Efficient Bayesian Additive Regression Models For Microbiome and Gene Expression Studies

Tinghua Chen[1], Michelle Pistner Nixon,[1], Justin D. Silverman[1,2,3]

[1]College of Information Science and Technology, Pennsylvania State University; [2]Department of Statistics, Pennsylvania State University; [3]Department of Medicine, Pennsylvania State University

1. Abstract

Analyzing sequence count data, such as microbiome or gene expression data, poses a challenge. Researchers often want to estimate linear and non-linear effects of covariates on microbial or gene composition. Bayesian multinomial logistic-normal (MLN) models have gained popularity due to their ability to account for the count compositional nature of these data. Moreover, compared to more common Multinomial-Dirichlet models, MLN models can represent both positive and negative covariation between taxa or genes. However, inferring MLN models can be challenging. Recently, we developed a computationally efficient and accurate approach to inferring MLN models with a Marginally Latent Matrix-T Process (MLTP) form. Our approach is based on a particle filter with marginal Laplace approximation – called the *Collapse-Uncollapse* (CU) sampler. Here, we introduce a new class of MLN Additive Gaussian Process models (*MultiAddGPs*) for additive deconvolution of overlapping linear and non-linear effects. MultiAddGPs can be efficiently inferred using the CU sampler. Furthermore, we develop an efficient approach to hyperparameter selection via Maximum Marginal Likelihood estimation. We validate our approach via both simulated and real data studies.

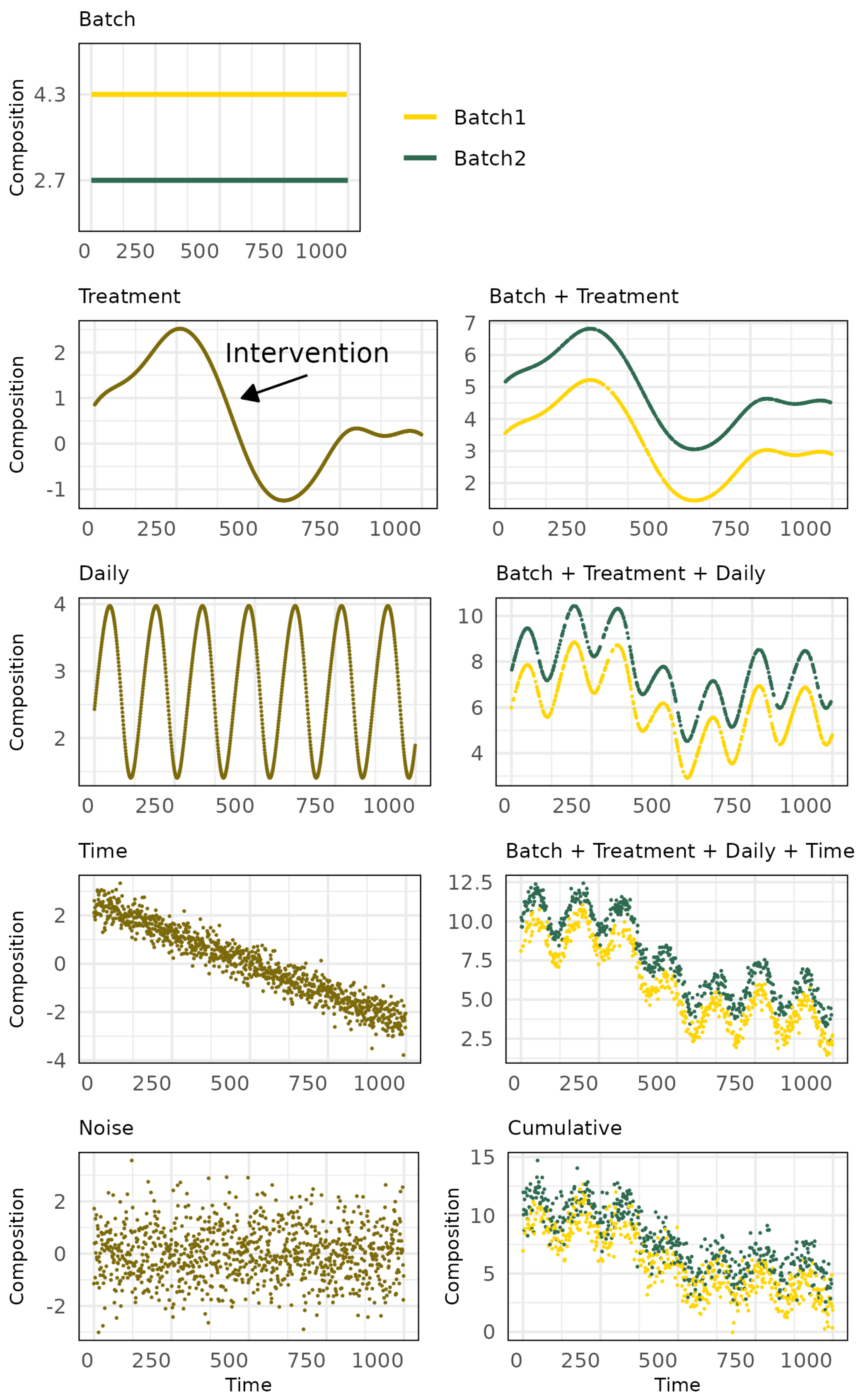
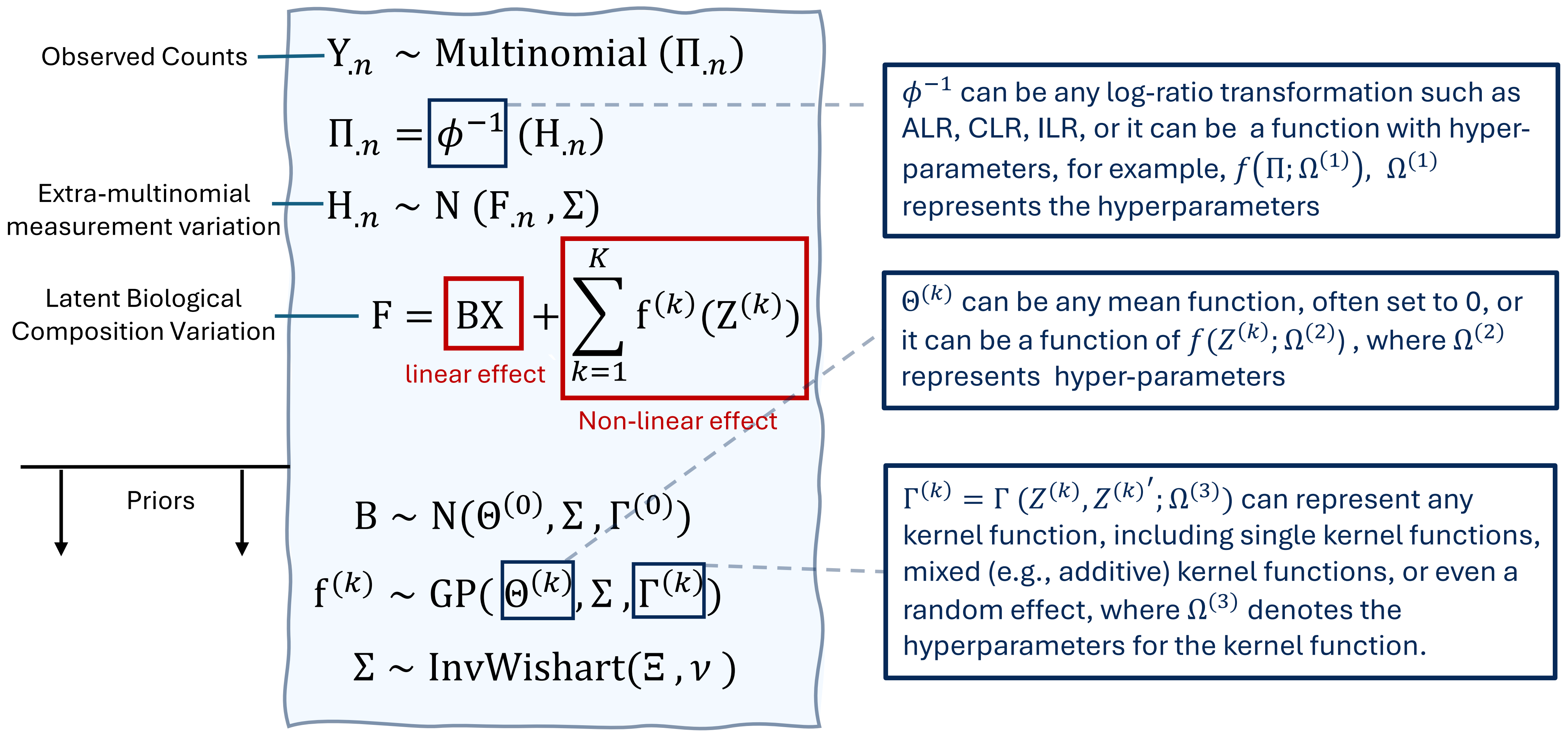


Figure 1. **Overlapping linear and non-linear effects can confound inference.** Host-associated microbiota experiences daily changes in response to factors such as diet, disease, and drugs. Robust statistical methods are necessary to disentangle the effects of multiple measured factors on microbiota.

2. MultiAddGPs Framework



Observed Counts — $Y_{.n} \sim \text{Multinomial}(\Pi_{.n})$

$\Pi_{.n} = \boxed{\phi^{-1}}(H_{.n})$

Extra-multinomial measurement variation — $H_{.n} \sim N(F_{.n}, \Sigma)$

Latent Biological Composition Variation — $F = \underbrace{\boxed{BX}}_{\text{linear effect}} + \underbrace{\boxed{\sum_{k=1}^{K} f^{(k)}(Z^{(k)})}}_{\text{Non-linear effect}}$

Priors

$B \sim N(\Theta^{(0)}, \Sigma, \Gamma^{(0)})$

$f^{(k)} \sim GP(\boxed{\Theta^{(k)}}, \Sigma, \boxed{\Gamma^{(k)}})$

$\Sigma \sim \text{InvWishart}(\Xi, \nu)$

ϕ^{-1} can be any log-ratio transformation such as ALR, CLR, ILR, or it can be a function with hyper-parameters, for example, $f(\Pi; \Omega^{(1)})$, $\Omega^{(1)}$ represents the hyperparameters

$\Theta^{(k)}$ can be any mean function, often set to 0, or it can be a function of $f(Z^{(k)}; \Omega^{(2)})$, where $\Omega^{(2)}$ represents hyper-parameters

$\Gamma^{(k)} = \Gamma(Z^{(k)}, Z^{(k)'}; \Omega^{(3)})$ can represent any kernel function, including single kernel functions, mixed (e.g., additive) kernel functions, or even a random effect, where $\Omega^{(3)}$ denotes the hyperparameters for the kernel function.

- The observed sequence count data ($\mathbf{Y}$) is the result of drawing from an unobserved composition (Π).
- The log-ratio transformed latent composition ($\mathbf{F}$) is modeled using additive linear and non-linear regression models.
- Posterior samples of $p(f^{(1)}, \ldots, f^{(K)}, B, \Sigma \mid Y)$ are obtained through the Collapse-Uncollapse sampler (CU sampler).

3. Hyperparameter Selection via Maximum Marginal Likelihood

Marginal Likelihood can be efficiently and accurately approximated via the CU sampler's Laplace approximation.

$$\log \int p(\mathbf{H}, \mathbf{Y} \mid \Omega) d\mathbf{H} \approx \frac{(D-1)N}{2} \log(2\pi) + \log p(\hat{\mathbf{H}}_\Omega, \mathbf{Y} \mid \Omega) - \frac{1}{2} \log(-|\nabla^2[vec(\hat{\mathbf{H}}_\Omega)]|)$$

- Ω : a set of hyperparameters, including parameters in the kernel, mean function, or even log-ratio transformation ϕ.
- $\hat{\mathbf{H}}_\Omega$: the MAP estimate for $\mathbf{H}$ evaluated at the observed set $\{1, \ldots, N\}$ which depends on the hyperparameters Ω

We use Bayesian optimization to optimize the above approximate marginal likelihood.

4. Low-Dimensional Simulation Study

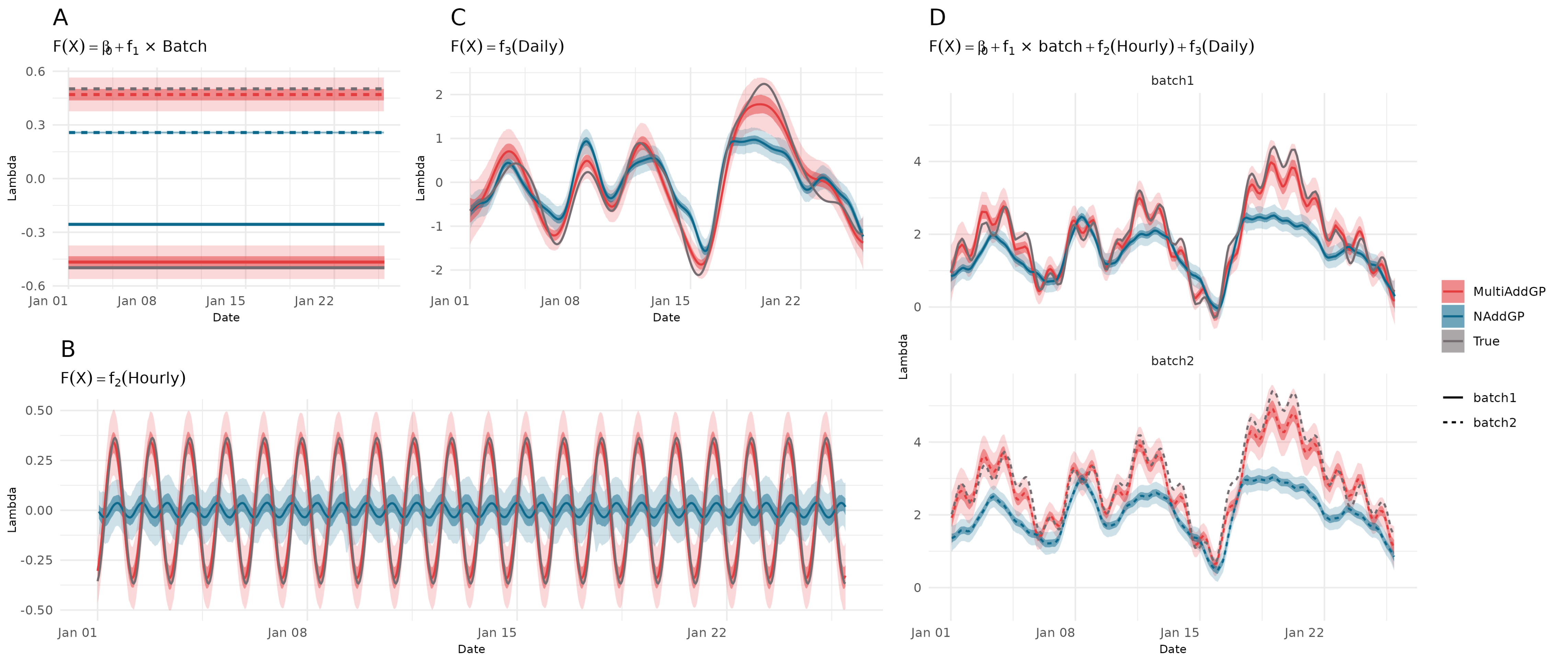


Figure 2. **Model comparison via simulated data**. MultiAddGPs faithfully partition variation due to multiple overlapping linear and non-linear effects. Others have proposed additive Gaussian Process models for analyzing microbiome data yet those models assume the data is transformed Gaussian (Cheng et al., 2019). By ignoring uncertainty due to counting, such approaches can lead to spurious inference. We illustrate this point using the NAddGP model which is the same as the MultiAddGP model but uses the following data normalization: $\mathbf{H}_{\cdot n} = \phi(Y_{\cdot n} + 0.5)$.

5. High-Dimensional Simulation Study

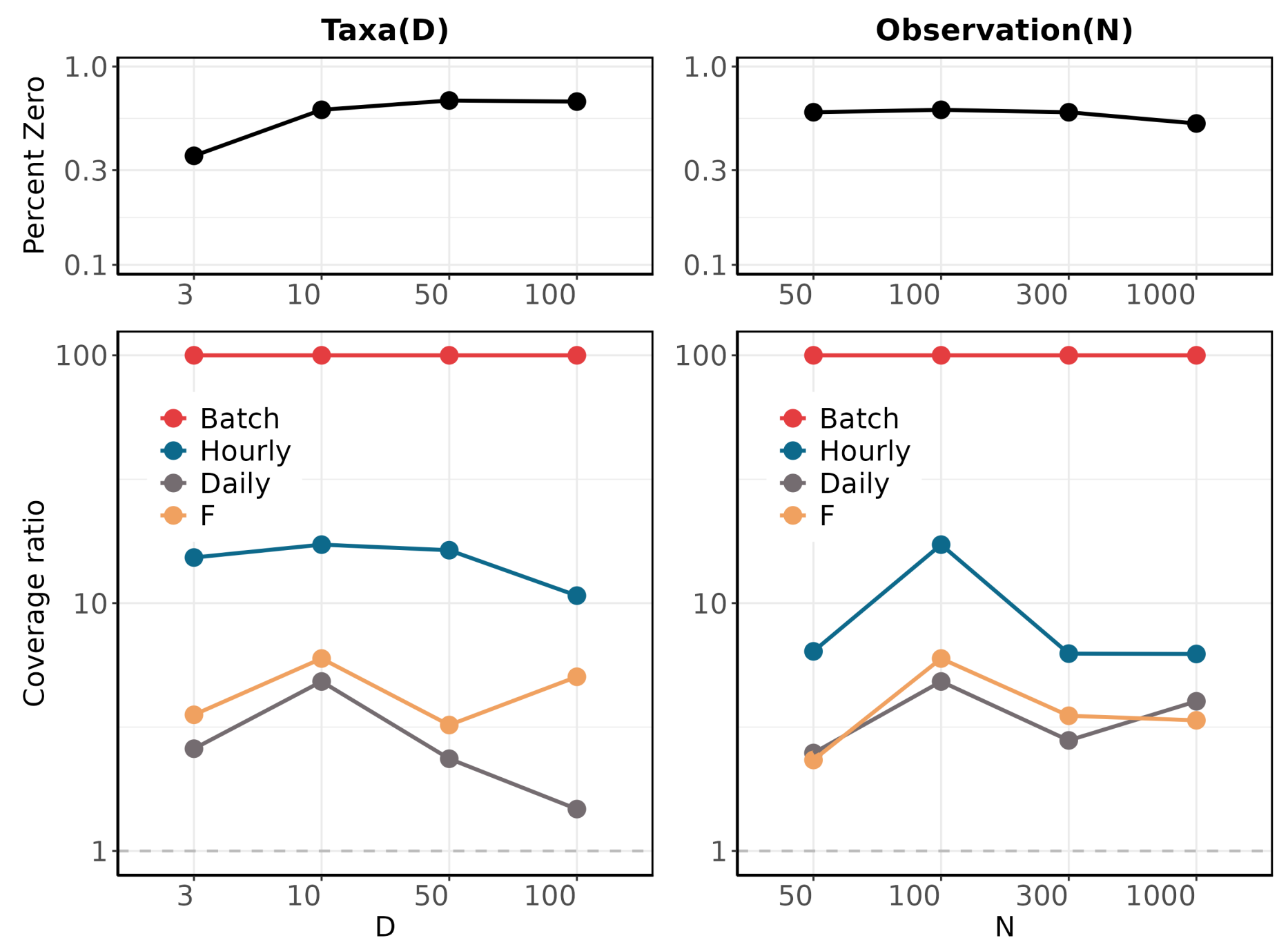


Figure 3. **The results of Figure 2 hold in higher dimensions**. We compare the performance of the MultiAddGPs and NAddGPs on higher-dimensional simulations based on the relative frequency with which posterior 95% confidence intervals cover the truth. We summarise this comparison as a coverage ratio – values greater than 1 show that posterior intervals of the MultiAddGPs cover the truth more frequently than corresponding intervals of the NAddGPs model. Data sparsity (percentage of zero counts) increases with increasing dimension D but does not vary with dimension N. In all simulations, MultiAddGPs covered the truth more frequently than NAddGPs.

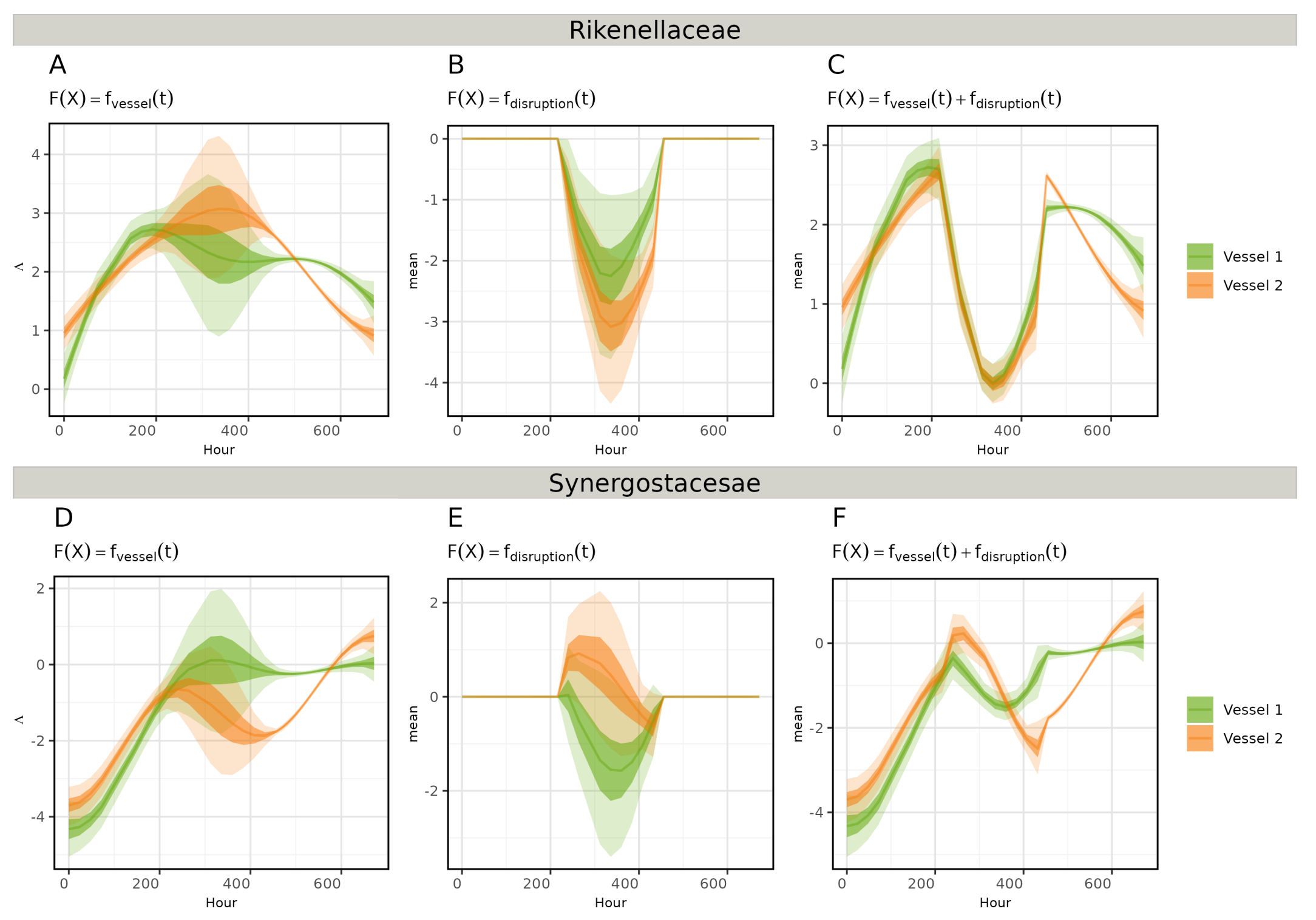


Figure 4. **MultiAddGPs model provides novel insight into the dynamic of artificial gut data**. This study analyzes mixed daily and hourly sampling from four artificial gut models over one month. Between experiment days 11 and 13, the feed lines supplying fresh media to Vessels 1 and 2 clogged, leading to partial starvation in those vessels, while Vessels 3 and 4 remained unaffected. The posterior 95% probability regions for log-ratio balances involving the families Rikenellaceae and Synergusistaceae are displayed.

Key Insights

- MultiAddGPs can deconvolve overlapping linear and non-linear effects in sequence count studies.
- We have developed efficient and accurate inference for the family of MultiAddGPs models.
- MultiAddGPs have been integrated into the *fido* software package which is available on CRAN.

Code: `https://jsilve24.github.io/fido/index.html`