seemed unable to estimate self-interaction terms regardless of error rates (Fig. 3.10). However, estimated self-interaction terms were consistent, with variance between runs often in the range of 0.001-0.1. Off-diagonal parameter estimates had higher percent error than those found with the gLV model. Estimated self-interaction terms consistently had the wrong sign, while estimated interspecies interactions had generally correct signs. When ξ was greater than 10^{-4}. Overall, parameter estimates from both models were more resilient to increases in noise from process error than measurement error.

Similarly to the gLV estimates, we also evaluated MAR-estimated parameters through the comparison of reconstructed time-series (Fig. 3.11; Fig. 3.12). When the initial time-series had high levels of either process or measurement noise ($\sigma_i > 4; \xi > 10^{-4}$), MAR-estimated time-series were roughly as accurate as gLV-estimated time-series. When time-series were reconstructed from parameters estimated from noiseless data, MSE fell in the range of 10-50, roughly ten units higher than noiseless estimates from the gLV model.

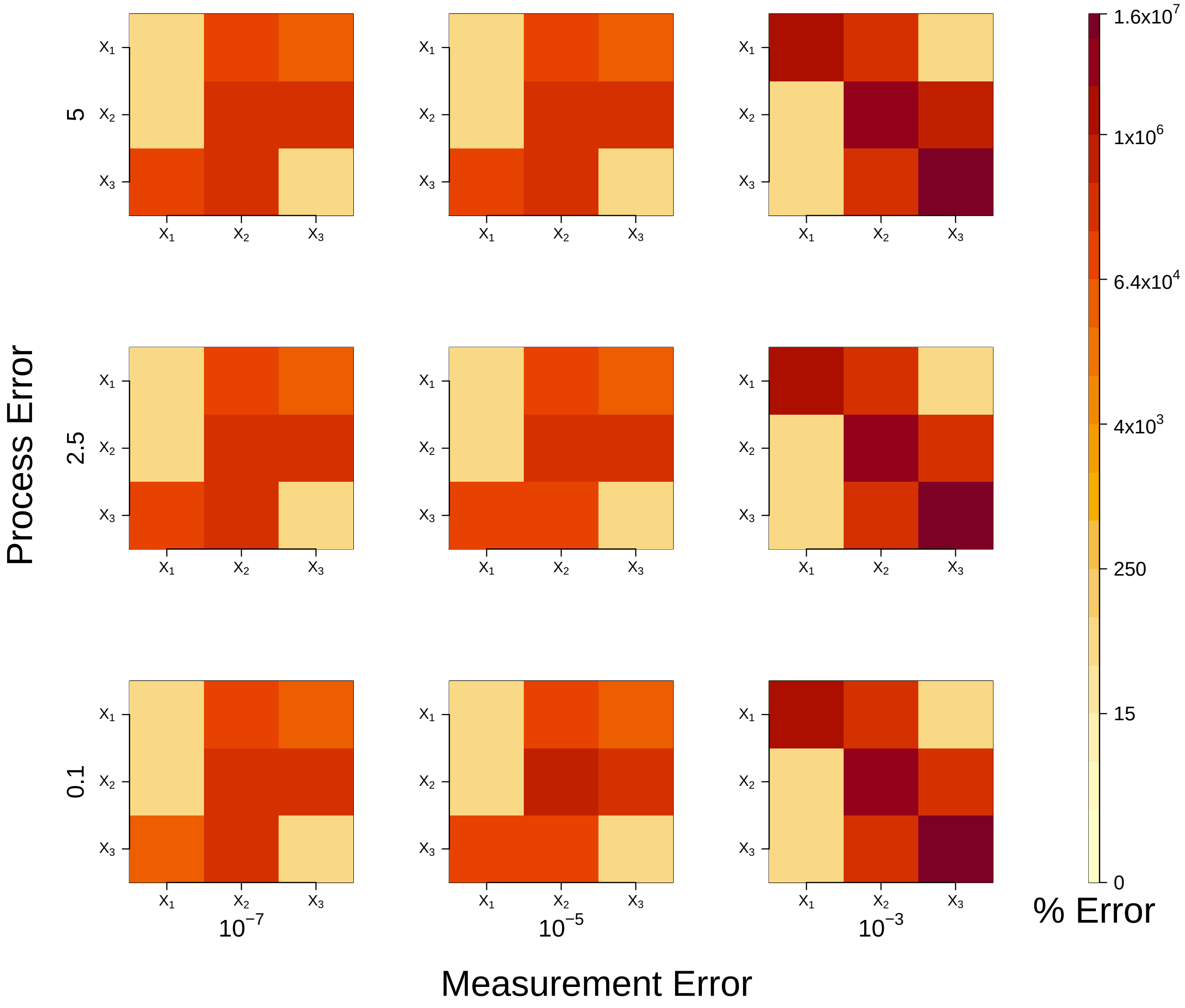

Figure 3.6: **Median error in estimated $b_{i,j}$ for relative abundance systems with varying levels of noise.** The figure shows the percent error between values of $b_{i,j}$ used to simulate time-series relative abundance data and estimations of each $b_{i,j}$ found using the gLV model. The simulation was run at each level of process and measurement error 10 times.

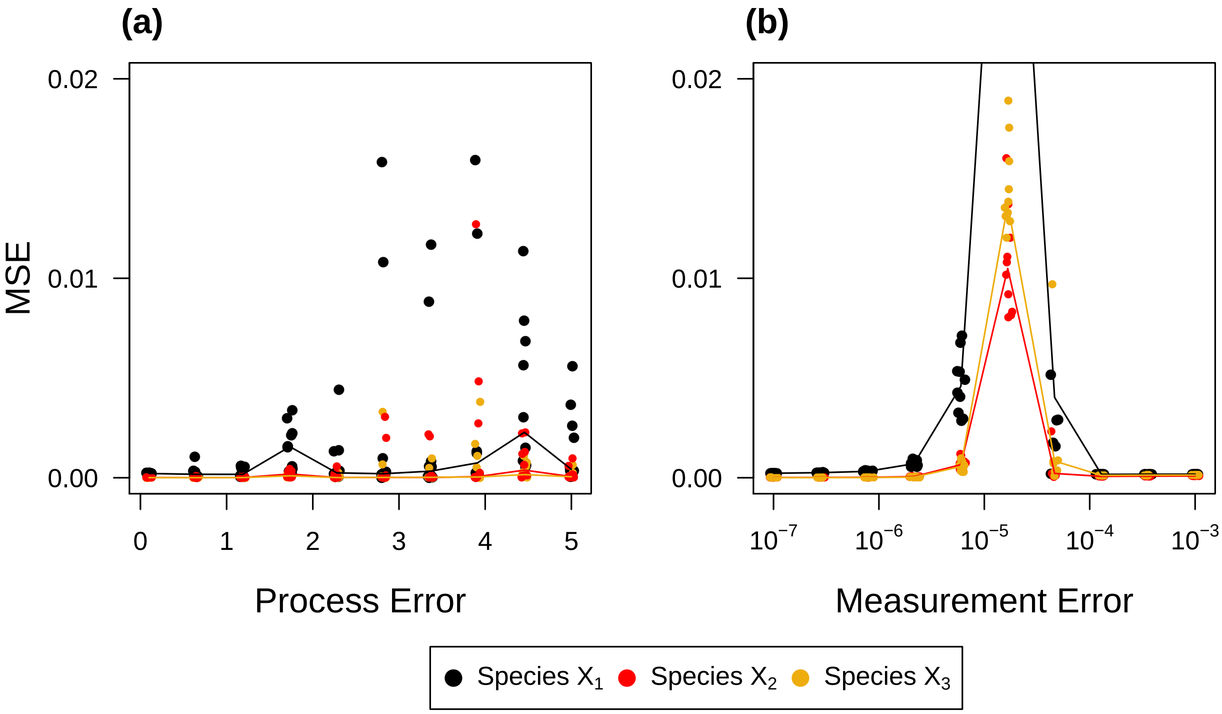

Figure 3.7: **Mean-square error in estimated relative abundance time-series derived from data with only one type of error.** Parameters were estimated by using the gLV method on simulated time-series data. For this particular test, data were simulated with only process (a) or measurement (b) error.

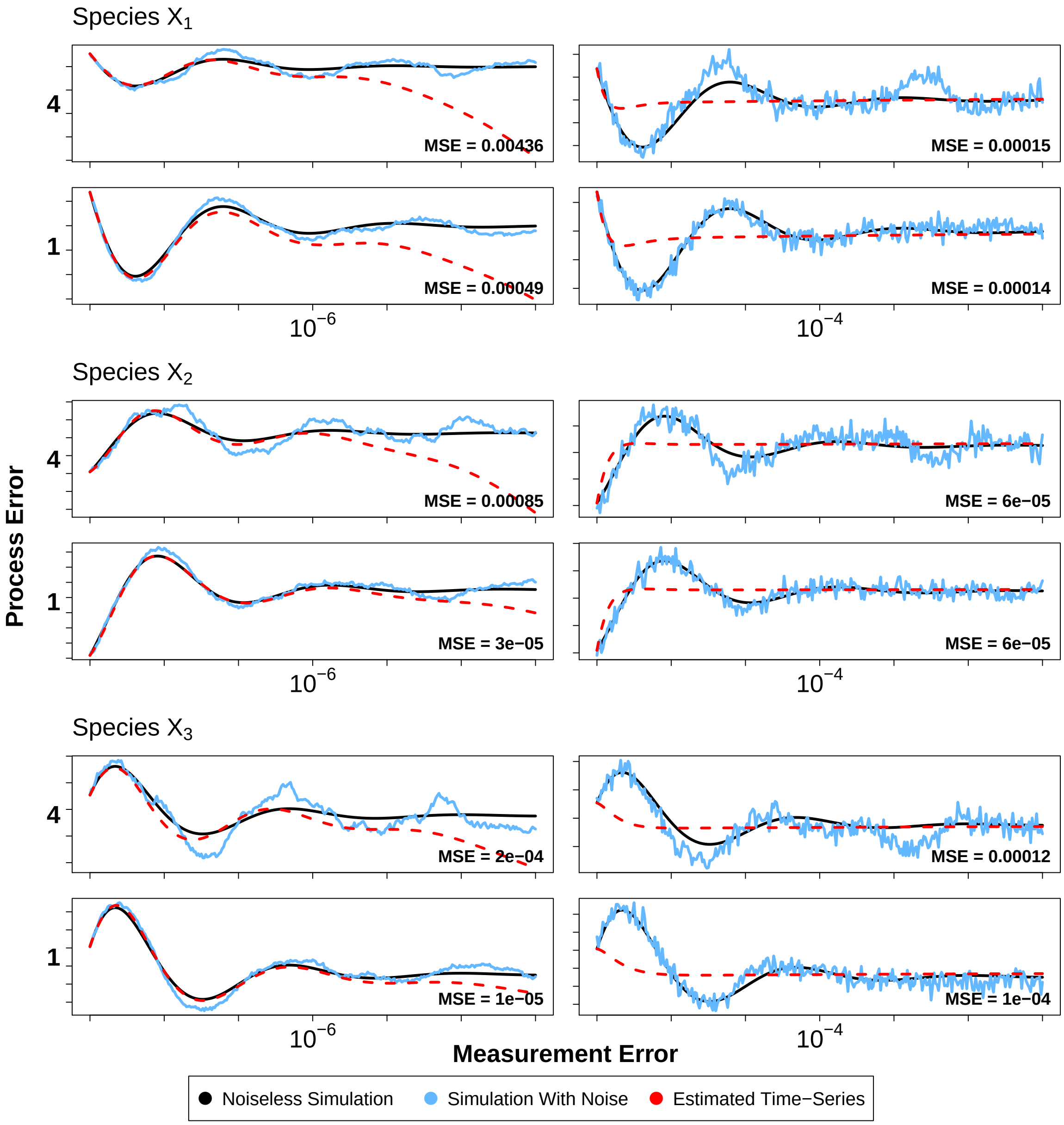

Figure 3.8: **Comparison of relative abundance time series reconstructed from gLV-estimated parameters.** For each species, four sample paths are presented with high and low levels of process and measurement error.

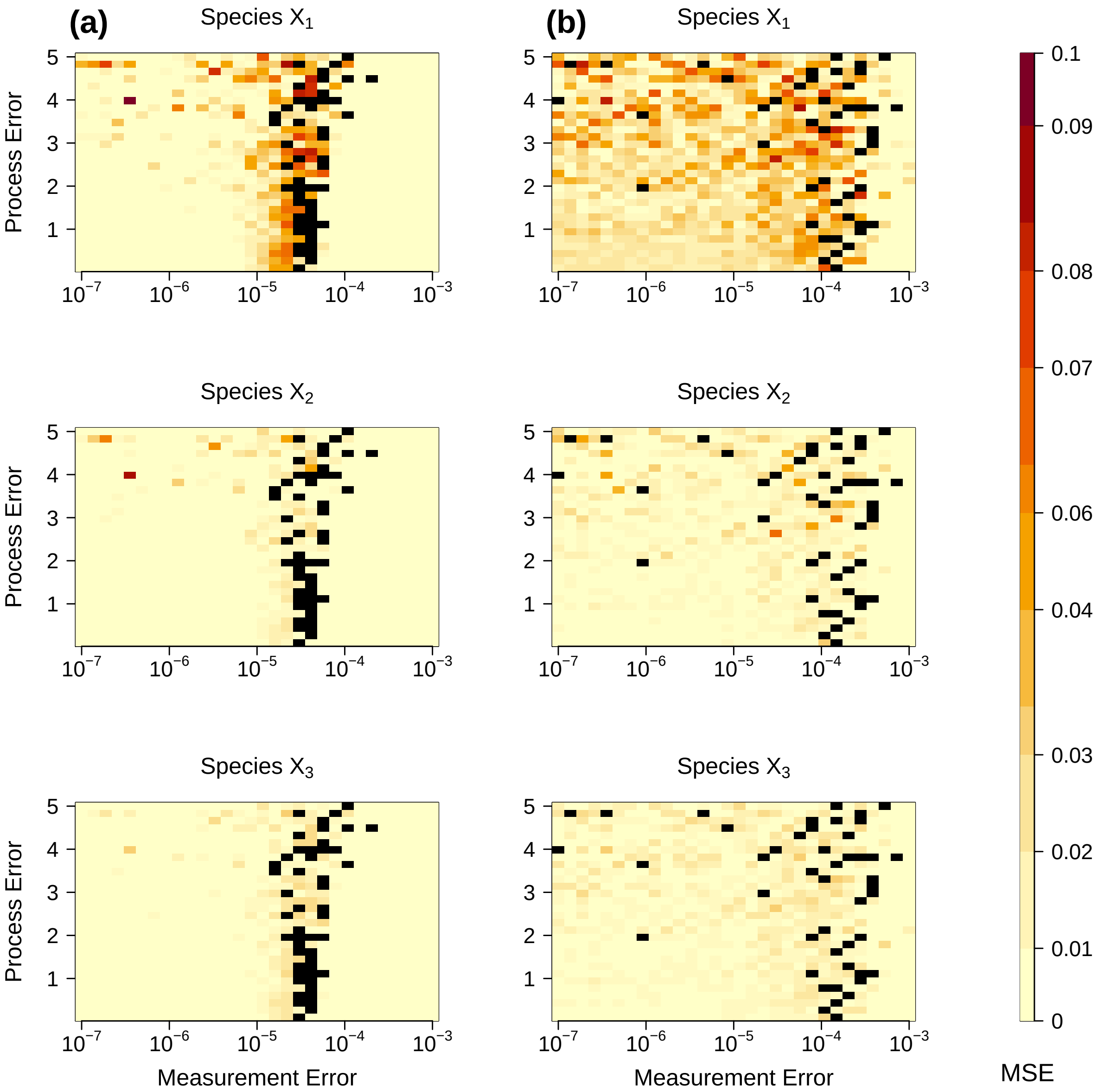

Figure 3.9: **Comparison of reconstructed time-series with differing numbers of time points; relative abundance gLV model.** Each heatmap shows the MSE for reconstructions of each species at varying levels of process and measurement error. Column (a) was estimated from a time-series with observations taken every 0.001 steps, while (b) has observations every 0.01 steps.

Unlike the gLV-estimated time series, MAR time series exhibited a small, consistent dip in MSE before steadily rising alongside increasing noise in the initial data. This appears to be a consequence of assuming noise in the data when noise levels are in fact very low ($\sigma_i < 1; \xi < 10^{-6}$). Further, while Fig. 3.11 seems to show a large dip in MSE, this is mostly a consequence of logarithmic scaling. At the low end of noise values, all but one reconstructed MAR species trajectory had a median MSE value less than 10. The odd trajectory out, species X_1 with only process error, had a low-noise median of 22.3. None of the reconstructed trajectories produced MSEs above 70 with σ_i and ξ in the above-specified range.

Time-series constructed from estimated MAR parameters were also less resilient to reductions in time point density than their gLV-based counterparts (Fig. 3.13). MAR parameters returned MSE levels on par with gLV parameters at high time-point density (step size 0.001 over a period of 0.3 units of time for a total of 300 points). When time point density was reduce to a step size of 0.01 (30 total points), however, no constructed MAR time-series had a MSE lower than 600%, regardless of noise in the system.

MAR-based Estimation with Relative Abundance Data

Similarly to the absolute abundance case, interaction parameters estimated using the MAR model with relative abundance data yielded values for $b_{i,j}$ with high levels of MSE (Fig. 3.14). The average percent error in estimated interaction parameters was over 100% for off-diagonal terms, and over 500% for values of $b_{i,i}$. Estimated self-interaction terms consistently had the wrong sign, while interspecies interaction terms had the correct signs at all noise levels.

When only one type of noise was present, time series reconstructions were fairly accurate (MSE $< 1.25 \times 10^{-4}$) when process error σ_i was less than 3 or measurement error ξ was less than 10^{-5} (Fig. 3.15. MAR-based estimates of relative abundance time series data did not produce the same spike in MSE at intermediate noise levels that was seen in MAR-based

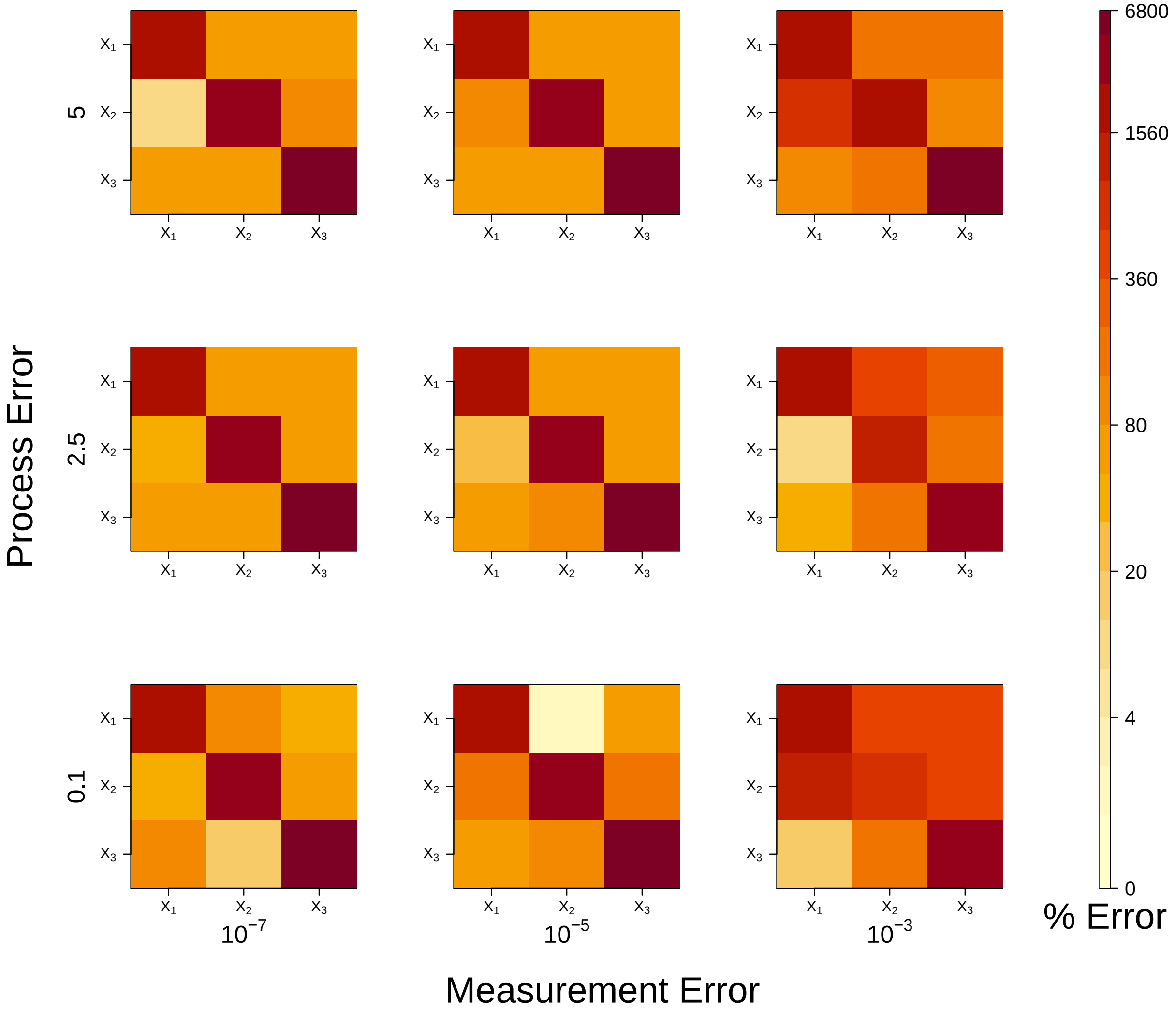

Figure 3.10: **Median error in estimated $b_{i,j}$ for systems with varying levels of noise.** The figure shows the percent error between values of $b_{i,j}$ used to simulate time-series abundance data and estimations of each $b_{i,j}$ found using the MAR model. The simulation was run at each level of process and measurement error 10 times.

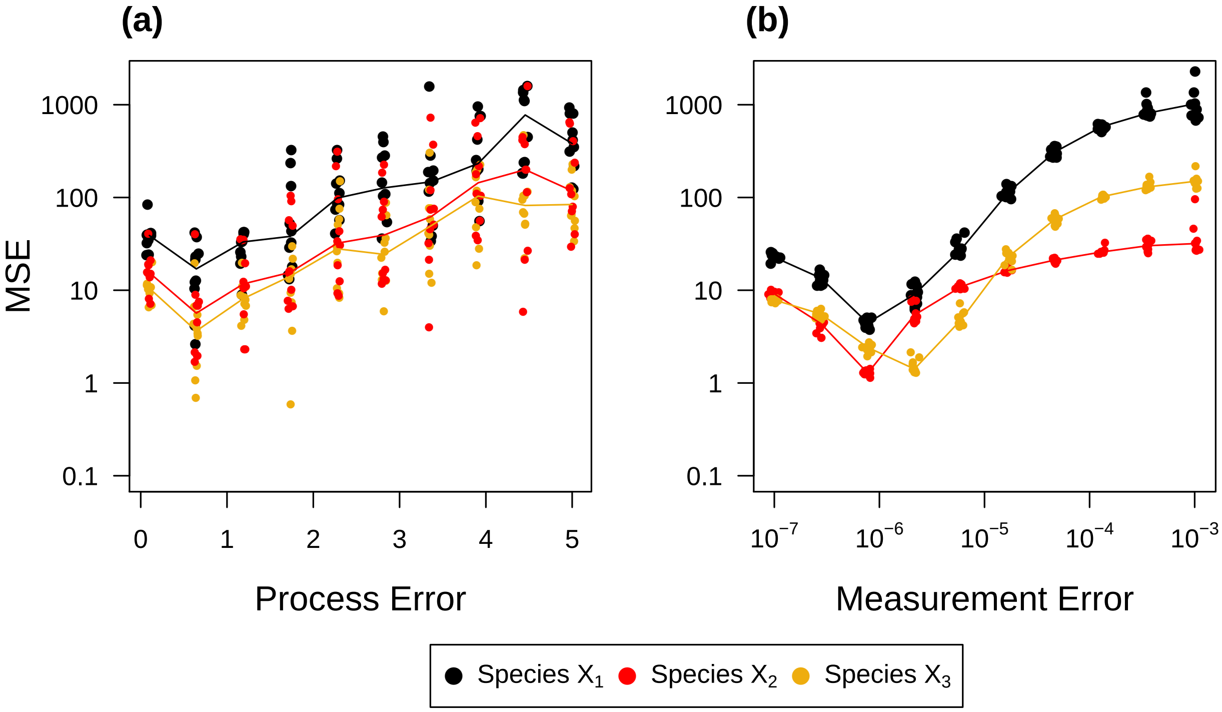

Figure 3.11: **Mean-square error in estimated time-series derived from data with only one type of error.** Parameters were estimated by using the MAR method on simulated time-series data. For this particular test, data were simulated with only process (a) or measurement (b) error.

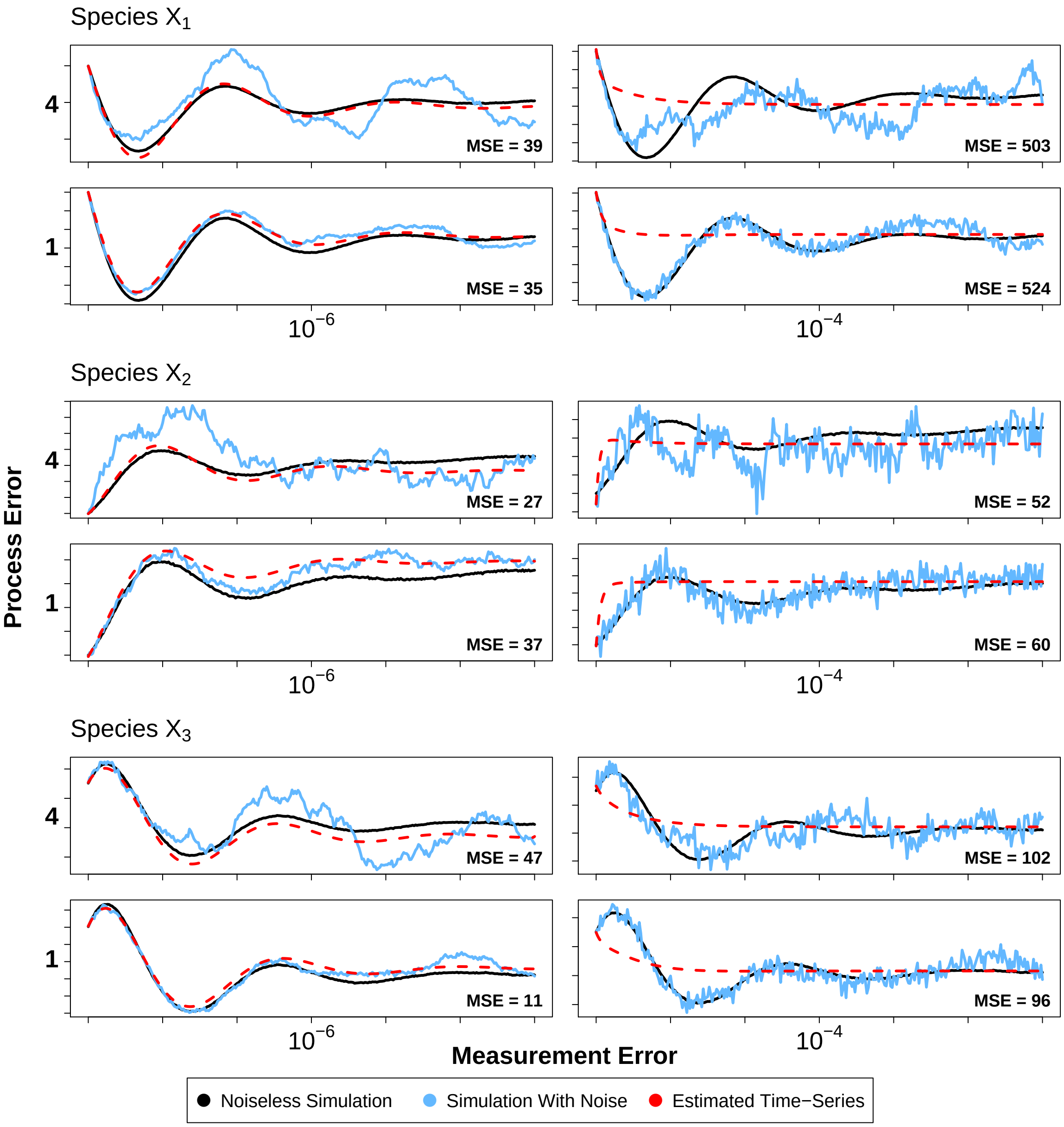

Figure 3.12: **Comparison of time-series reconstructed from estimated parameters and simulated time-series with varying levels of error.** For each species, four sample paths are presented with high and low levels of process and measurement error.

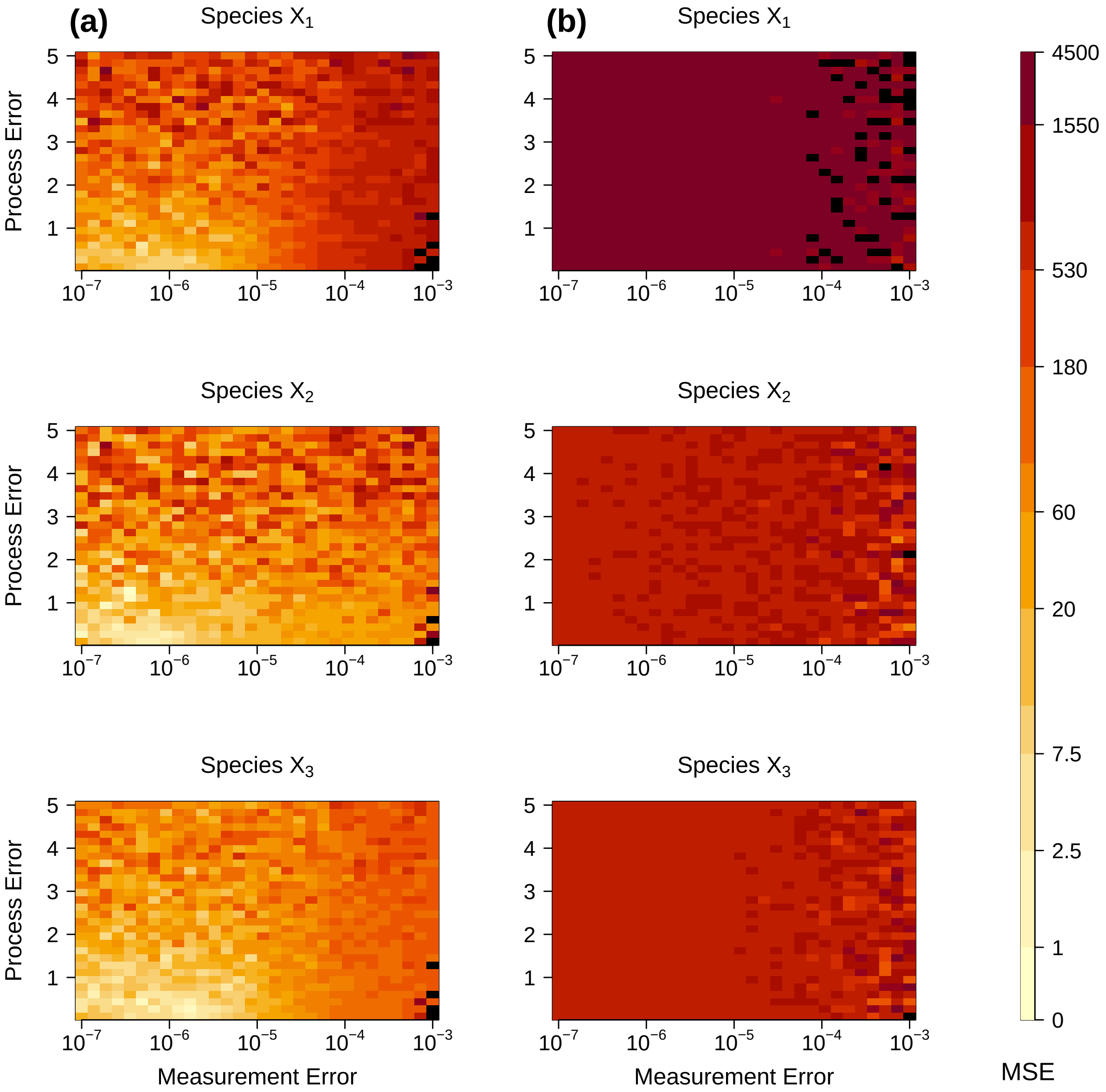

Figure 3.13: **Comparison of reconstructed time-series with differing numbers of time points.** Each heatmap shows the MSE for reconstructions of each species at varying levels of process and measurement error. Column (a) was estimated from a time-series with observations taken every 0.001 steps, while (b) has observations every 0.01 steps.

estimates of absolute abundance data (Fig. 3.11).

Similarly to reconstructions of time-series from relative abundance gLV data, reconstructions from MAR relative abundance data were not resilient to larger values of ξ (Fig. 3.16). When $\xi = 10^{-4}$, reconstructions leveled off at the mean almost immediately. Process noise had a much smaller impact on reconstructions, which closely followed the species' actual paths and consistently leveled off at the mean. When the number of time points was reduced, MSE in time-series reconstructions increased significantly.

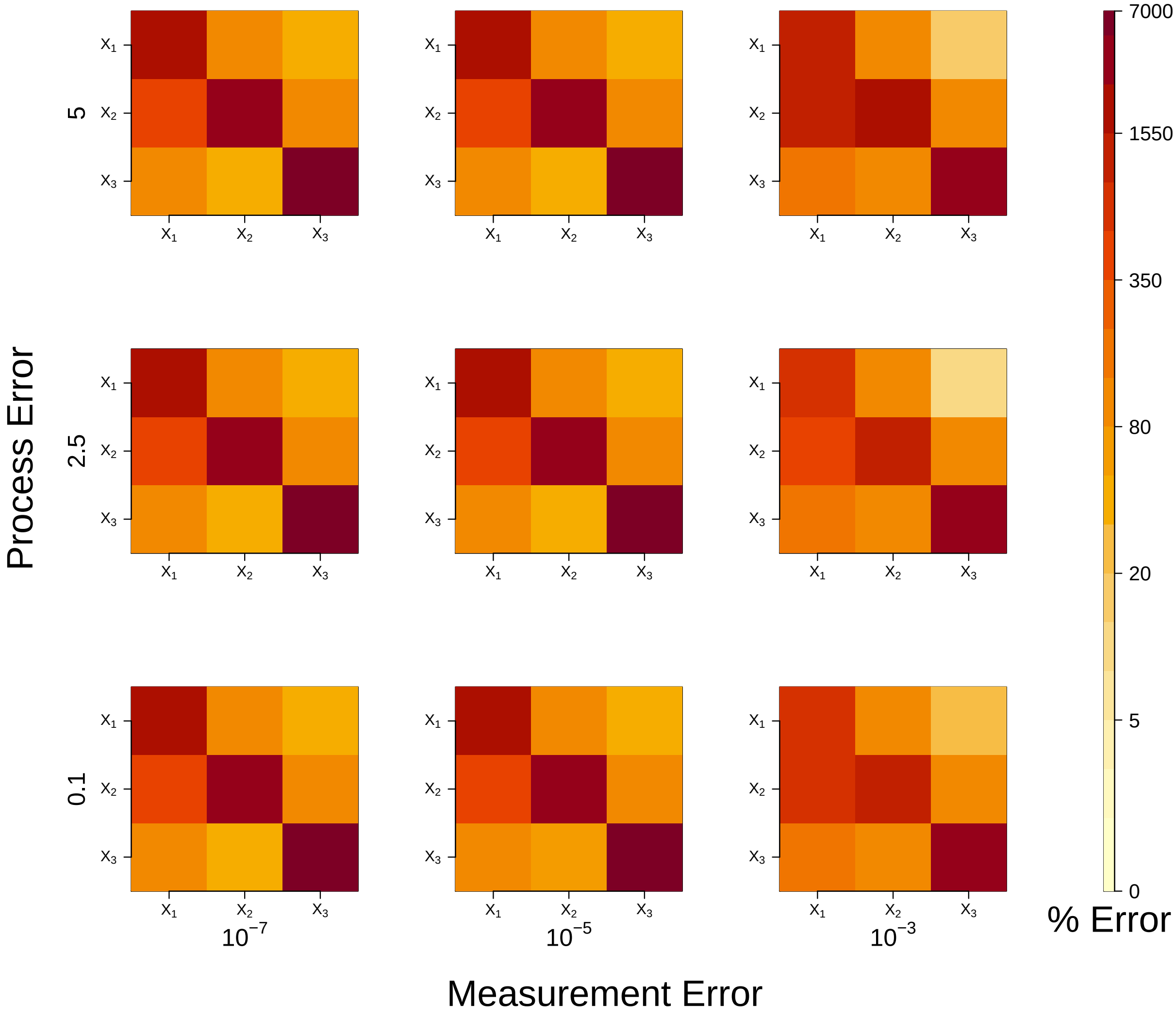

Figure 3.14: **Median error in estimated $b_{i,j}$ for systems with varying levels of noise; relative abundance case.** The figure shows the percent error between values of $b_{i,j}$ used to simulate time-series relative abundance data and estimations of each $b_{i,j}$ found using the MAR model. The simulation was run at each level of process and measurement error 10 times.

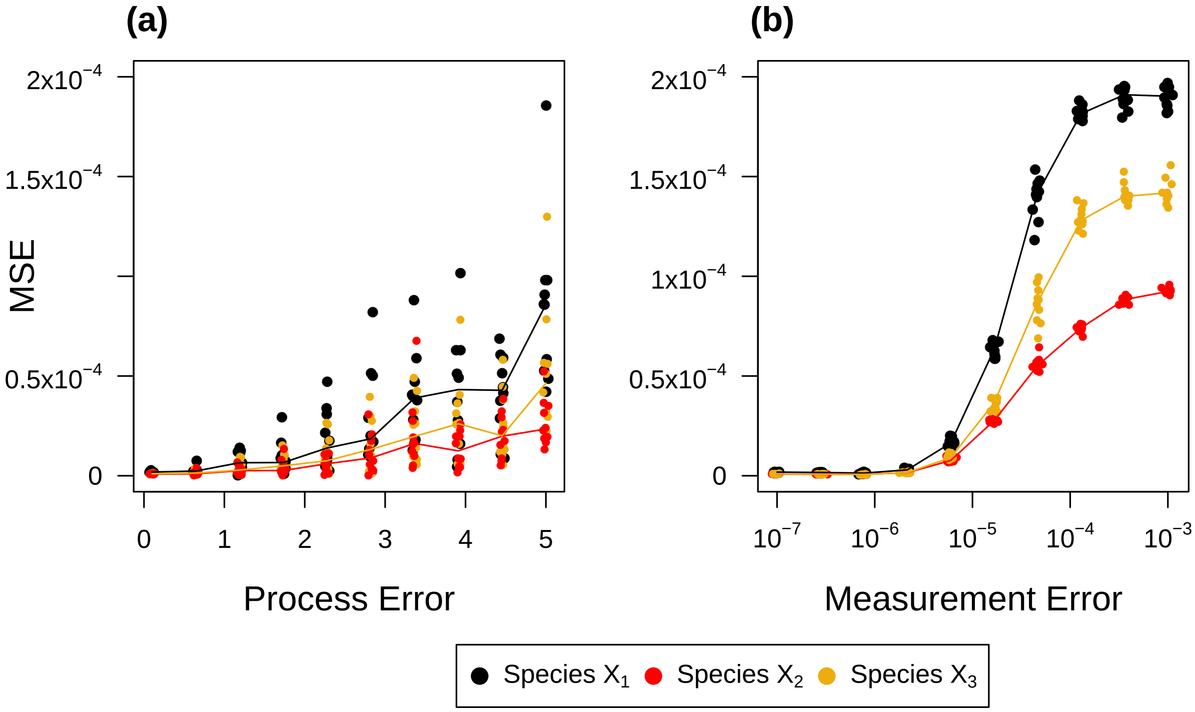

Figure 3.15: **Mean-square error in estimated time-series derived from data with only one type of error; relative abundance case.** Parameters were estimated by using the MAR method on simulated time series data. For this particular test, data were simulated with only process (a) or measurement (b) error.

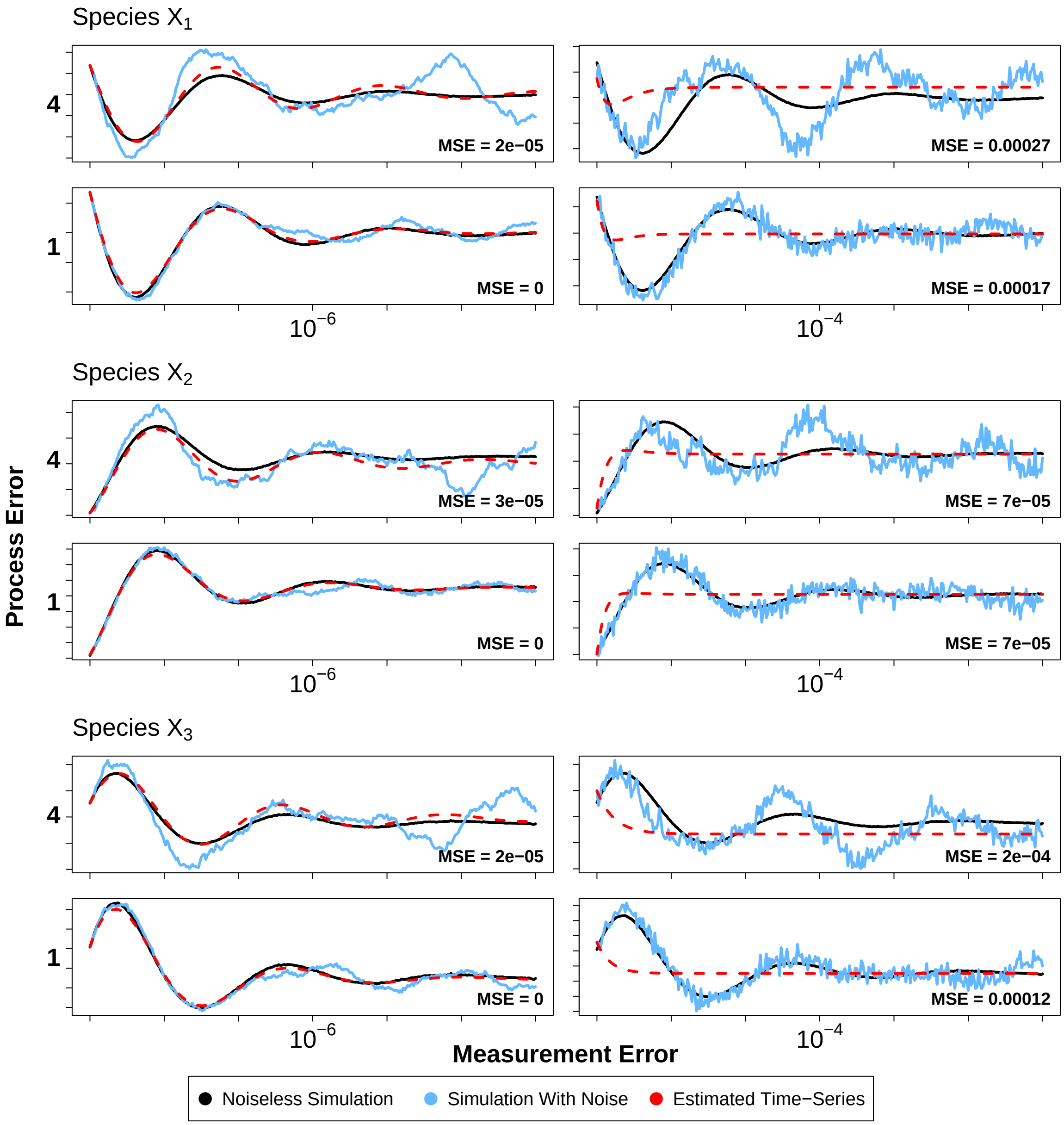

Figure 3.16: **Comparison of time-series reconstructed from MAR-estimated parameters and simulated time-series with varying levels of error, relative abundance case.** For each species, four sample paths are presented with high and low levels of process and measurement error.

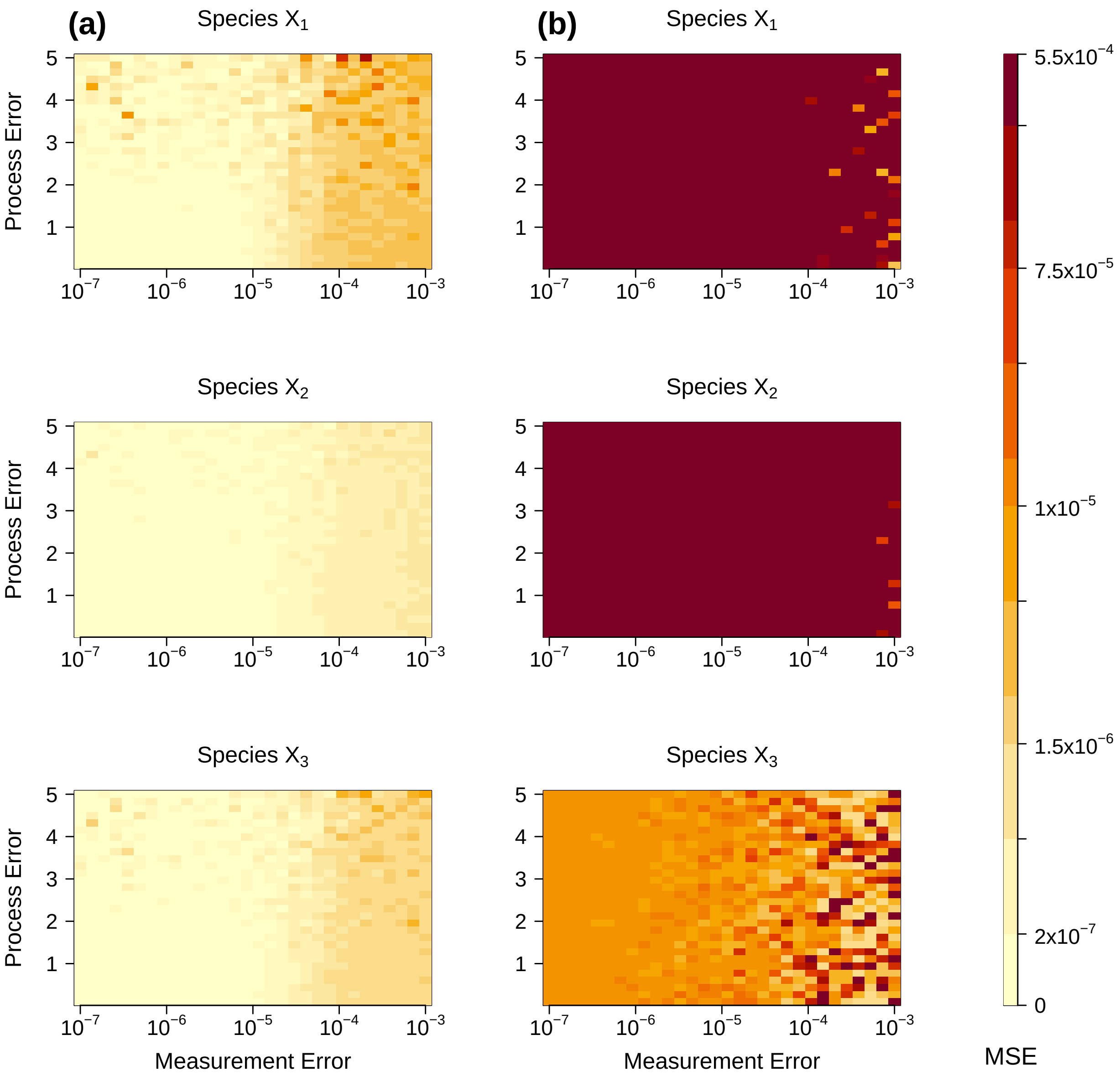

Figure 3.17: **Comparison of reconstructed time-series with differing numbers of time points, using MAR-estimated parameters and relative abundance data.** Each heatmap shows the MSE for reconstructions of each species at varying levels of process and measurement error. Column (a) was estimated from a time-series with observations taken every 0.001 steps, while (b) has observations every 0.01 steps.

CHAPTER 4

Discussion

Our synthetic time-series read count data were developed by assuming underlying interactions meant to mimic the community and error structure of real microbiomes. The gLV model was chosen because it describes direct interactions between species and has been a commonly used model to describe microbial communities [4, 5, 40]. If species i negatively affects the growth of species j, it is because species i has some direct negative impact on species j, for example as a predator-prey relation. However, it is not always the case that all microbial interactions are direct in nature. Some microbial species may interact through competition for a resource, or by consuming substances excreted by other species in the system (e.g., cross-feeding) [41]. The gLV model may not be capable of emulating some types of interactions within microbial communities [42]. The strengths of the gLV framework in modeling interspecies interactions are the simplicity of the interaction terms and linearity with respect to a focal species. These properties combined make the gLV model a powerful tool for efficiently fitting large systems and facilitating easy interpretation of results.

We found that parameters estimated using the gLV model were consistently more accurate than those estimated with the MAR model. Because the underlying synthetic data was generated using the gLV model, it is expected that gLV-estimated parameters would be more accurate. The accuracy of reconstructed time-series from estimated interaction parameters was most affected by the level of measurement error in the data. This discrepancy in accuracy of estimates brings up an interesting point about the purpose of estimating a community: Is it more important to know the "true" underlying parameters, or is it enough to find any set of parameters which adequately describes the system? If the goal is the accurate estimation of the nature of interactions, then the MAR method of estimation appears insufficient. In cases where the interest lies in examining the trajectories of species abundance, an inaccurate set of interaction parameters which describes a similar system may be enough, particularly given the opportunities for analysis afforded by MAR modeling. Of particular note, Ives et al. [31] showed that MAR parameters describing species interactions for population abundances taken on a log scale can be used to assess an ecological community's resilience to perturbations, and its ability to return to stationarity after such a perturbation.

It is not unexpected that the parameter estimation methods presented here struggle with large amounts of measurement error, as one-step ahead fitting with linear regression assumes that the noise is purely from process error. Models exist to account for both process and measurement error [43]. If data quality and quantity are sufficient, these so-called state-space models can be fit using the Kalman filter; however, fitting such models can be computationally expensive for large systems.

Another limitation which has become apparent when estimating parameters is the need for the simulated data to fall in a specific numeric range relative to the equilibrium. When the initial conditions were too close to the steady state, there wasn't enough movement in species' abundance to accurately estimate parameters, and estimated trajectories were flat. In contrast, if the initial conditions were too far from the steady state, the elastic net would overfit the data. Relative stationarity is already a requirement for comparing gLV- and MAR-estimated parameters for population dynamics, but this reinforces the importance of data being near an equilibrium.

In our work, we calculated mean squared error by comparing the entire reconstructed time series to the synthetic data. Alternatively, we could have calculated the mean squared error by comparing an estimate for the data at time $t + 1$ given synthetic data at time t to the synthetic data at time $t + 1$ for the entire time series. These two methods are different because in the latter case, process error is not allowed to propagate over time. If interest lies in the immediate future (e.g., predicting one time point ahead), then short-term prediction may provide a high level of accuracy. However, short-term prediction will retain error from the original dataset, which will lead to greater inaccuracies the farther out the predictions go. Long-term prediction may be less accurate on a point-by-point basis, but may be more successful at predicting trends in the behavior of species trajectories.

While our examination of only one small synthetic system is in some ways limiting, it allowed us to more thoroughly understand how different discrepancies in data affect results. For example, species X_1 covers a greater range of abundances than species' X_2 or X_3, and in the noiseless case hits the highest absolute abundance of the three species. Species X_1 also had higher levels of mean squared error. We may expect a similar trend in larger systems.

As the body of research surrounding microbiomes continues to grow, it becomes more vital to understand the limitations of the methods used to quantify them. Here, we see that models commonly used to assess microbial interactions can describe systems with a high degree of accuracy, but struggle under certain conditions. Specifically, we see that both the gLV and MAR fitting routines become less accurate in cases with even moderate levels of measurement noise, or when time points become sparse. These are both shortcomings that can be addressed at the data collection stage with more frequent sampling, or in the analysis stage by choosing models which are better suited to handling measurement error. As long as researchers recognize how even moderate decreases in data resolution can impact the accuracy of their results, they can take steps to account for these shortcomings, or be mindful of the potential error in their results.

The purpose of this study was to shed light on the accuracy of one-step-ahead fitting for estimating parameters of gLV-based microbiome data. A valuable next step will be examination of a wider range of simulated datasets with more variation in community size, length and number of time steps, and so on. With a more thorough understanding of when and how parameter estimation for microbiomes succeeds and fails, previous microbiome studies can be reexamined with a more precise understanding of accuracy. The structure of future microbiome studies can also be informed by a knowledge of the circumstances under which accurate results can be obtained.

Bibliography

[1] Susannah G Tringe and Philip Hugenholtz. A renaissance for the pioneering 16s rrna gene. *Current Opinion in Microbiology*, 11(5):442 – 446, 2008. ISSN 1369-5274. doi: https://doi.org/10.1016/j.mib.2008.09.011. URL `http://www.sciencedirect.com/science/article/pii/S1369527408001264`. Antimicrobials/Genomics.

[2] James J. Kozich, Sarah L. Westcott, Nielson T. Baxter, Sarah K. Highlander, and Patrick D. Schloss. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Applied and Environmental Microbiology*, 2013. ISSN 0099-2240. doi: 10.1128/AEM.01043-13. URL `https://aem.asm.org/content/early/2013/06/17/AEM.01043-13`.

[3] Gregory B. Gloor, Ruben Hummelen, Jean M. Macklaim, Russell J. Dickson, Andrew D. Fernandes, Roderick MacPhee, and Gregor Reid. Microbiome profiling by illumina sequencing of combinatorial sequence-tagged pcr products. *PLOS ONE*, 5(10):1–15, 10 2010. doi: 10.1371/journal.pone.0015406. URL `https://doi.org/10.1371/journal.pone.0015406`.

[4] Jeremy E. Koenig, Aymé Spor, Nicholas Scalfone, Ashwana D. Fricker, Jesse Stombaugh, Rob Knight, Largus T. Angenent, and Ruth E. Ley. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1000081107. URL `https://www.pnas.org/content/108/Supplement_1/4578`.

[5] Lawrence A. David, Arne C. Materna, Jonathan Friedman, Maria I. Campos-Baptista, Matthew C. Blackburn, Allison Perrotta, Susan E. Erdman, and Eric J. Alm. Host lifestyle affects human microbiota on daily timescales. *Genome Biology*, 15(7):R89, Jul 2014. ISSN 1474-760X. doi: 10.1186/gb-2014-15-7-r89. URL `https://doi.org/10.1186/gb-2014-15-7-r89`.

[6] J. Gregory Caporaso, Christian L. Lauber, William A. Walters, Donna Berg-Lyons, Catherine A. Lozupone, Peter J. Turnbaugh, Noah Fierer, and Rob Knight. Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4516–4522, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1000080107. URL `https://www.pnas.org/content/108/Supplement_1/4516`.

[7] Hong-Wei Zhou, Dong-Fang Li, Nora Fung-Yee Tam, Xiao-Tao Jiang, Hai Zhang, Hua-Fang Sheng, Jin Qin, Xiao Liu, and Fei Zou. Bipes, a cost-effective high-throughput method for assessing microbial diversity. *The ISME Journal*, 5(4):741–749, Apr 2011. ISSN 1751-7370. doi: 10.1038/ismej.2010.160. URL `https://doi.org/10.1038/ismej.2010.160`.

[8] Les Dethlefsen, Sue Huse, Mitchell L Sogin, and David A Relman. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rrna sequencing. *PLOS Biology*, 6(11):1–18, 11 2008. doi: 10.1371/journal.pbio.0060280. URL `https://doi.org/10.1371/journal.pbio.0060280`.

[9] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R. Mende, Adriana Alberti, Francisco M. Cornejo-Castillo, Paul I. Costea, Corinne Cruaud, Francesco d'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M. Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T. Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, , Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B. Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G. Acinas, and Peer Bork. Structure and function of the global ocean micro-

biome. *Science*, 348(6237), 2015. ISSN 0036-8075. doi: 10.1126/science.1261359. URL http://science.sciencemag.org/content/348/6237/1261359.

[10] Jacqueline M. Chaparro, Amy M. Sheflin, Daniel K. Manter, and Jorge M. Vivanco. Manipulating the soil microbiome to increase soil health and plant fertility. *Biology and Fertility of Soils*, 48(5):489–499, Jul 2012. ISSN 1432-0789. doi: 10.1007/s00374-012-0691-4. URL https://doi.org/10.1007/s00374-012-0691-4.

[11] J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, Jeffrey I Gordon, and Rob Knight. Moving pictures of the human microbiome. *Genome Biology*, 12(5):R50–R50, 2011. doi: 10.1186/gb-2011-12-5-r50. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3271711/.

[12] Vanni Bucci, Belinda Tzen, Ning Li, Matt Simmons, Takeshi Tanoue, Elijah Bogart, Luxue Deng, Vladimir Yeliseyev, Mary L. Delaney, Qing Liu, Bernat Olle, Richard R. Stein, Kenya Honda, Lynn Bry, and Georg K. Gerber. Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses. *Genome Biology*, 17(1): 121, Jun 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0980-6. URL https://doi.org/10.1186/s13059-016-0980-6.

[13] Jonathan Friedman, Logan M. Higgins, and Jeff Gore. Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology & Evolution*, 1(5):0109, Mar 2017. ISSN 2397-334X. doi: 10.1038/s41559-017-0109. URL https://doi.org/10.1038/s41559-017-0109.

[14] Ophelia S Venturelli, Alex V Carr, Garth Fisher, Ryan H Hsu, Rebecca Lau, Benjamin P Bowen, Susan Hromada, Trent Northen, and Adam P Arkin. Deciphering microbial interactions in synthetic human gut microbiome communities. *Molec-*

ular Systems Biology, 14(6):e8157, 2018. doi: 10.15252/msb.20178157. URL https://www.embopress.org/doi/abs/10.15252/msb.20178157.

[15] Ainslie E.F. Little, Courtney J. Robinson, S. Brook Peterson, Kenneth F. Raffa, and Jo Handelsman. Rules of engagement: Interspecies interactions that regulate microbial communities. *Annual Review of Microbiology*, 62(1):375–401, 2008. doi: 10.1146/annurev.micro.030608.101423. URL https://doi.org/10.1146/annurev.micro.030608.101423. PMID: 18544040.

[16] M. Claire Horner-Devine, Jessica M. Silver, Mathew A. Leibold, Brendan J. M. Bohannan, Robert K. Colwell, Jed A. Fuhrman, Jessica L. Green, Cheryl R. Kuske, Jennifer B. H. Martiny, Gerard Muyzer, Lise Øvreås, Anna-Louise Reysenbach, and Val H. Smith. A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology*, 88(6):1345–1353, 2007. doi: 10.1890/06-0286. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/06-0286.

[17] Albert Barberán, Scott T Bates, Emilio O Casamayor, and Noah Fierer. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal*, 6(2):343–351, 2012. doi: 10.1038/ismej.2011.119. URL https://doi.org/10.1038/ismej.2011.119.

[18] Richard R. Stein, Vanni Bucci, Nora C. Toussaint, Charlie G. Buffie, Gunnar RÃtsch, Eric G. Pamer, Chris Sander, and JoÃ£o B. Xavier. Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLOS Computational Biology*, 9(12):1–11, 12 2013. doi: 10.1371/journal.pcbi.1003388. URL https://doi.org/10.1371/journal.pcbi.1003388.

[19] Charles K. Fisher and Pankaj Mehta. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLOS ONE*,

9(7):1–10, 07 2014. doi: 10.1371/journal.pone.0102451. URL https://doi.org/10.1371/journal.pone.0102451.

[20] Grace Tzun-Wen Shaw, Yueh-Yang Pao, and Daryi Wang. Metamis: a metagenomic microbial interaction simulator based on microbial community profiles. *BMC Bioinformatics*, 17(1):488, Nov 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1359-0. URL https://doi.org/10.1186/s12859-016-1359-0.

[21] Mustafa Alshawaqfeh, Erchin Serpedin, and Ahmad Bani Younes. Inferring microbial interaction networks from metagenomic data using sglv-ekf algorithm. *BMC Genomics*, 18(3):228, Mar 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-3605-x. URL https://doi.org/10.1186/s12864-017-3605-x.

[22] Xuefeng Gao, Bich-Tram Huynh, Didier Guillemot, Philippe Glaser, and Lulla Opatowski. Inference of significant microbial interactions from longitudinal metagenomics data. *Frontiers in Microbiology*, 9:2319, 2018. ISSN 1664-302X. doi: 10.3389/fmicb.2018.02319. URL https://www.frontiersin.org/article/10.3389/fmicb.2018.02319.

[23] Rajith Vidanaarachchi, Marnie Shaw, Sen-Lin Tang, and Saman Halgamuge. Imparo: inferring microbial interactions through parameter optimisation. *BMC Molecular and Cell Biology*, 21(1):34, Aug 2020. ISSN 2661-8850. doi: 10.1186/s12860-020-00269-y. URL https://doi.org/10.1186/s12860-020-00269-y.

[24] Jonathan Friedman and Eric J. Alm. Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, 8(9):1–11, 09 2012. doi: 10.1371/journal.pcbi.1002687. URL https://doi.org/10.1371/journal.pcbi.1002687.

[25] Kun-Nan Tsai, Shu-Hsi Lin, Wei-Chung Liu, and Daryi Wang. Inferring microbial interaction network from microbiome data using rmn algorithm. *BMC Systems Biology*,

9(1):54, Sep 2015. ISSN 1752-0509. doi: 10.1186/s12918-015-0199-2. URL https://doi.org/10.1186/s12918-015-0199-2.

[26] Jens Christian Claussen, Jurgita Skiecevičienė, Jun Wang, Philipp Rausch, Tom H. Karlsen, Wolfgang Lieb, John F. Baines, Andre Franke, and Marc-Thorsten Hütt. Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLOS Computational Biology*, 13(6):1–21, 06 2017. doi: 10.1371/journal.pcbi.1005361. URL https://doi.org/10.1371/journal.pcbi.1005361.

[27] Ye Wang, Tathagata Bhattacharya, Yuchao Jiang, Xiao Qin, Yue Wang, Yunlong Liu, Andrew J Saykin, and Li Chen. A novel deep learning method for predictive modeling of microbiome data. *Briefings in Bioinformatics*, 05 2020. ISSN 1477-4054. doi: 10.1093/bib/bbaa073. URL https://doi.org/10.1093/bib/bbaa073. bbaa073.

[28] Holmes Elizabethe., Ward Ericj., and Kellie Wills. Marss: Multivariate autoregressive state-space models for analyzing time-series data. *The R Journal*, 4(1):11, 2012. doi: 10.32614/rj-2012-002.

[29] Ralph Mac Nally, James R. Thomson, Wim J. Kimmerer, Frederick Feyrer, Ken B. Newman, Andy Sih, William A. Bennett, Larry Brown, Erica Fleishman, Steven D. Culberson, and Gonzalo Castillo. Analysis of pelagic species decline in the upper san francisco estuary using multivariate autoregressive modeling (mar). *Ecological Applications*, 20(5):1417–1430, 2010. doi: 10.1890/09-1724.1. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/09-1724.1.

[30] Colette Mair, Sema Nickbakhsh, Richard Reeve, Jim McMenamin, Arlene Reynolds, Rory N. Gunson, Pablo R. Murcia, and Louise Matthews. Estimation of temporal covariances in pathogen dynamics using bayesian multivariate autoregressive models, 2019. URL https://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1007492.

[31] A. R. Ives, B. Dennis, K. L. Cottingham, and S. R. Carpenter. Estimating community stability and ecological interactions from time-series data. *Ecological Monographs*, 73 (2):301–330, 2003. ISSN 1557-7015. doi: 10.1890/0012-9615(2003)073[0301:ECSAEI] 2.0.CO;2. URL `http://dx.doi.org/10.1890/0012-9615(2003)073[0301:ECSAEI]2.0.CO;2`.

[32] Linda J S Allen. *An introduction to stochastic processes with applications to biology*. Upper Saddle River, N.J. : Pearson/Prentice Hall, 2003.

[33] Aaron King, Dao Nguyen, and Edward Ionides. Statistical inference for partially observed markov processes via the r package pomp. *Journal of Statistical Software, Articles*, 69(12):1–43, 2016. ISSN 1548-7660. doi: 10.18637/jss.v069.i12. URL `https://www.jstatsoft.org/v069/i12`.

[34] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010. doi: 10.1111/j.1467-9868.2009.00736.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2009.00736.x`.

[35] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL `https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634`.

[36] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x`.

[37] H Zou and T Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[38] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

[39] Christopher H Remien, Mariah J Eckwright, and Benjamin J Ridenhour. Parameter identifiability of the generalized lotka-volterra model for microbiome studies. *bioRxiv*, 2018. doi: 10.1101/463372. URL `https://www.biorxiv.org/content/early/2018/11/06/463372`.

[40] Emily L. Kara, Paul C. Hanson, Yu Hen Hu, Luke Winslow, and Katherine D. McMahon. A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic lake mendota, wi, usa. *ISME J*, 7(3):680–684, Mar 2013. ISSN 1751-7362. doi: 10.1038/ismej.2012.118. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3578560/`. 23051691[pmid].

[41] Michael A. Henson and Poonam Phalak. Byproduct cross feeding and community stability in an in silico biofilm model of the gut microbiome. *Processes*, 5(1), 2017. ISSN 2227-9717. doi: 10.3390/pr5010013. URL `https://www.mdpi.com/2227-9717/5/1/13`.

[42] Babak Momeni, Li Xie, and Wenying Shou. Lotka-volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *eLIFE*, pages 1–34, 2017. doi: 10.7554/eLife.25051. URL `https://elifesciences.org/articles/25051`.

[43] Justin D. Silverman, Heather K. Durand, Rachael J. Bloom, Sayan Mukherjee, and Lawrence A. David. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome*, 6(1):202, 2018. doi: 10.1186/s40168-018-0584-3. URL `https://doi.org/10.1186/s40168-018-0584-3`.