mbtransfer: Microbiome Intervention Analysis using Transfer Functions and Mirror Statistics

Kris Sankaran[1,2] and Pratheepa Jeganathan[3]

[1]Department of Statistics, University of Wisconsin - Madison
[2]Wisconsin Institute for Discovery
[3]Department of Mathematics & Statistics, McMaster University

June 12, 2023

Abstract

Microbiome interventions provide valuable data about microbial ecosystem structure and dynamics. Despite their ubiquity in microbiome research, few rigorous data analysis approaches are available. In this study, we extend transfer function-based intervention analysis to the microbiome setting, drawing from advances in statistical learning and selective inference. Our proposal supports the simulation of hypothetical intervention trajectories and False Discovery Rate-guaranteed selection of significantly perturbed taxa. We explore the properties of our approach through simulation and re-analyze three contrasting microbiome studies. An R package, mbtransfer, is available at https://go.wisc.edu/crj6k6/. Notebooks to reproduce the simulation and case studies can be found at https://go.wisc.edu/dxuibh and https://go.wisc.edu/emxv33.

1 Introduction

Figure 1 gives three examples of microbial community dynamics under environmental perturbations. Part (a) shows the shift in the gut microbiome of a subject from David *and others* (2013) during a five-day shift to an animal-based diet. Part (b) shows the postpartum change in the vaginal microbiome from one participant tracked by Costello *and others* (2022). Part (c) gives the dynamics of an aquaculture microbiome in a tank following shifts in environmental pH, as examined by Yajima *and others* (2022). The shifts in these few cases represent general phenomena – the interventions they describe have reproducible effects on the microbiome, consistently altering the abundance of specific taxa on a predictable time scale. Similar microbial community studies are widespread in microbiome research efforts, especially those with the long-term goal of engineering microbial systems that promote health in a dynamic environment.

Statistical inference of intervention effects on the microbiome must account for temporal dependence – otherwise, there is a risk of overinflating the effective sample size. We will see in Section 3 that, though the practice of two-sample testing of intervention effects is common, it leads to inflated false discovery rates. Several microbial community dynamics models have been proposed in response to this issue. Among the most widely used is the generalized Lotka-Volterra (gLV) model, which discretizes an ordinary differential equation model for competitive predator-prey dynamics, optionally including covariates to model environmental influences (Gerber, 2014; Gibbons *and others*, 2017). Specifically, let $\mathbf{y}(t)$ be the microbial community profile at time t, and let $\mathbf{w}(t)$ be the state of the associated intervention. Then, the gLV supposes $\frac{\partial \mathbf{y}(t)}{\partial t} = A\mathbf{y}(t) + D\mathbf{w}(t) + \epsilon(t)$, and it is typically estimated by first log-transforming the observed taxonomic abundances $\log(1 + \mathbf{y}_t)$ and fitting an elastic net regression of $\log(1 + \mathbf{y}_{t+1}) - \log(1 + \mathbf{y}_t)$ onto $\mathbf{w}_t$. The main limitations of this model are (1) that it assumes linearity in the relationship of $\log(1 + \mathbf{y}_{t+1}) - \log(1 + \mathbf{y}_t)$ onto

$\mathbf{w}_t$ and (2) it can only refer to the immediate past of and $\mathbf{w}_t$. Moreover, it does not quantify the uncertainty or stability of any estimated effects.

To address this, Bucci *and others* (2016); Gibson *and others* (2021) and Silverman *and others* (2018, 2022) developed explicit probabilistic models of community dynamics. Silverman *and others* (2022)'s models, MALLARD and fido, are based on a multinomial logistic-normal autoregressive and multinomial logistic-normal Gaussian Process model, respectively. The environmental shifts can be included as covariates to support intervention analysis. MDSINE and MDSINE2 (Bucci *and others*, 2016; Gibson *and others*, 2021) are negative binomial dynamical systems models that extend the gLV and focus on the discovery of interspecies interactions and perturbation effects. Autoregressive dynamics are clustered using a Dirichlet Process, and a Gaussian Process prior is used to regularize species abundance trajectories. These models are closely related to our work in their application of a dynamical systems model and inference of environmental intervention effects. We provide a quantitative comparison in Section 3, and Supplementary Section 7.2 summarizes existing methods.

We make two contributions to the tapestry of currently available models. First, we demonstrate that nonparametric transfer function models lead to more accurate forecasts than current models, especially when environmental shifts are large. We leverage an existing gradient boosting package (Chen *and others*, 2018) and achieve competitive performance, most likely due to their relatively weak modeling assumptions and our data-rich setting. Second, we provide an algorithm to support the precise inference of intervention effects on individual taxa at specific temporal lags. Decoupling community dynamical modeling from inference makes our interpretations of environmental effects more robust to model misspecification. Section 2 explains how to guarantee False Discovery Rate (FDR) control of the selected taxa using only a symmetry assumption.

The key ingredients of our approach are transfer function models (Box and Tiao, 1975), which summarize community dynamics, and mirror statistics (Dai *and others*, 2020), which enable precise inference. Transfer functions relate an "input" series to an "output" one. These models were originally developed to support intervention analysis in time series data, for example, the influence of a new automobile emissions regulation on local ozone levels. Section 2 adapts this framework to the high-dimensional microbiome setting. Mirror statistics are an approach to selective inference that leverages data splitting to rank differential effects while controlling the FDR, and we develop an instance of this algorithm using partial dependence profiles of the fitted boosting models. This approach is analogous to recent microbiome approaches based on knockoffs (Xie and Lederer, 2021; Zhu *and others*, 2021), but it does not depend on the simulation of appropriate knockoff features.

Taken together, transfer function models and mirror-based inference provide answers to the following central questions in microbiome intervention analysis:

1. Which taxa are the most affected by the intervention? Our mirror statistics identify taxa with differential trajectories across counterfactual environmental conditions.

2. When are these taxa affected? We can distinguish between transient and sustained shifts in the community by simulating alternative counterfactuals from our fitted transfer function models.

3. Which factors mediate the shift? Flexible transfer function models can detect interactions between interventions and environmental features.

The mbtransfer R package computes artifacts directly related to these questions. Specifically, it supports training transfer function models, testing for significant taxon-level effects, and simulation under counterfactual alternatives. The package's implementation and documentation, including the code to reproduce the data analysis of Section 4, can be found at https://go.wisc.edu/crj6k6.

2 Method

Section 2.1 discusses a flexible generalization of transfer function models (Box and Tiao, 1975). Here, the input series is a measure of the intervention strength. The resulting model can summarize and simulate

intervention effects on microbiome communities, accounting for baseline composition and mediating host features. Section 2.2 develops mirror statistics (Dai *and others*, 2020) to formally infer which taxa are the most strongly influenced by the interventions and whether the effects are immediate or delayed. The overall workflow supports statistically-guaranteed discovery of intervention effects in microbial time series.

2.1 Transfer function models

Transfer function models were introduced as a linear autoregressive model applied to two concurrent time series, a series $w_t \in \mathbb{R}$ that serves as the intervention and a series $y_t \in \mathbb{R}$ that changes in response. We consider the generalization,

$$y_{tj}^{(i)} = f_j\left(\mathbf{y}_{(t-P-1):(t-1)}^{(i)}, \mathbf{w}_{(t-Q+1):t}^{(i)}, \mathbf{z}^{(i)}\right) + \epsilon_{jt}^{(i)}, \tag{1}$$

where we have used the following notation:

- $\mathbf{y}_t^{(i)} \in \mathbb{R}^J$: The (potentially transformed) abundances of all taxa $j \in \{1, \ldots, J\}$ at time t in subject i. This vector has jth coordinate $y_{tj}^{(i)}$.
- $\mathbf{w}_t^{(i)} \in \mathbb{R}^D$: The strength of D different interventions at time t in subject i.
- $\mathbf{z}^{(i)} \in \mathbb{R}^S$: The characteristics of subject i that do not vary over time.
- $\epsilon_{jt}^{(i)}$: Random error in the abundance of taxon j for timepoint t in subject i. In Section 2.2, this noise is assumed symmetric.

We learn each f_j separately using gradient boosting models (Friedman, 2001; Chen and Guestrin, 2016). For training, we extract nonoverlapping temporal segments, and the last P lags of $\mathbf{y}_t^{(i)}$ and Q lags of $\mathbf{w}_{t+1}^{(i)}$ are vectorized and combined with $\mathbf{z}^{(i)}$ to form a $\mathbb{R}^{PJ+QD+S}$-dimensional feature vector. Once trained, this model can simulate expected counterfactual trajectories under hypothetical interventions $\tilde{\mathbf{w}}_{(t+1):(t+h)}$ given initial compositions $\mathbf{y}_{(t-P+1):t}$ and subject characteristics $\tilde{\mathbf{z}}$. The one-step forecast is as follows:

$$\hat{\mathbf{f}}\left(\mathbf{y}_{(t-P+1):t}, \tilde{\mathbf{w}}_{(t-Q+2):(t+1)}, \tilde{\mathbf{z}}\right) := \begin{bmatrix} \hat{f}_1\left(\mathbf{y}_{(t-P+1):t}, \tilde{\mathbf{w}}_{(t-Q+2):(t+1)}, \tilde{\mathbf{z}}\right) \\ \vdots \\ \hat{f}_J\left(\mathbf{y}_{(t-P+1):t}, \tilde{\mathbf{w}}_{(t-Q+2):(t+1)}, \tilde{\mathbf{z}}\right) \end{bmatrix}$$

and longer time horizons can be forecast by substituting intermediate predictions:

$$\hat{\mathbf{f}}^{+h}\left(\hat{\mathbf{y}}_{(t-P+h):(t+h-1)}, \tilde{\mathbf{w}}_{(t-Q+h+1):(t+h)}, \tilde{\mathbf{z}}\right) := \begin{bmatrix} \hat{f}_1\left(\hat{\mathbf{y}}_{(t-P+h):(t+h-1)}, \tilde{\mathbf{w}}_{(t-Q+h+1):(t+h)}, \tilde{\mathbf{z}}\right) \\ \vdots \\ \hat{f}_J\left(\hat{\mathbf{y}}_{(t-P+h):(t+h-1)}, \tilde{\mathbf{w}}_{(t-Q+h+1):(t+h)}, \tilde{\mathbf{z}}\right) \end{bmatrix}$$

where we used the convention that $\hat{\mathbf{y}}_{t'} = \mathbf{y}_{t'}$ if $t' \leq t$ is observed and $\hat{\mathbf{y}}_{t'} = \hat{\mathbf{f}}^{+h'}\left(\hat{\mathbf{y}}_{(t-P+h'):(t+h'-1)}, \tilde{\mathbf{w}}_{(t-Q+h'+1):(t+h')}, \tilde{\mathbf{z}}\right)$ for intermediate predictions $t' = t + h'$ with $h' < h$.

The two advantages of this formulation are: (1) it can estimate nonlinear relationships between past microbial community profiles, interventions, and host features with taxon j's current abundance, and (2) it can detect interaction effects between covariates that improve predictive power and which may have valuable scientific interpretations. Note that each taxon j is trained separately. On the one hand, this means that information is not shared between related taxa. On the other, this allows us to use existing, reliable boosting implementations, and if many taxa are of interest, the implementation is easily parallelizable.

2.2 Mirror statistics

The transfer function model in equation (1) summarizes the effects of interventions $\mathbf{w}_t$ on taxonomic abundances $\mathbf{y}_t$. However, this model may not provide statistical guarantees and can lead to ambiguous results. To address this, we propose a mirror statistics implementation. First, we randomly split subjects into subsets, $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$. Then, for each split s, we train models $\hat{\mathbf{f}}^{(s)}$. Next, we estimate the counterfactual difference between interventions $\tilde{\mathbf{w}}_{(t-Q+2):(t+1)} = \mathbf{1}_Q$ and $\tilde{\mathbf{w}}_{(t-Q+2):(t+1)} = \mathbf{0}_Q$ for each taxon j using:

$$\mathrm{PD}_j^{(s)} = \frac{1}{\left|\mathcal{D}^{(s)}\right|} \sum_{\text{segments} \in \mathcal{D}^{(s)}} \left[\hat{f}_j^{(s)} \left(\mathbf{y}_{(t-P+1):t}^{(i)}, \mathbf{1}_Q, \mathbf{z}^{(i)} \right) - \hat{f}_j^{(s)} \left(\mathbf{y}_{(t-P+1):t}^{(i)}, \mathbf{0}_Q, \mathbf{z}^{(i)} \right) \right]. \tag{2}$$

This equation is a partial dependence profile applied to the fitted model of taxon j (Friedman, 2001; Biecek and Burzykowski, 2021). Note that this definition toggles all D interventions. For an isolated intervention, we can use $(0, \ldots, 0, 1, 0, \ldots, 0)$ instead of $\mathbf{1}_Q$ to get the analogous statistic. We also define the corresponding mirror statistic as:

$$M_j = \mathrm{sign}\left(\mathrm{PD}_j^{(1)} \mathrm{PD}_j^{(2)} \right) \left[\left| \mathrm{PD}_j^{(1)} \right| + \left| \mathrm{PD}_j^{(2)} \right| \right], \tag{3}$$

which measures the consistency between estimated effects across separate splits. We assume that for taxon j with no true intervention effects, $\mathrm{PD}_j^{(s)}$ is symmetrically distributed around 0. This assumption is plausible because in the absence of an intervention effect on taxon j, any differences between $\hat{f}_j^{(s)} \left(\mathbf{y}_{(t-P+1):t}^{(i)}, \mathbf{1}_Q, \mathbf{z}^{(i)} \right)$ and $\hat{f}_j^{(s)} \left(\mathbf{y}_{(t-P+1):t}^{(i)}, \mathbf{0}_Q, \mathbf{z}^{(i)} \right)$ are due to noise.

Given mirror statistics M_j, we estimate the false discovery proportion using the same procedure of Dai *and others* (2020), viewing $\mathrm{PD}_j^{(s)}$ as analogous to $\hat{\beta}_j^{(s)}$. Specifically, we compute:

$$\widehat{\mathrm{FDP}}(t) = \frac{\left| \{ j : M_j < -t \} \right|}{\left| \{ j : M_j > t \} \right|}, \tag{4}$$

where the choice of t defines a selection set $\hat{J}_1$ for the current pair of splits. Given FDR control level q, we choose the largest t such that $\widehat{\mathrm{FDP}}(t) \leq q$. We aggregate across multiple splits to improve power, following Dai *and others* (2020)'s Algorithm 2. Our examples always aggregate across 25 splits. For delayed effects, we define analogous $\mathrm{PD}_j^{(s),+h}$ and M_j^{+h} using f_j^{+h} instead of f_j, and the estimate in equation 4 is modified to use mirrors across all lags h.

3 Simulations

We perform simulation experiments to examine the effectiveness of mbtransfer relative to competitive baselines and to identify data characteristics that lead to better or worse performance. We evaluate both forecasting ability and inferential quality.

3.1 Data generating mechanism

We simulate data from a negative binomial vector autoregressive model:

$$\begin{aligned}
\mathbf{y}_t^{(i)} | \theta_t^{(i)}, \varphi_i &\sim \mathrm{NB}\left(\exp\left(\theta_t^{(i)}\right), \varphi_i\right) \\
\theta_t^{(i)} | \epsilon_t^{(i)} &= \sum_{p=1}^{P} A_p \theta_{t-p}^{(i)} + \sum_{q=1}^{Q} \left(B_q + C_q \odot z^i\right) \mathbf{w}_{t-q}^i + \epsilon_t^{(i)} \\
\varphi_{ij} &\sim \Gamma(\alpha, \lambda) \\
\epsilon_t^{(i)} &\sim \mathcal{N}\left(0, \sigma_\epsilon^2 I_J\right)
\end{aligned}$$

Here, $\theta_t^{(i)} \in \mathbb{R}^J$, and NB refers to a negative binomial distribution applied coordinate-wise to each taxon j using a mean-dispersion parameterization. $A_p \in \mathbb{R}^{J \times J}$ parameterizes the lag-p autoregressive dynamics between pairs of taxa and $B_q \in \mathbb{R}^{J \times D}$ parameterizes the lag-q effect of the D interventions. We have chosen a negative binomial generative mechanism because this distribution has previously been found to fit 16S rRNA gene sequencing data well, especially after correcting for library size differences (Calgaro *and others*, 2020). C_q represents an interaction between host characteristics and the intervention, where some taxa may be more strongly affected by an intervention when their host has particular features.

The parameters B_q, C_q, and A_p are simulated as follows. The first J_1 taxa have true intervention effects and the remaining $J_0 = J - J_1$ rows of B_q and C_q are set to 0. Among the nonnull taxa $j \in J_1$, we draw $B_{q,jd} \sim \mathrm{Unif}([-2b, -b] \cup [b, 2b])$ where b encodes the signal strength. Using two intervals ensures that nonnull effects are bounded away from 0. Entries $C_{q,jd}$ are drawn similarly, except entire rows $C_{q,j\cdot}$ are set to 0 with an additional probability p_c. Such rows represent taxa with real intervention effects but no interaction with host characteristics, represented by $z^{(i)} \sim \mathcal{N}(0, \sigma_z^2)$. Finally, we simulate $A \in \mathbb{R}^{J \times J}$ as a sparsified version of a random, low-rank matrix. Specifically, we first set $\tilde{A}^{(0)} \sim QQ^T$ where $Q \in \mathbb{R}^{J \times K}$ has entries drawn independently from $\mathcal{N}(0, \sigma_A^2)$. Entries of $\tilde{A}^{(0)}$ are randomly set to 0 with probability p_A, yielding $\tilde{A}^{(1)}$, and the result is normalized: $A = \frac{\tilde{A}^{(1)}}{\|\tilde{A}^{(1)}\|_2}$.

We simulate random, one-dimensional interventions $w_t^{(i)} \in \{0, 1\}$ by first randomly sampling a starting point $t^{\text{start}} \sim \mathrm{Unif}\left[\frac{T}{3}, \ldots, \frac{2T}{3}\right]$. The intervention length is drawn from $\ell \in \mathrm{Unif}[L, 2L]$. If $t^{\text{start}} + \ell > T$, we truncate the intervention series at T. A visualization of the trajectories for null and nonnull taxa is given in Figure 2. For inference, we compare the mirror algorithm to DESeq2 (Love *and others*, 2014) with the formula $\sim$ `intervention` $+ z^{(i)} + z^{(i)} \times$ `intervention`. DESeq2 is a negative binomial-based generalized linear model originally developed for hypothesis testing in bulk RNA-seq data. Nonetheless, it is often recommended in 16S sequencing analysis and has exhibited strong performance in benchmarks against methods specifically built for 16S gene sequencing data (Callahan *and others*, 2016; Calgaro *and others*, 2020). Supplementary Section 7.1 provides further details for examining and reproducing the simulation setup.

Given this simulation mechanism, we vary the following data parameters:

1. The number of taxa $\in \{100, 200, 400\}$.
2. The fraction of null taxa $\pi_0 \in \{0.1, 0.2, 0.4\}$. Given π_0, we set $J_0 = \lfloor \pi_0 J \rfloor$.
3. The signal strength $b \in \{0.25, 0.5, 1\}$.

Across all runs, we simulate 50 subjects with 30 timepoints each. We fix $p_c = 0.2$ and $p_A = 0.4$. Considering all combinations of these parameters yields 27 simulated datasets. These can be downloaded from https://go.wisc.edu/8ey754, and simulation outputs discussed below are available at[1] https://go.wisc.edu/3gc982.

[1]Reproducibility details are discussed in Supplementary Section 7.1. Results were obtained using Center for High Throughput Computing (2006).

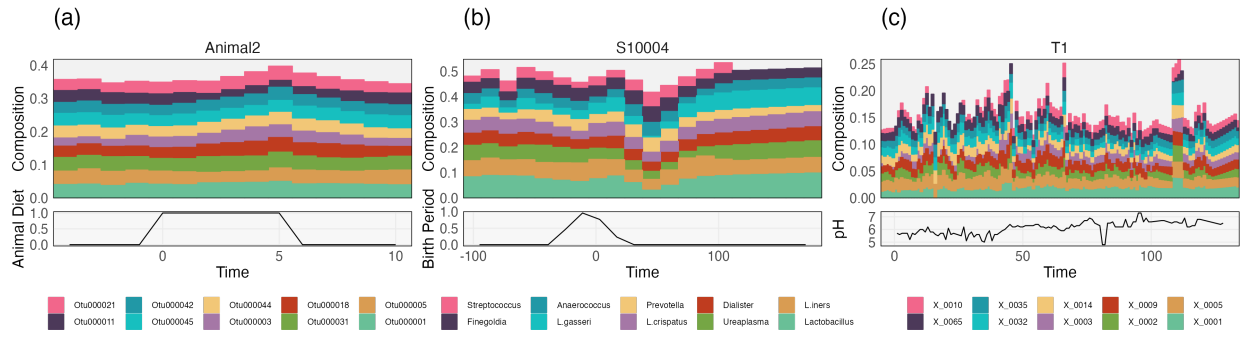

Figure 1: Examples of microbial community shifts in response to environmental change. Part (a) describes the gut microbiome of a subject undergoing a diet intervention (David *and others*, 2013), (b) shows remodeling of a mother's vaginal microbiome following birth (Costello *and others*, 2022), and (c) profiles an aquaculture tank microbiome together with environmental pH (Yajima *and others*, 2022). Section 4 explores data from these studies in depth.

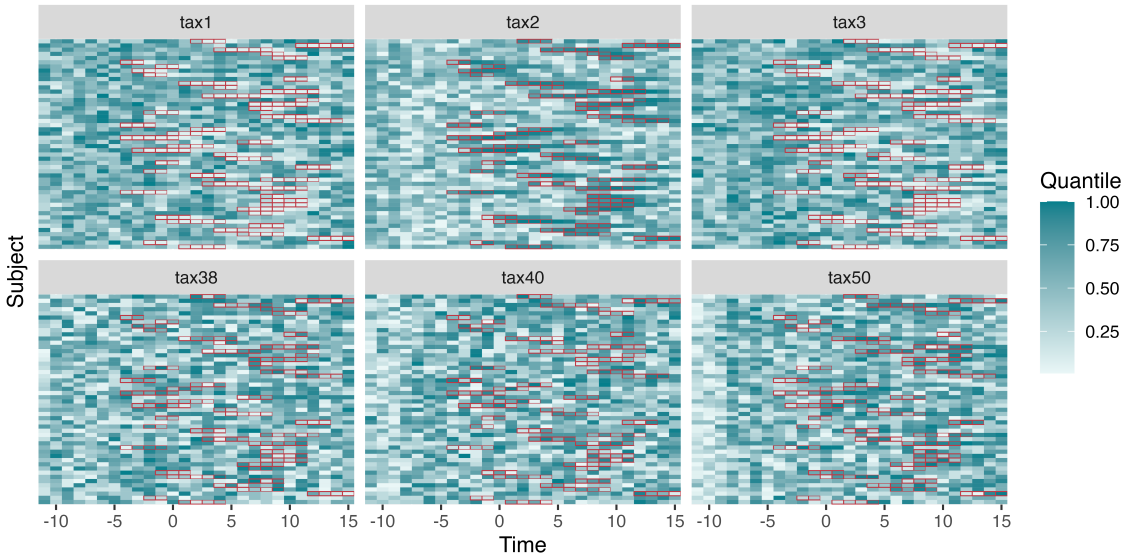

Figure 2: Time series simulated according to the mechanism described in Section 3.1. Each panel shows the trajectories for one taxon, and each row is the time series for one subject. The colors of the tiles encode changes in abundance. Red borders indicate the samples where the intervention is present. The first row of taxa (tax1, tax2, tax3) are nonnull with negative, positive, and negative effects, respectively, and the bottom row of taxa are all null.

3.2 Model settings and metrics

Given these data, we gather performance metrics associated with normalization, forecasting, and inference approaches. For normalization, we consider working with the original, untransformed data, the DESeq2 size-factor normalized data (Love *and others*, 2014), and the size-factor normalized data followed by an asinh transformation (Callahan *and others*, 2016; Jeganathan and Holmes, 2021). The latter two transformations account for potentially different library sizes across simulated samples and the fact that negative binomial data can be heavily skewed. For forecasting, we apply MDSINE2, fido with kernel parameters $\rho = \sigma = 0.5$ and $\rho = \sigma = 1$, and mbtransfer with $Q = P = 2$ and $Q = P = 4$. As mentioned above, MDSINE2 is an alternative to the gLV model designed to account for microbiome community dynamics (Gibson *and others*, 2021), and fido is a logistic-normal Gaussian Process model (Silverman *and others*, 2022) (see Supplementary Section 7.2 for details). All the models were provided with host and perturbation-related covariates. For MDSINE2, we forecast by integrating the learned dynamics over future timepoints, setting the initial conditions equal to the current test sample's microbiome community profile[2]. We consider 3 normalizations $\times$ 5 models $\times$ 27 datasets, but exclude MDSINE2 on runs with 400 taxa due to consistently long computation times. This results in 378 simulation configurations.

We compare the cross-validated forecasting performance of the models across the simulation settings. We divide the 50 simulated subjects into $K = 4$ folds. For each iteration k, models are trained with the $\mathbf{y}_t^{(i)}, \mathbf{w}_t^{(i)}$ from all subjects except those in the holdout fold. On holdout folds, we reveal all timepoints up to the first intervention t^* in the currently held-out subject. The trained models then forecast the community profiles up to a time horizon of $H = 5$. We provide access to intermediate interventions $\mathbf{w}_{t+h}^{(i)}$, but not community compositions $\mathbf{y}_{t+h}^{(i)}$ for $h > t^*$. For each iteration k, we compute the mean absolute error across lags, holdout subjects, and taxa:

$$MAE_k = \frac{1}{JH}\frac{1}{|\mathcal{D}_{-k}|}\sum_{j=1}^{J}\sum_{h=1}^{H}\sum_{i\in\mathcal{D}_{-k}}\left|y_{j(t^*+h)}^{(i)} - \hat{f}_j^{+h}\left(y_{j(t^*-P+h):(t^*+h-1)}^{(i)}, \mathbf{w}_{(t^*-Q+h+1):(t^*+h)}^{(i)}, \mathbf{z}^{(i)}\right)\right|.$$

We also evaluate inferential quality using false discovery proportions and power. Specifically, for instantaneous effects at lag $h = 0$, we compute the false discovery proportion and power as:

$$\text{FDP}(0) = \frac{\left|J_0 \cap \hat{J}(0)\right|}{\left|\hat{J}(0)\right|}, \qquad \text{Power}(0) = \frac{\left|J_1 \cap \hat{J}(0)\right|}{|J_1|},$$

where $\hat{J}(0)$ are the taxa flagged as having immediate intervention effects and $J_1(0)$ are the rows of B_0 with at least one nonnull effect: $\cup_d\{j : B_{0,jd} \neq 0\}$. For delayed effects, we must account both for taxa with nonzero entries of B_q for $q > 0$ and also those taxa that, though not directly influenced through B_q, are indirectly shifted by autoregressive links A_p with taxa that are affected by the intervention. To this end, we recursively define:

$$J_1(h) = \left\{j : \text{row } j \text{ of } \prod_{p=1}^{h} A_p \mathbf{1}_{J_1(h-p)} \text{ has at least one nonzero element}\right\} \bigcup \left\{j : B_{h,jd} \neq 0 \text{ for some } d\right\},$$

where $\mathbf{1}_{J_1(h-p)} \in \{0,1\}^J$ is an indicator over taxa that are nonnull at lag $h - p$. $J_0(h)$ is defined as the complement of $J_1(h)$. The mirror-selected taxa at delay h are denoted $\hat{J}(h)$, and they can be compared with $J_1(h)$ and $J_0(h)$ to define FDP(h) and Power(h).

[2]We use `md2.integrate` as discussed in this MDSINE2 documentation.

3.3 Results

Figure 4 summarizes cross-validated forecasting performance on DESeq2-asinh transformed data. We also discuss alternative transformations below. Error rates increase with the proportion of nonnull taxa $1-\pi_0$ and signal strength b. This increase is likely a consequence of the high variance shifts in $\mathbf{y}_t$ during interventions for these settings. MDSINE2's performance is consistently worse than either fido or mbtransfer's. Figure 3 sheds light on this. It shows prediction error for individual holdout subjects in one of the simulation settings; residual error in other simulation settings is qualitatively similar. We average errors across all taxa and truncate those with a magnitude greater than 50. This figure shows that minor errors in MDSINE2's initial forecast become amplified at larger horizons. This behavior is not universal, but its effects on the subset of subjects where it does appear are strong enough to explain MDSINE2's deterioration in our simulation setup. In retrospect, such behavior is unsurprising – MDSINE2 can only refer to one step in the past, and it must have either exponential growth or decay until the community reaches its carrying capacity. Though this behavior does not affect inferences for taxon-perturbation relationships, which are the main focus of (Gibson *and others*, 2021), it can limit the usefulness of the forecasts needed to simulate hypothetical trajectories. In contrast, mbtransfer and fido can refer to historical windows, supporting more realistic intervention analysis: The second day of a microbiome intervention does not necessarily have the same consequences as the first day.

When the intervention effect has a smaller magnitude or is limited to fewer taxa, fido and mbtransfer perform comparably. In other cases, mbtransfer is more accurate. We interpret this by noting that, despite its ability to incorporate interventions as covariates, fido's Gaussian Process assumption enforces smoothness in the predicted values. This smoothness prevents the model from capturing the sharp changes in abundance within these simulation settings. Since the simulation's true autoregressive dynamics have $P = Q = 3$, the fitted mbtransfer models are misspecified. Interestingly, the $P = Q = 2$ model slightly outperforms the $P = Q = 4$ model. In this context, the reduction in variance from having a slightly less flexible model outweighs the reduction in bias from modeling larger lags.

Analogous results for alternative transformations are available in Supplementary Figures 10 and 11. When the data are not asinh transformed, the mbtransfer model performs worse than either MDSINE2 or fido. This reversal is consistent with the use of a squared-error loss in the underlying gradient boosting model, which is not adapted to count data. fido should be preferred if data must be modeled on the original scale. However, we note that transformations are often well-justified in microbiome analysis, and an increasing number of formal methods implement them (McKnight *and others*, 2018; Chen *and others*, 2018; Jiang *and others*, 2021). Finally, Supplementary Figure 12 gives the average computation time across folds. MDSINE2 is slower than either fido or mbtransfer. fido and mbtransfer have comparable computation times except when using DESeq2-asinh transformed data. In this setting, fido is noticeably faster, but mbtransfer provides better forecasts.

Figure 5 summarizes inferential performance. When considering longer time horizons, all methods have improved FDR control because more taxa become truly nonnull as instantaneous effects propagate across the community. However, DESeq2 never appears to control the FDR at the prespecified level of $q = 0.2$. Though DESeq2 assumes a negative binomial generative mechanism across taxa, it treats samples as independent. This misspecification likely contributes to the poor FDR control seen in this simulation. For larger b, the mirror statistics may appear conservative, with many $\widehat{\text{FDP}} \ll 0.2$. These settings often correspond to high power, though, and the signal may have simply become easy to detect. Mirrors control the FDR when the DESeq2-asinh transformation is applied, and the number of taxa is large. This is expected, because the DESeq2-asinh transformation improves forecasts, and the need for a large number of taxa is consistent with Proposition 3.3 of Dai *and others* (2020), which demonstrates FDR control only asymptotically. We attribute the strong performance of mirror statistics to the fact that its false discovery rate control is adapted to the current dataset of interest rather than a previously defined probabilistic model.

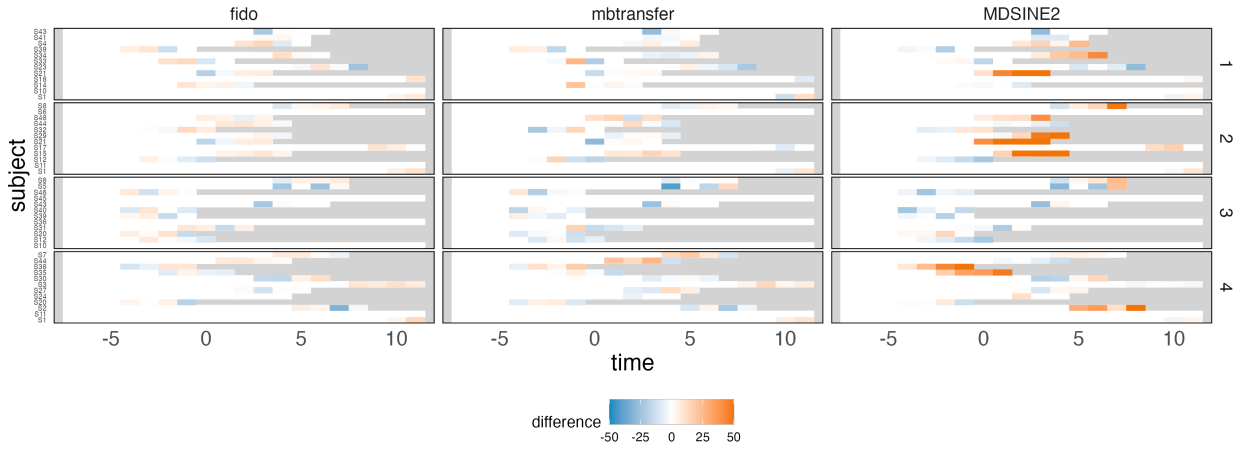

Figure 3: A comparison of forecasting residuals across four folds (rows) in one simulation run suggests that forward integrating the MDSINE2 model can lead to exponential increases in forecasting errors.

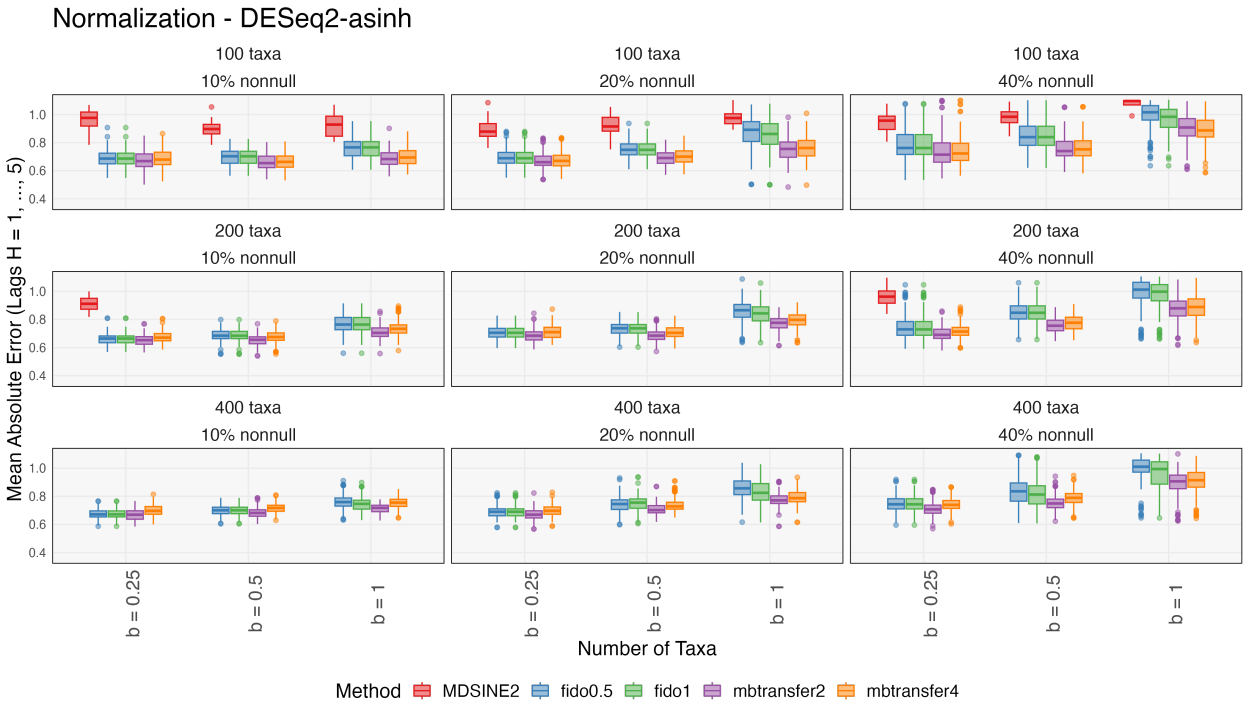

Figure 4: Forecasting errors across simulation settings. The y-axis corresponds to the average of MAE_k across folds. Within each panel, the signal strength b increases from left to right. Column panels give the proportion of taxa affected by the intervention, and rows have different numbers of taxa. Errors beyond 3× the interquartile range of those from fido and mbtransfer have been omitted from the view, excluding some outliers from MDSINE2. Runs that did not complete within 72 hours are not included – this explains the missing boxplot for MDSINE2 in the 200 taxa, 20% nonnull panel. The fido package is comparable to mbtransfer when the intervention strength is weak but deteriorates when the intervention is strong.

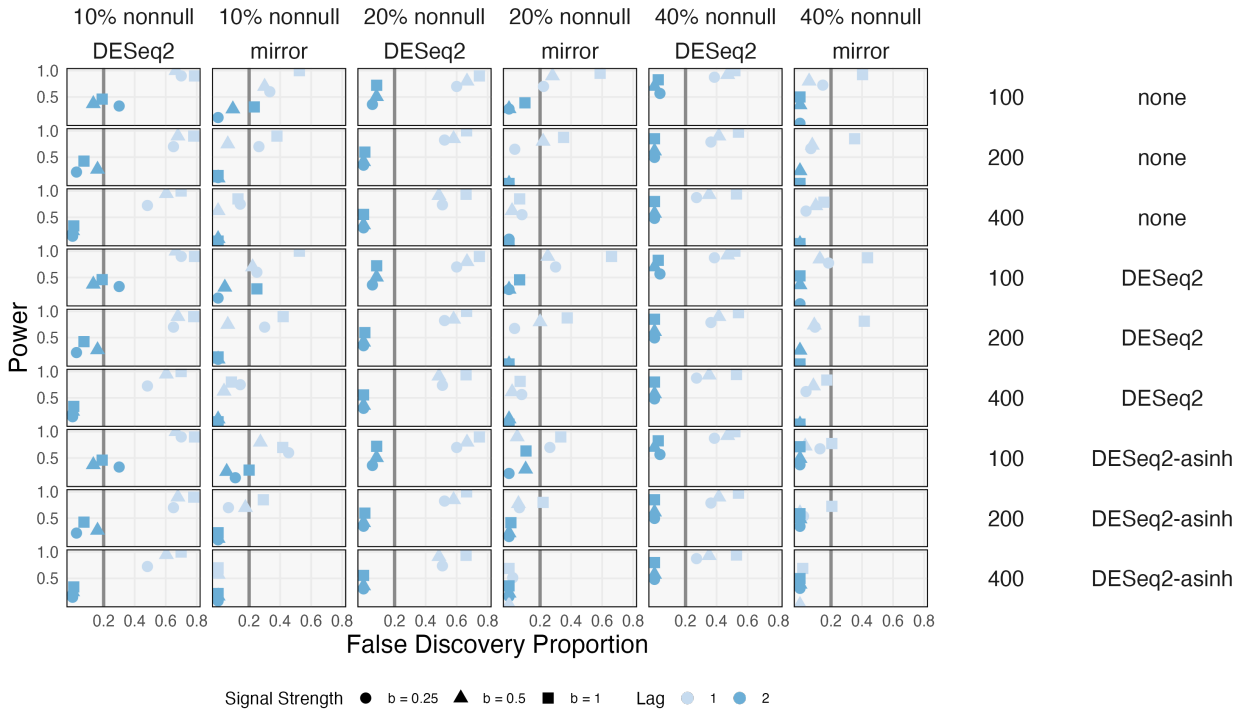

Figure 5: Inferential performance in the simulation experiment. Rows index different normalization methods and the total number of taxa. The target FDR in each case has been set to 0.2. Columns have varying proportions of nonnull hypotheses and compare DESeq2-based inferences with those from mirror statistics. DESeq2 does not provide FDR control for lag one effects in any simulation context. mbtransfer's mirror algorithm controls the FDR when using the DESeq2-asinh transformation data and when the number of taxa is sufficiently large.

4 Data Analysis

We next illustrate mbtransfer using three microbiome datasets. The data are drawn from two human and one animal microbiome studies. They include an experimentally-defined intervention, one that arises in the natural progression of a prospective study, and a shift in continuous ecosystem parameters. The studies also vary in their taxonomic richness, number of subjects, and total number of timepoints. Despite the various domains and intervention types considered, each study focuses on how environmental change remodels the microbiome.

4.1 Diet and the gut microbiome

David *and others* (2013) investigated the sensitivity of the human gut microbiome to brief diet interventions. To this end, they recruited 20 participants and randomly assigned them to either "plant" or "animal" interventions. Subjects in the two groups were required to maintain a plant- or animal-based diet during a five-day intervention window. Samples were collected for two weeks surrounding the intervention, typically at a daily frequency. Ultimately, 8 - 15 samples were collected for each participant since some timepoints were never successfully sampled. We linearly interpolate timepoints onto an even, daily sampling grid, motivated by the cubic spline interpolation adopted by Ruiz-Perez *and others* (2019). Regularly spaced timepoints are a fundamental limitation of discrete, autoregressive models. Initially, the data contained 17310 taxa. We filter to those present in at least 40% of the samples, reducing the number of taxa to 191 – a drastic reduction, but one consistent with distinguishing a "core" microbiome for more focused analysis (Shade and Handelsman, 2012; Neu *and others*, 2021).

We fit mbtransfer with $P = Q = 2$ and $\mathbf{w}_t^{(i)} = (\mathbb{I}(t \in \text{Animal Shift}), \mathbb{I}(t \in \text{Plant Shift})) \in [0,1]^2$, setting two intervention series[3] and omitting any host features $z^{(i)}$. Figure 6 shows in- and out-of-sample forecasts. Forecasting performance deteriorates out of sample, highlighting the between-participant heterogeneity in this study. This challenge is most pronounced within the lowest quantile of abundance. Nonetheless, out-of-sample forecasts are still clearly correlated with the truth. In both the in and out-of-sample contexts, forecasts on shorter time horizons are more accurate. In addition, performance seems better in the more highly abundant taxa. The gradient boosting model's least squares training objective likely deteriorates in sparse data.

We compute mirror statistics for time lags $h = 1, \ldots, 4$ to evaluate the effect of a four day shift to an animal diet. Supplementary Figure 15 shows the associated mirror statistic distributions. The increasing magnitude across lags for some taxa suggests that the diet intervention effects are not instantaneous but build up over consecutive treatment days. To support interpretation, Figure 7a shows the median difference between counterfactual trajectories for a subset of significant taxa. The taxa were chosen by applying principal component analysis to the simulated trajectory differences, projecting onto the first component, and selecting every sixth taxon according to that ordering. Some taxa (e.g., OTU000006) have more immediate but transient effects, while others (e.g., OTU000065) have more gradual but sustained changes. Further, in several taxa (e.g., OTU000118, OTU000012), a long-run increase follows an initial decrease, which is corroborated by the associated subject-level data. Within taxa, we found that the first and third quartiles of the counterfactual differences across subjects tended to agree. This suggests that the model has not learned interactions between the intervention effects and past composition – the effect of the diet intervention may be uniform across various initial community states. The main benefit of a transfer function modeling approach is the model's capacity to learn different shapes of counterfactual trajectories while still controlling a precise notion of FDR.

For comparison, the original, interpolated data for a subset of taxa is shown in Figure 7b. These views are consistent with the counterfactual trajectories, but they are obfuscated by the higher degree of sampling noise and require more space. Our results are in line with David *and others* (2013), but we can

[3]It is possible for these series to lie between 0 and 1 because some interpolated timepoints lie in the transitions between active and inactive periods – see Figure 1.

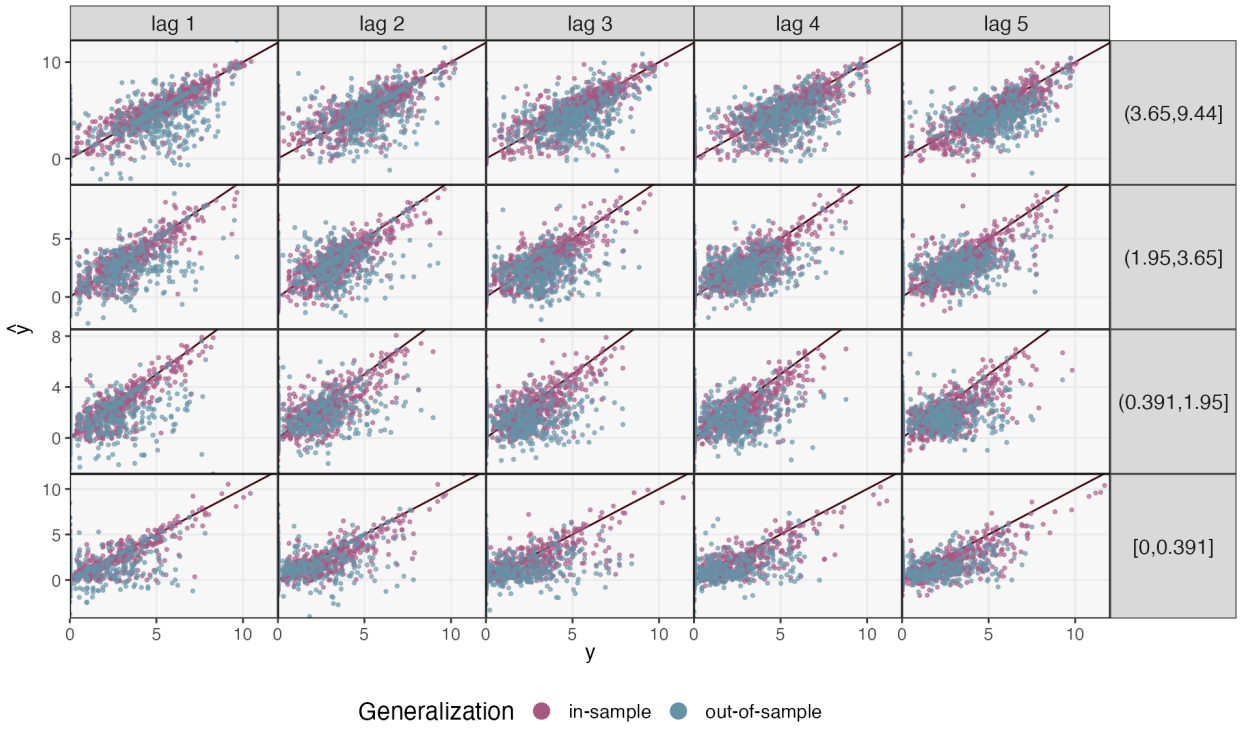

Figure 6: Forecasting error for an mbtransfer model applied to the diet intervention dataset of Section 4.1. The y-axis is faceted by quantiles of abundance and the x-axis is faceted by time horizon h. In-sample error refers to errors made on new timepoints for individuals who appeared in the training data, while out-of-sample predictions are made on individuals who did not appear during training. Performance is strongest in shorter time horizons and for more abundant taxa.

clearly describe ecosystem dynamics by modeling temporal dependence between the diet intervention and microbiome community profiles.

4.2 Birth and the vaginal microbiome

We next re-analyze data from Costello *and others* (2022), which studied how birth influences the composition of the mother's vaginal microbiome. Supplementary Figure 13 shows in and out-of-sample forecasting accuracy. Compared to Figure 6 in the previous analysis, in and out-of-sample performances are more comparable, reflecting the larger sample size of this study. The derived mirror statistics are shown in Supplementary Figure 16. Compared to the diet intervention, more of the ecosystem is shifted by the birth intervention. We generate four counterfactual trajectories for all subjects to understand how birth influences individual taxa and whether any effects are modulated by contraception use. Specifically, we compute $\hat{f}^{+h}\left(\mathbf{y}_{(t^*-P-1):(t^*-1)}\right), \tilde{\mathbf{w}}_{(t^*-Q):t^*}, \tilde{z}\right)$ for $\tilde{\mathbf{w}}_{(t^*-Q):t^*)} \in \{\mathbf{1}_Q, \mathbf{0}_Q\}$ representing presence or absence of the birth event and $\tilde{z} \in \{0, 1\}$ denoting re-initiation of contraceptive use following birth. Figure 8a suggests the absence of an interaction with contraceptive use. This may be a consequence of the fact that 57% of subjects were missing any data on contraceptive use – though the examples discussed in Costello *and others* (2022) consider plausible mechanisms for how contraception can influence the postpartum microbiome, our model does not detect a generalizable enough association to learn the interaction.

Like in the diet intervention, we can distinguish between response trajectories. Members of genus *Lactobacillus* are clearly depleted, while other taxa appear to take advantage of the postpartum environment. For example, *Porphyromonas* appears briefly during the same window that the *Lactobacilli* disappear. Figure 8b compares these trajectories with real data. As before, we see that the learned trajectories denoise the original data, and consistent with the lack of interaction, we do not observe obvious, systematic associations between postpartum community trajectory and contraception use.

4.3 pH and the aquaculture microbiome

We next use mbtransfer in a problem with continuous intervention values. Yajima *and others* (2022) studied the taxonomic composition of the eel aquaculture microbiome, collecting water samples every 24 hours for 128 days from five aquaculture tanks. We can view the tank's pH and eel activity scores as continuous inputs $\mathbf{w}_t$ to a transfer function model. Based on the five tanks' longitudinal data, Yajima *and others* (2022) concluded that the microbiome composition changes over time and is related to various environmental factors. Moreover, there was a substantial shift in pH in three of the tanks, and we analyze how this shift influenced community composition.

After preprocessing the 16S data, we have 345 samples and 128 taxa. We interpolate the sampling times to fill in some missing days. Then, we fit a mbtransfer with $P = 4$ and $Q = 4$. We simulate counterfactual pH series that are constant for ten timepoints, with values varying between pH = 2 and = 9. Figure 9 shows scatterplots of pH vs. abundance for a subset of taxa that were found to be significant when contrasting the two extremes, pH = 2 and 9. Moreover, we can also simulate taxa trajectories for each counterfactual pH series (Figure 9b). The effects seem additive, with counterfactual abundances smoothly varying as a function of pH. The taxa with the clearest associations (e.g., X_0001, X_0010) appear to have larger variation in their simulated trajectories, which agrees with their being more sensitive to changes in pH.

5 Software

We created the mbtransfer R package for analyzing interventions using transfer function modeling. This package provides various functionalities, including the creation of an S4 object called `ts_inter`, handling intervention windows, conducting counterfactual simulations, and performing inference based on mirror

[5]Specifically, we include panels for every sixth selected taxon after sorting according to the first dimension of the PCA taken on simulated trajectories.

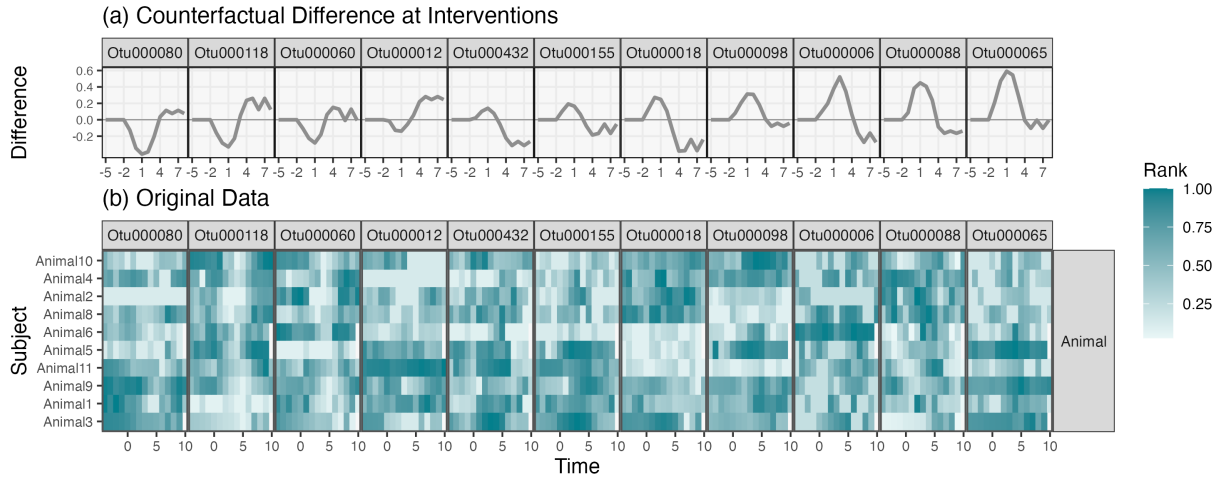

Figure 7: (a) Counterfactual difference in simulated trajectories for a subset[5]of the selected taxa in the diet data in Section 4.1. (b) Subject-level data from a subset of taxa appearing in (a). Each heatmap row is a subject, and each column is a timepoint. These data are consistent with the interpretations from the counterfactual simulation. For example, OTU000006 has more transient increases in abundance (e.g., Animal1, and Animal6) while OTU000065 has more prolonged departures (e.g., Animal3 and Animal 9).

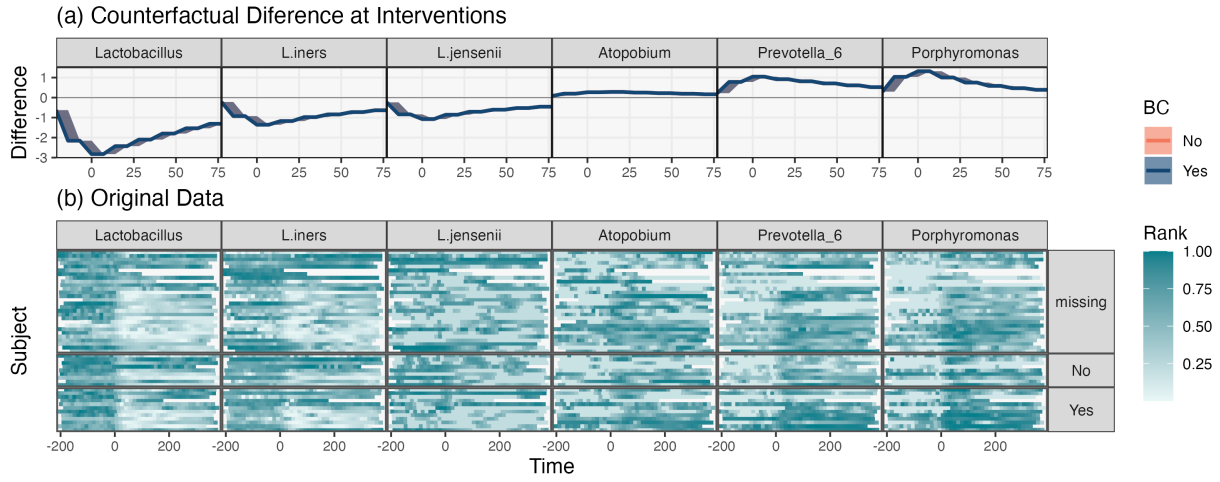

Figure 8: (a) Counterfactual differences for a subset of selected taxa from the re-analysis of (Costello *and others*, 2022). Counterfactual differences are computed for each subject in the data, and bands represent the first and third quartiles of differences across subjects. Since the bands for birth control reinitiation overlap, we conclude that the model does not learn the interaction effects between the intervention and contraception use. (b) The corresponding subject-level data. Rows are grouped according to the birth control reinitiation survey response.

statistics. The block below creates an example `ts_inter` container that unifies intervention time series and microbiome profiles,

We can generalize the time delay for any number of input covariates. In addition, if there are nonuniform sampling time points, we can use `interpolate()` to get different resolutions of sampling:

After setting up the data, we can train a transfer function-based boosting model. The model assumes that interventions at different time lags impact microbial communities within a specified range of time lags. For instance, we can train the model with $P = 3$ and $Q = 3$, where P represents the maximum time lags affecting the communities and Q denotes the range of intervention time lags,

Once the model is trained, we can forecast future community composition. `predict()` will fill in any time points that are present in the `intervention` slot, which allows forecasts across different time horizons. For example, the block below fills in predicted compositions from the ninth time point on,

Following model evaluation, we can identify taxa with instantaneous or delayed intervention effects. This can be achieved by simulating counterfactual alternatives and employing partial dependence mirror statistics to control the false discovery rate. We can obtain a list of intervention series by specifying the characteristics of the hypothetical interventions (e.g., step interventions starting at specific time points and with defined lengths).

Using these series and the trained model, we can select taxa with significant effects based on a specified false discovery rate.

As transfer function-based boosting models can incorporate various covariates, this framework enables the identification of factors that influence shifts in each taxon.

6 Discussion

mbtransfer adapts transfer function models to the dynamic microbiome context. The approach is flexible and interpretable, enabling intervention analysis without assuming a restrictive functional form and supporting the simulation of counterfactual trajectories. We have complemented our modeling approach with a formal inferential mechanism, leveraging recent advances in selective inference. A simulation study illuminated our method's properties across data-generating settings, and our data analysis highlighted its practical application in three contrasting microbiome studies.

We anticipate several directions for further study. First, we have focused attention on developing mirror statistics for detecting temporal effects, but the same strategy could be generalized to support inference for inter-species relationships and host-microbiome interactions. Indeed, the construction of mirror statistics via partial dependence profiles depends only on having access to a simulator f that can generate hypothetical responses. This simulator could be used to contrast profiles with alternative initial states $\mathbf{y}_t$ or host features $\mathbf{z}$. Similarly, the procedure could clarify interactions between several concurrent interventions or scales (Fukuyama *and others*, 2021; Sankaran and Holmes, 2022). Second, it would be valuable to develop a transfer function model that learns the entire distribution of responses $p\left(\mathbf{y}_t | \mathbf{y}_{(t-P-1):(t-1)}, \mathbf{w}_{(t-Q+1):t}\right)$, rather than simply the mean. Such a probabilistic analog would allow us to quantify uncertainty in intervention effects. Characterizing the uncertainty in intervention effects is especially valuable in the design of potential probiotics – an intervention with moderate but consistent effects may be preferable to one with strong but erratic ones (Thompson *and others*, 2022; Fannjiang *and others*, 2022; Jeganathan *and others*, 2018). Finally, an extension to continuous time autoregressive processes would allow us to model irregular sampling frequencies, removing the need for the interpolation steps performed above.

We have synthesized a variety of statistical concepts to address a recurring microbiome data analysis challenge: How can we quantify the influence of environmental shifts on a microbial ecosystem? The transfer function perspective has guided our intervention analysis, and we linked the resulting nonlinear models with a modern computational inference technique based on mirror statistics. This facilitates the stability and attribution analysis critical for forming scientific conclusions (Efron, 2020; Yu, 2018). As microbiome studies continue to investigate more and more nuanced questions about ecosystem dynamics, similarly formal simulation and inference methods will likely play an essential role.

7 Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgements

Support for this research was provided by the University of Wisconsin - Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

P. Jeganathan received funding from the Faculty of Science at McMaster University.

Conflict of Interest: None declared.

References

Biecek, Przemyslaw and Burzykowski, Tomasz. (2021, March). *Explanatory model analysis*, Chapman & Hall/CRC Data Science Series. London, England: CRC Press.

Box, G. E. P. and Tiao, George C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association* **70**, 70–79.

Bucci, Vanni, Tzen, Belinda, Li, Ning, Simmons, Matthew, Tanoue, Takeshi, Bogart, Elijah, Deng, Luxue, Yeliseyev, Vladimir, Delaney, Mary L., Liu, Qing, Olle, Bernat, Stein, Richard R., Honda, Kenya, Bry, Lynn *and others*. (2016). Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses. *Genome Biology* **17**.

Calgaro, Matteo, Romualdi, Chiara, Waldron, Levi, Risso, Davide and Vitulo, Nicola. (2020). Assessment of statistical methods from single cell, bulk rna-seq, and metagenomics applied to microbiome data. *Genome Biology* **21**.

Callahan, Benjamin J., Sankaran, Kris, Fukuyama, Julia, McMurdie, Paul J. and Holmes, Susan P. (2016). Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research* **5**, 1492.

Center for High Throughput Computing. (2006). Center for high throughput computing.

Chen, Li, Reeve, James, Zhang, Lu, Huang, Shengbing, Wang, Xuefeng and Chen, Jun. (2018). Gmpr: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* **6**.

Chen, Tianqi and Guestrin, Carlos. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Costello, Elizabeth K., DiGiulio, Daniel B., Robaczewska, Anna, Symul, Laura, Wong, Ronald J., Shaw, Gary M., Stevenson, David K., Holmes, Susan P., Kwon, Douglas S. and Relman, David A. (2022). Longitudinal dynamics of the human vaginal ecosystem across the reproductive cycle. *bioRxiv*.

Dai, Chenguang, Lin, Buyu, Xing, Xin and Liu, Jun S. (2020). False discovery rate control via data splitting. *Journal of the American Statistical Association*.

David, Lawrence A., Maurice, Corinne F., Carmody, Rachel N., Gootenberg, David B., Button, Julie E., Wolfe, Benjamin E., Ling, Alisha V., Devlin, A. Sloan, Varma, Yug, Fischbach, Michael A., Biddinger, Sudha B., Dutton, Rachel J. *and others*. (2013). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559 – 563.

Efron, Bradley. (2020). Prediction, estimation, and attribution. *International Statistical Review* **88**, S28 – S59.

Fannjiang, Clara, Bates, Stephen, Angelopoulos, Anastasios Nikolas, Listgarten, Jennifer and Jordan, Michael I. (2022). Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences of the United States of America* **119**.

Friedman, Jerome H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* **29**, 1189–1232.

Fukuyama, Julia, Sankaran, Kris and Symul, Laura. (2021). Multiscale analysis of count data through topic alignment. *Biostatistics*.

Gerber, Georg K. (2014). The dynamic microbiome. *FEBS Letters* **588**.

Gibbons, Sean M., Kearney, Sean M., Smillie, Chris S. and Alm, Eric J. (2017). Two dynamic regimes in the human gut microbiome. *PLoS Computational Biology* **13**.

Gibson, Travis E., Kim, Younhun, Acharya, Sawal, Kaplan, David E., DiBenedetto, Nicholas, Lavin, Richard, Berger, Bonnie, Allegretti, Jessica R., Bry, Lynn and Gerber, Georg K. (2021). Intrinsic instability of the dysbiotic microbiome revealed through dynamical systems inference at scale. *bioRxiv*.

Jeganathan, Pratheepa, Callahan, Benjamin J., Proctor, Diana M, Relman, David A. and Holmes, Susan P. (2018). The block bootstrap method for longitudinal microbiome data. *arXiv: Methodology*.

Jeganathan, Pratheepa and Holmes, Susan P. (2021). A statistical perspective on the challenges in molecular microbial biology. *Journal of Agricultural, Biological and Environmental Statistics* **26**, 131 – 160.

Jiang, Ruochen, Li, Wei Vivian and Li, Jingyi Jessica. (2021). mbimpute: an accurate and robust imputation method for microbiome data. *Genome Biology* **22**.

Love, Michael I., Huber, Wolfgang and Anders, Simon. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology* **15**.

McKnight, Donald T., Huerlimann, Roger, Bower, Deborah S., Schwarzkopf, Lin, Alford, Ross A. and Zenger, Kyall R. (2018). Methods for normalizing microbiome data: An ecological perspective. *Methods in Ecology and Evolution* **10**, 389 – 400.

Neu, Alexander T., Allen, Eric E. and Roy, Kaustuv. (2021). Defining and quantifying the core microbiome: Challenges and prospects. *Proceedings of the National Academy of Sciences of the United States of America* **118 51**.

Ruiz-Perez, Daniel, Lugo-Martinez, Jose, Bourguignon, Natalia, Mathee, Kalai, Lerner, Betiana, Bar-Joseph, Ziv and Narasimhan, Giri. (2019). Dynamic Bayesian networks for integrating multi-omics time-series microbiome data. *bioRxiv*.

Sankaran, Kris and Holmes, Susan P. (2022). Generative models: An interdisciplinary perspective. *Annual Review of Statistics and Its Application*.

Shade, Ashley and Handelsman, Jo. (2012). Beyond the venn diagram: the hunt for a core microbiome. *Environmental microbiology* **14 1**, 4–12.

Silverman, Justin D., Durand, Heather K., Bloom, Rachael J., Mukherjee, Sayan and David, Lawrence A. (2018). Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* **6**.

Silverman, Justin D, Roche, Kim, Holmes, Zachary C, David, Lawrence A and Mukherjee, Sayan. (2022). Bayesian multinomial logistic normal models through marginally latent matrix-t processes. *Journal of Machine Learning Research* **23**.

Thompson, Jaron, Zavala, Victor M. and Venturelli, Ophelia S. (2022). Integrating a tailored recurrent neural network with Bayesian experimental design to optimize microbial community functions. *bioRxiv*.

Xie, Fang and Lederer, Johannes. (2021). Aggregating knockoffs for false discovery rate control with an application to gut microbiome data. *Entropy* **23**.

Yajima, Daii, Fujita, Hiroaki, Hayashi, Ibuki, Shima, Genta, Suzuki, Kenta and Toju, Hirokazu. (2022). Core species and interactions prominent in fish-associated microbiome dynamics. *Microbiome* **11**.

Yu, Bin. (2018). Three principles of data science: predictability, computability, and stability (pcs). *2018 IEEE International Conference on Big Data (Big Data)*, 4–4.

Zhu, Zifan, Fan, Yingying, Kong, Yinfei, Lv, Jinchi and Sun, Fengzhu. (2021). Deeplink: Deep learning inference using knockoffs with applications to genomics. *Proceedings of the National Academy of Sciences* **118**.

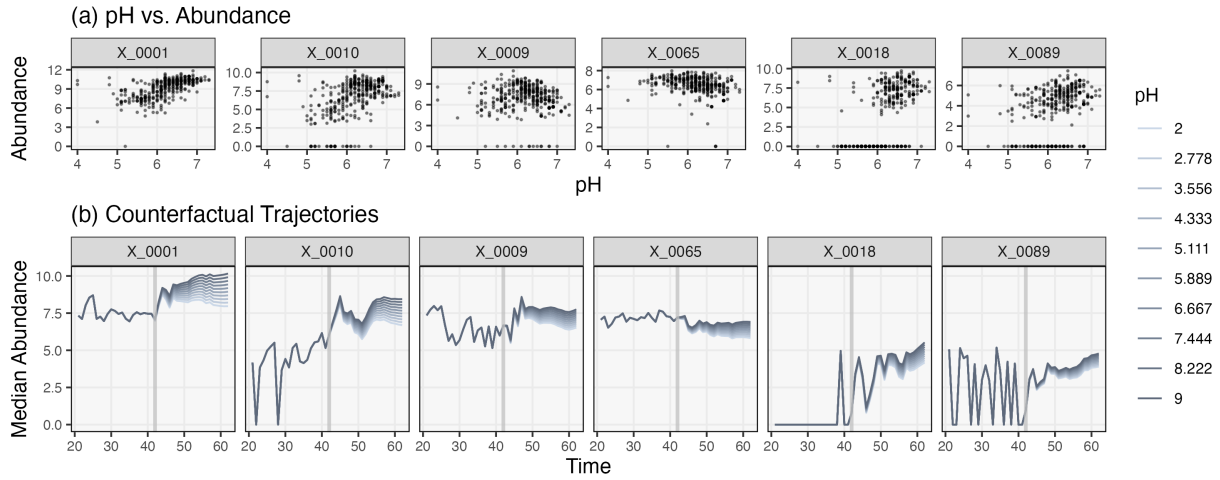

Figure 9: (a) Associations between pH and abundance for a subset of taxa that are selected by the mirror algorithm. Taxa have been sorted in decreasing order according to the magnitude of their associated mirror statistics. (b) Counterfactual trajectories under pH shifts. A sustained pH shift is imagined starting at day 42. Different values of pH tend to influence the magnitude, but not the shape, of the forecast trajectories

Supplementary Materials

7.1 Reproducibility

Simulation experiments

- A Dockerfile that installs software used in the experiments is available at https://go.wisc.edu/eovk4b. The image can be pulled from DockerHub using `docker pull krisrs1128/mi:20230506`.
- Simulation inputs have been saved at https://go.wisc.edu/8ey754. They were generated using https://go.wisc.edu/37y6ny. This script also includes source code for Figure 2.
- Each forecasting and inference simulation run corresponds to one `run_id` of these Rmarkdown notebooks: I, II. Simulation outputs have been saved at https://go.wisc.edu/3gc982.
- Figures 3, 5, and Supplementary Figures 10, 11, 12 were generated using this script https://go.wisc.edu/87876s applied to the previous outputs. Figure 4 was generated using https://go.wisc.edu/l1n79m.

Case studies

- The case studies appear as vignettes in our accompanying R package (https://go.wisc.edu/crj6k6, I, II, III). They can also be rerun without installing the package by visiting this binder notebook: https://go.wisc.edu/emxv33.
- Processed versions of the data used in all case studies can be found on figshare: https://go.wisc.edu/7ig8q8, https://go.wisc.edu/q827o9, https://go.wisc.edu/83l84r. The data were processed according to this script https://go.wisc.edu/37x8hh.

7.2 Summary of MDSINE2 and FIDO

This subsection briefly describes the MDSINE2 (Bucci *and others*, 2016; Gibson *and others*, 2021) and fido (Silverman *and others*, 2022) methods that were used in the simulation study.

MDSINE2

MDSINE2 is a recently proposed Bayesian model of microbiome dynamics. It adapts the generalized Lotka-Volterra dynamics in the following ways,

1. Taxa are assigned to clusters. This effectively reduces the dimensionality of the gLV's autoregressive dynamics – taxa influence one another's growth rates via their cluster membership. Moreover, perturbation effects are constrained to be identical across all taxa within the same cluster.
2. The approach is probabilistic and models each sample's total count and relative abundance structure. This contrasts with alternative gLV estimators, which often proceed by initially transforming abundance and applying regularized least squares.

We provide a high-level overview of the model's generative mechanism. If we assume uniform sampling over time, then the model generates latent taxonomic abundances according to

$$\log x_j^{(i)}(t+1) \,|\mu_j^{(i)}(t) \sim \mathcal{N}\left(\log \mu_{s,k}(t+1), \sigma^2\right)$$

where i, j and t index subjects, taxa, and time. The mean vector $\mu_j^{(i)}$ is a deterministic, gLV-like function of random clustering and growth parameters,

$$\log \mu_j^{(i)}(t+1) := \log x_j^{(i)}(t) + a_{1,j}\left[1 + \sum_{p=1}^{P} \gamma_{c_j} \mathbf{z}_{c_j,p}^{(\gamma)} \mathbf{1}\left\{(i,t) \in \texttt{Perturbation } p\right\}\right] - a_{2,j} x_j^{(i)}(t) + \sum_{j':c_{j'} \neq c_j} b_{c_j c_{j'}} \mathbf{z}_{c_j c_{j'}}^{(b)} x_{j'}^{(i)}(t)$$

The terms $a_{1,j}$ and $a_{2,j}$ are the growth and decay rates for taxon j, as in the standard gLV. c_j is the cluster index of taxon j. The summation of perturbations p describes the influence of different perturbations on taxon j's abundance. γ_{c_j} and $\mathbf{z}_{c_j,p}^{(\gamma)}$ parameterized the strength and presence of a perturbation p effect on cluster c_j – note that the perturbation influences are shared across all members of the same cluster. The final summation describes autoregressive dynamics between pairs of taxa j, j'. The autoregressive coefficients are shared between all pairs of taxa with the same cluster assignments $c_j, c_{j'}$. In this way, the autoregressive dynamics operate at the cluster level.

These latent taxonomic abundances $x_j^{(i)}(t)$ are transformed into the observed relative $y_j^{(i)}(t)$ and total $r_i(t)$ abundances for taxon j in sample t of subject i using a negative binomial measurement model,

$$y_j^{(i)}(t) \mid \left(x_j^{(i)}(t)\right)_{j=1}^{J} \sim \text{NB}\left(\frac{r_i(t)\, x_j^{(i)}(t)}{\sum_{j'} x_{j'}^{(i)}(t)}, d_1 + d_0\left(\frac{x_j^{(i)}(t)}{\sum_{j'} x_{j'}^{(i)}(t)}\right)^{-1}\right)$$

where d_0 and d_1 are hyperparameters set in advance to account for dispersion in the observed samples.

Finally, priors are placed on the parameters $a_{1,j}, a_{2,j}, c_j, \gamma_l, \mathbf{z}_l^{(b)}, \mathbf{z}_l^{(\gamma)}$, and $b_{ll'}$. A stick-breaking process prior is placed on c_j, allowing for the number of clusters to adapt to the evidence in the data. Inference is performed through MCMC, cycling over these parameters and those used within hyperpriors.

fido

fido combines a multinomial logistic-normal model, matrix-normal process, and Bayesian inference to model the effect of covariates on taxa abundance.

- The multinomial proportions are transformed into real space. Then, the mean of the transformed data is assumed to be latent matrix-normal processes. This approach allows for the modeling of latent factors that capture the shared information across the taxa.
- The Bayesian framework starts with specifying priors for taxa covariance. Then, during each iteration of the collapse-uncollapse sampler, the `fido` updates the latent factors that capture the shared information across the taxa using latent - T process (LTP). In addition, the regression coefficients are updated that relate the latent factors and covariates, including perturbations.

Let's assume there are J taxa, Q covariates, and N total samples. The model generative process for k-th taxa abundance at time t in subject i, $y_j^{(i)}(t)$ is as follows.

1. The prior distribution of the covariance between additive log-ratio transformed (ALR) taxa is inverse Wishart distribution, $\Sigma_{J-1\times J-1} \sim W^{-1}(\Xi, \upsilon)$ with a scale matrix $\Xi_{J-1\times J-1}$ and degrees of freedom υ.
2. Then, the smooth mean function $\Lambda[X] \sim \text{GP}(\Theta[X], \Sigma, \Gamma[X])$ relates the covariates $X_{Q\times N}$ to ALR-transformed η with the mean function $\Theta[X]$, row (taxa) covariance Σ, and column (samples) covariance $\Gamma[X]$.

- Covariates X includes time t.
- $\Gamma[X]$ evaluates the kernel K at time points t and $t-1$ for a given subject s, $K(X_s(t), X_s(t-1))$ otherwise, it is zero.

3. Next, $\eta \sim \mathrm{N}(\Lambda[X], \Sigma, I_N)$ is a normal-matrix distribution, where $\pi = \phi^{-1}(\eta)$ and η is a $(J-1) \times N$ real valued matrix.

4. For the subject i at time t, we compute the inverse of ALR, $\pi^{(i)}(t)$ and generate $\mathbf{y}^{(i)}(t)$

$$\mathbf{y}^{(i)}(t) | \pi^{(i)}(t) \sim \text{Multinomial}\left(n^{(i)}(t), \pi^{(i)}(t)\right),$$

where $n^{(i)}(t)$ is the library size.

7.3 Supplementary Figures

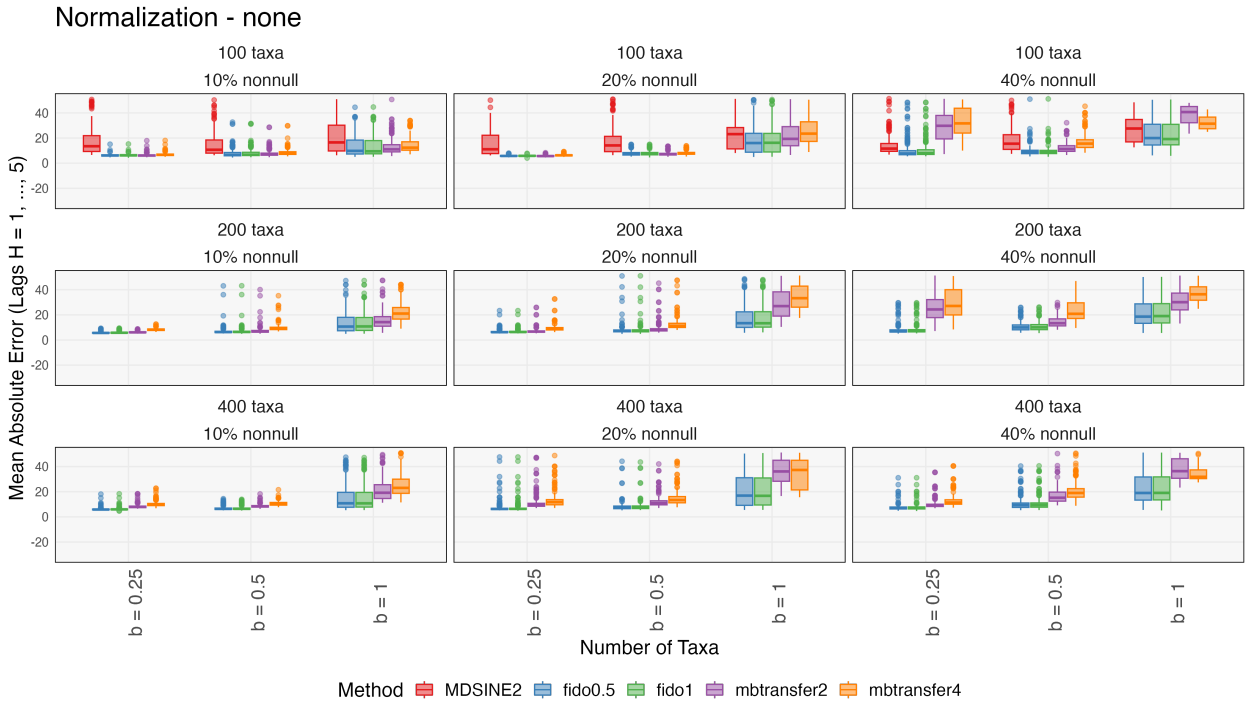

Figure 10: The analog of Figure 3 when not using any normalization.

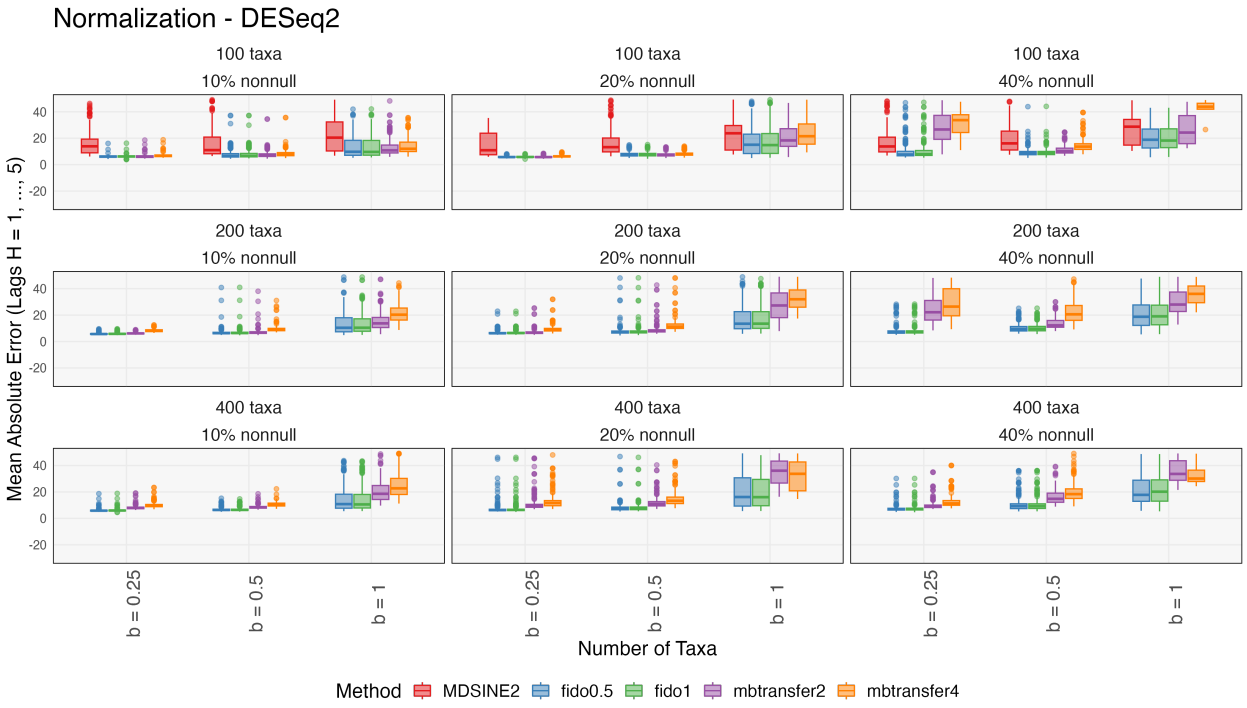

Figure 11: The analog of Figure 3 when using DEseq2 size-factor normalization.

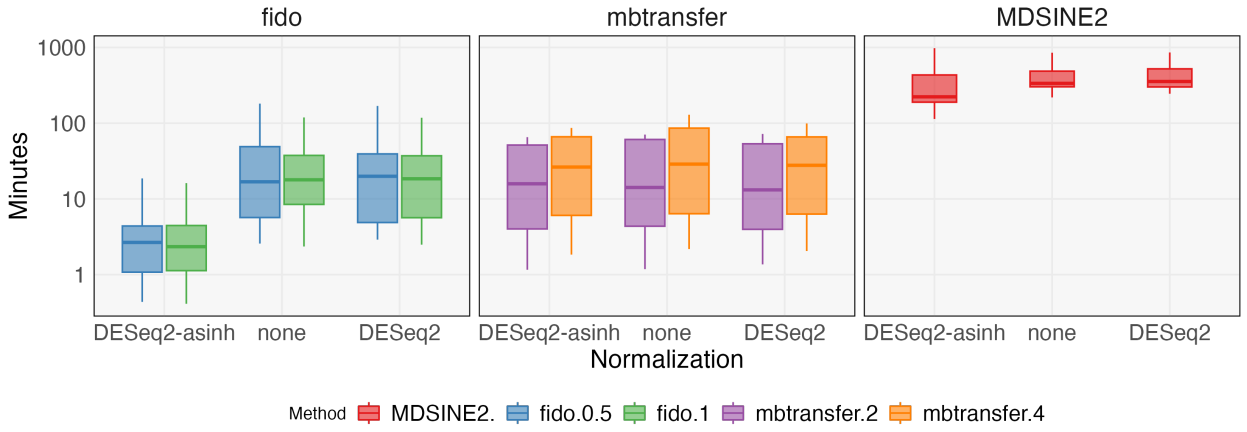

Figure 12: Computation times for methods considered in the simulation experiment. fido is fast on untransformed, count data, which is the context in which it was originally designed. mbtransfer is comparable to fido on transformed data. Both packages are an order of magnitude faster than MDSINE2.

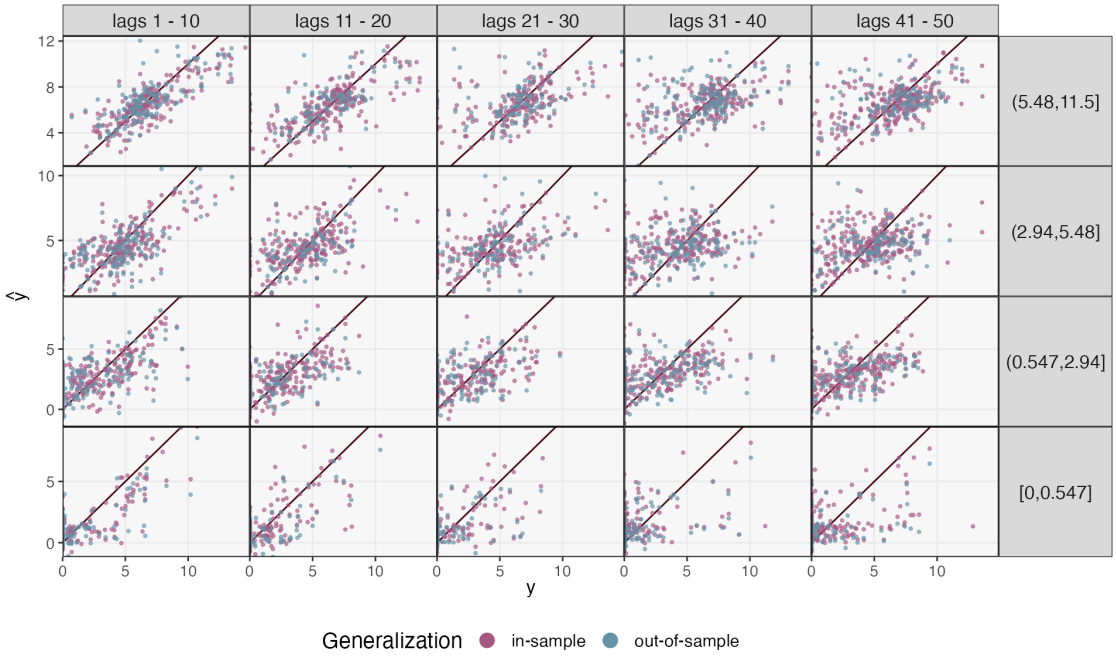

Figure 13: The analog of Figure 6 for the postpartum case study in Section 4.2. Each point is one sample. In-sample error refers to errors from future timepoints of subjects observed in the training data. Out-of-sample errors are those on previously unobserved subjects. As before, errors predictions are most accurate on nearby time lags and more abundant taxa.

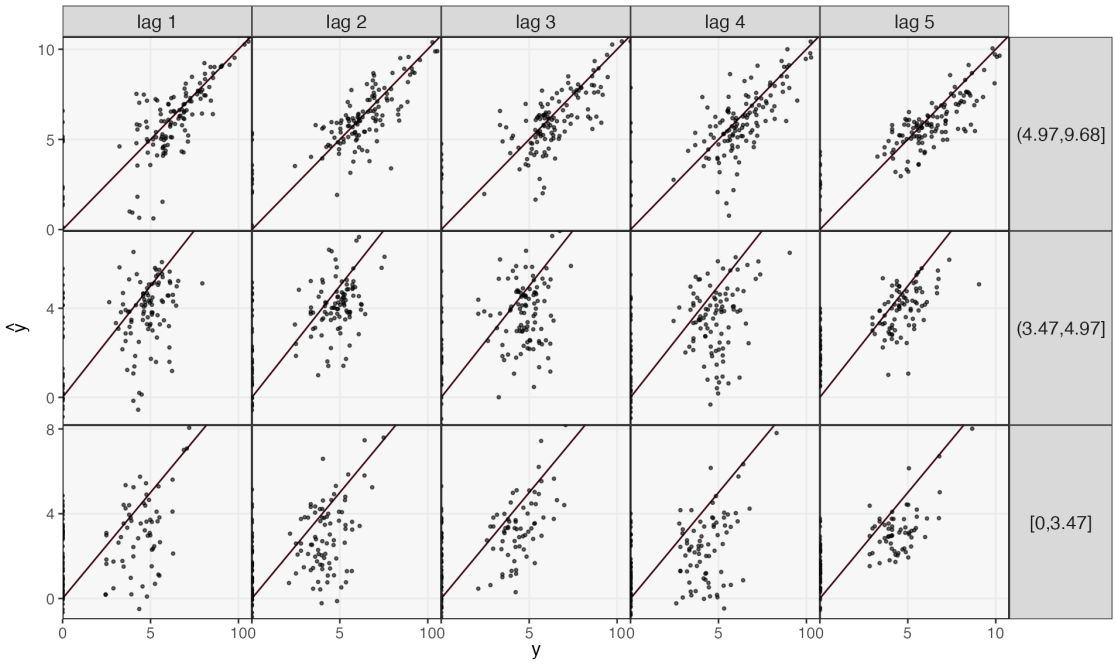

Figure 14: The analog of Figure 6 for the aquaculture case study in Section 4.3. We only show in-sample errors, because our analysis only considers three unique tanks.

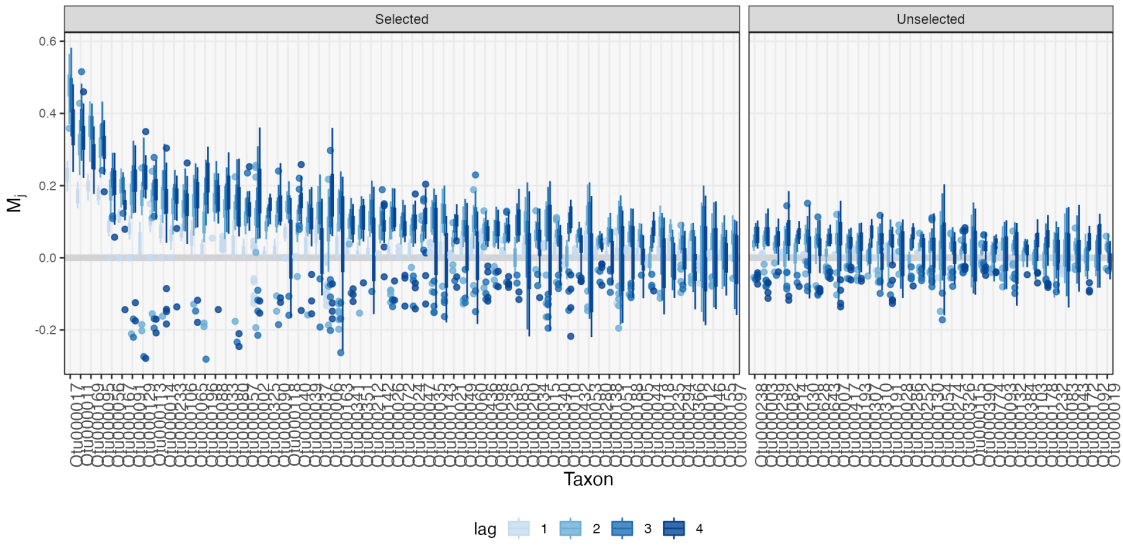

Figure 15: The distribution of mirror statistics M_j for all selected and a subset of unselected taxa. Larger statistics indicate strong, consistent lag-0 effects (specifically, $PD_j(0)$ for taxon j) across data splits. The selection threshold is chosen adaptively according to a false discovery proportion estimate. The unselected taxa shown are those with the largest median M_j, and we have shown as many as possible while limiting the total number of boxplots to 100.

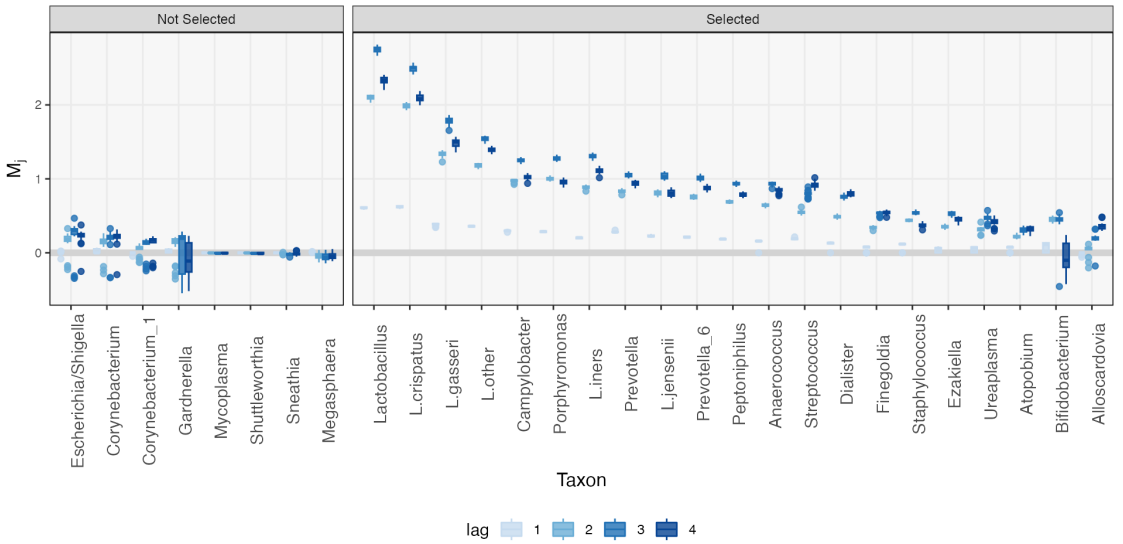

Figure 16: The analog of Supplementary Figure 15 for the mirror statistics in the postpartum case study in Section 4.2. Mirror statistics appear more concentrated, likely a consequence of the larger sample size and stronger effects visible in this dataset.

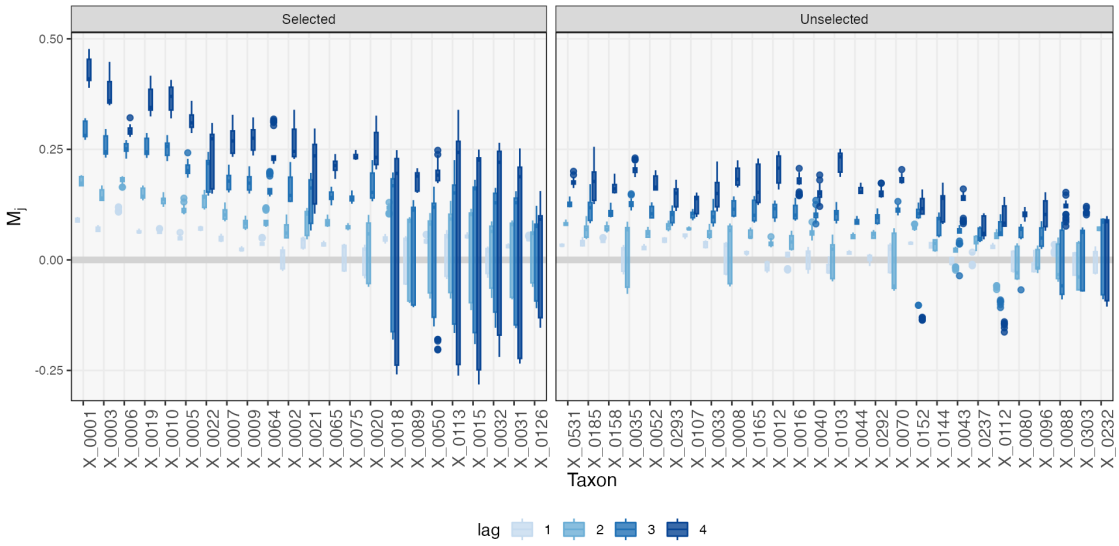

Figure 17: The analog of Supplementary Figure 15 for the aquaculture case study in Section 4.3.Note that we plot a taxon in the selected panel if it is significant for any lag. This explains why some taxa are selected despite having higher-lag mirror distributions that are symmetric around zero.