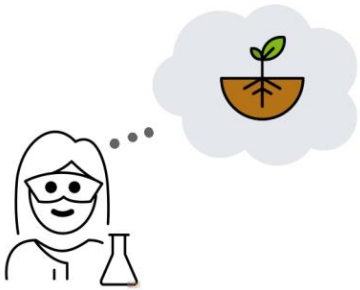
Efficient Bayesian Additive Regression Models For Microbiome Studies

Tinghua Chen, Michelle Nixon, Justin Silverman

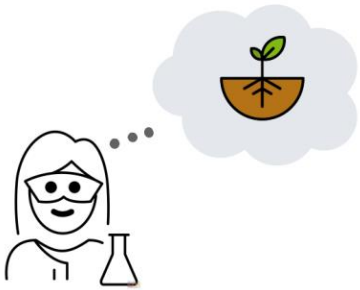
Silverman-Lab
Penn State University
tuc579@psu.edu

Oct 2 2024

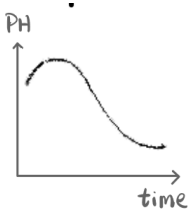
Motivated Example 1: Soil Microbiome



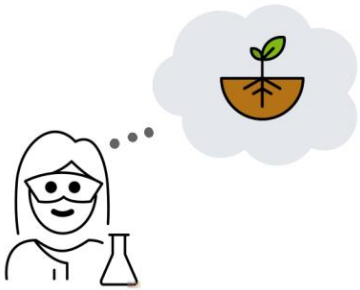
Motivated Example 1: Soil Microbiome



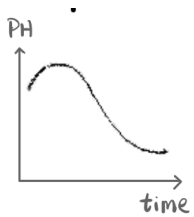
Soil Ph



Motivated Example 1: Soil Microbiome

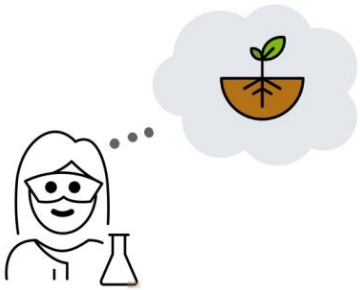


Soil Ph

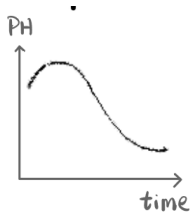


Soil Humidity

Motivated Example 1: Soil Microbiome

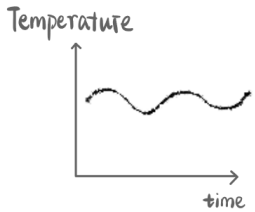


Soil Ph

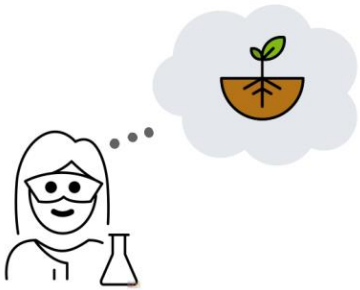


Soil Humidity

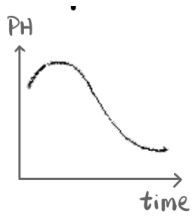
Temperature



Motivated Example 1: Soil Microbiome

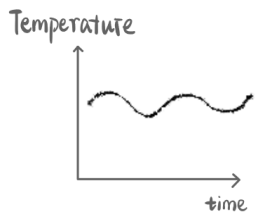


Soil Ph

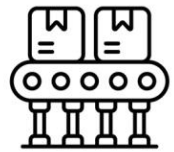


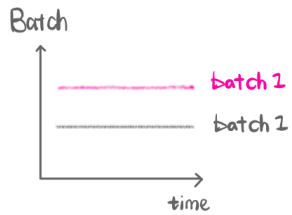
Soil Humidity

Temperature

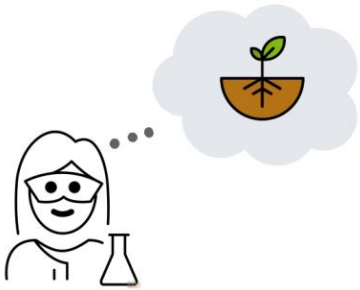


Batch

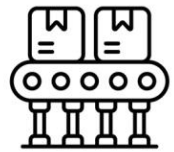


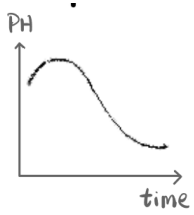


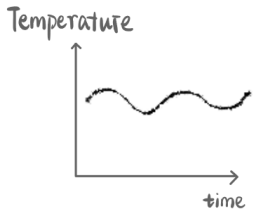
Motivated Example 1: Soil Microbiome

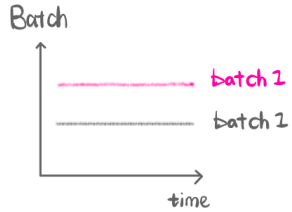


$=$ $f(\text{Soil Ph}) + f(\text{Soil Humidity}) + f(\text{Temperature}) + \beta \times \text{Batch}$

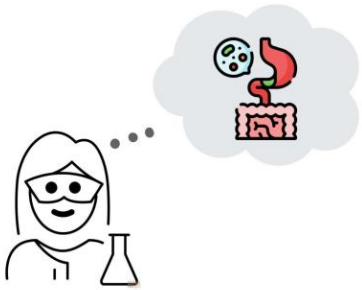




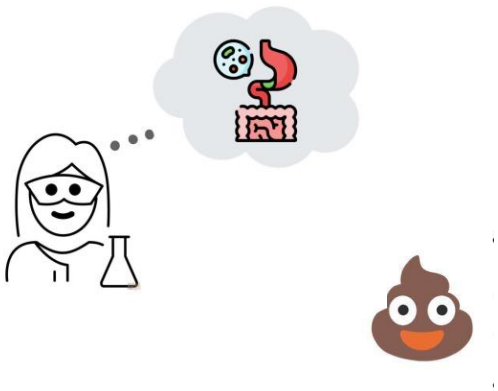




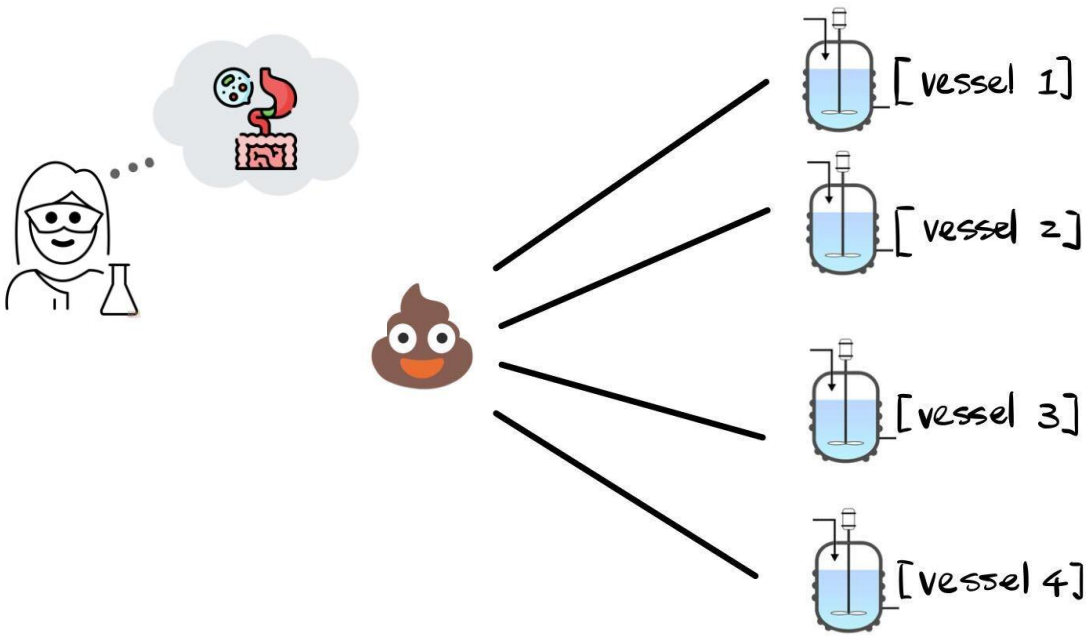
Motivated Example 2: Artificial Gut Data



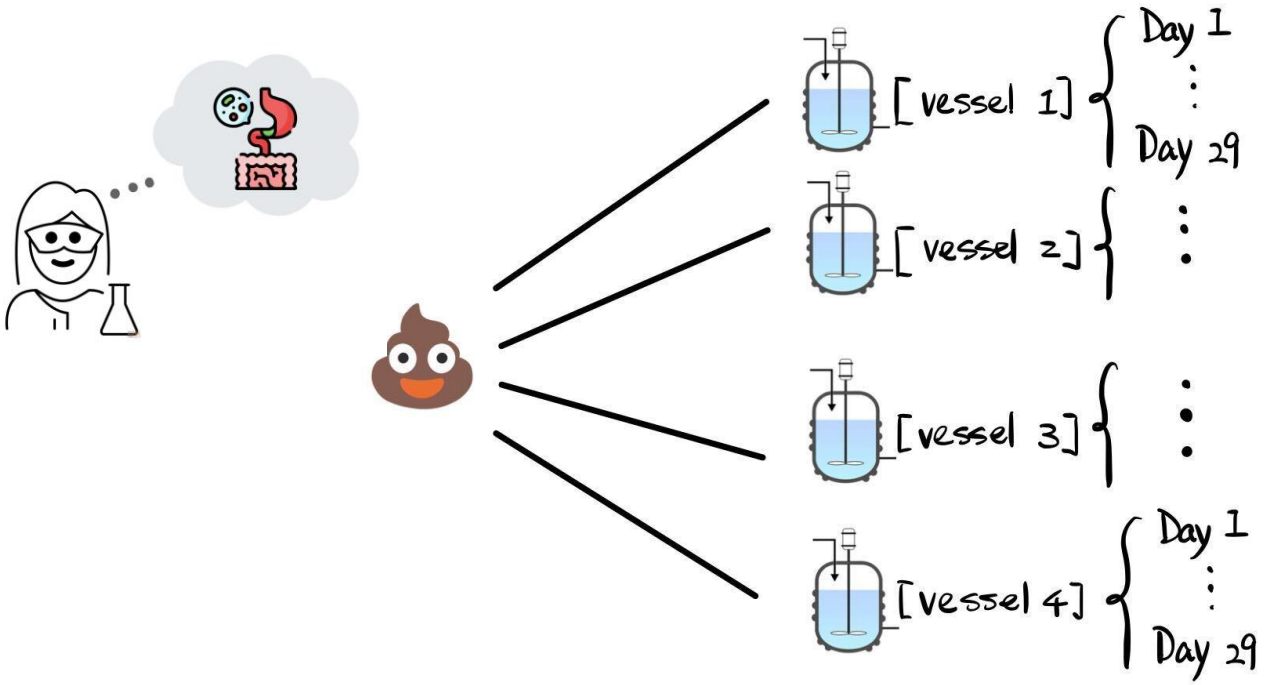
Motivated Example 2: Artificial Gut Data



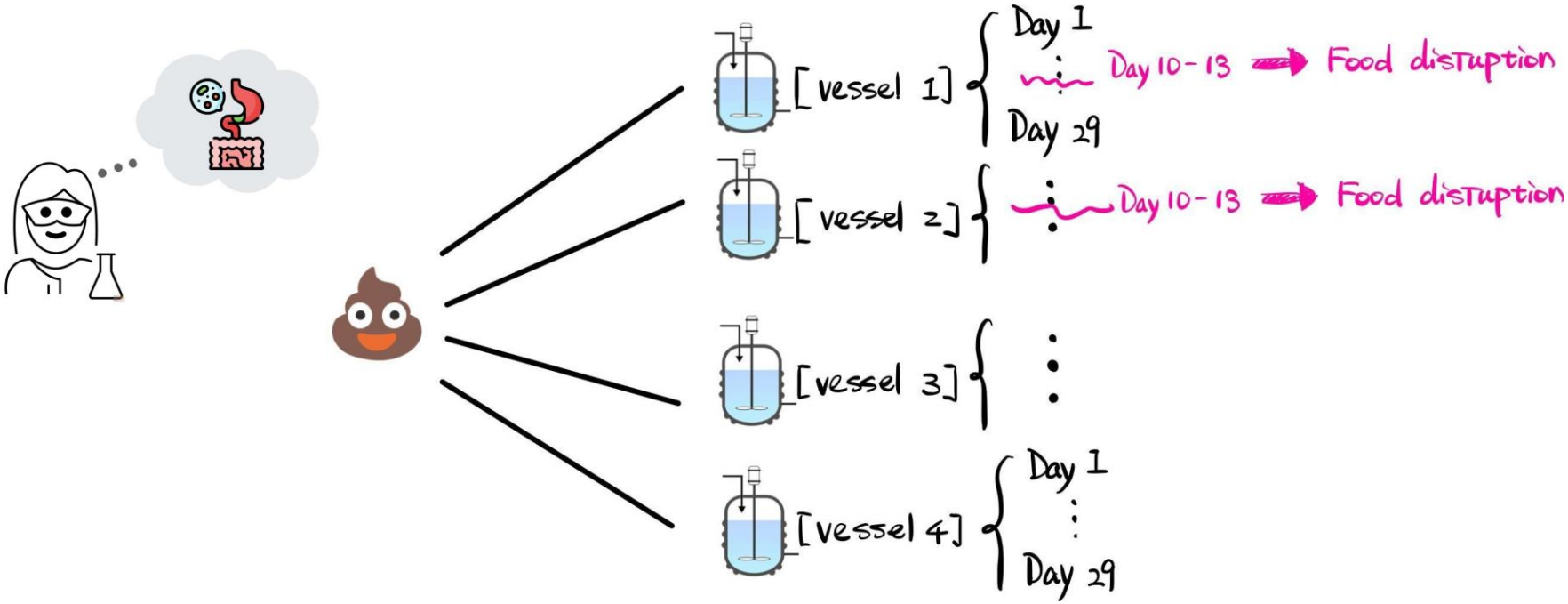
Motivated Example 2: Artificial Gut Data



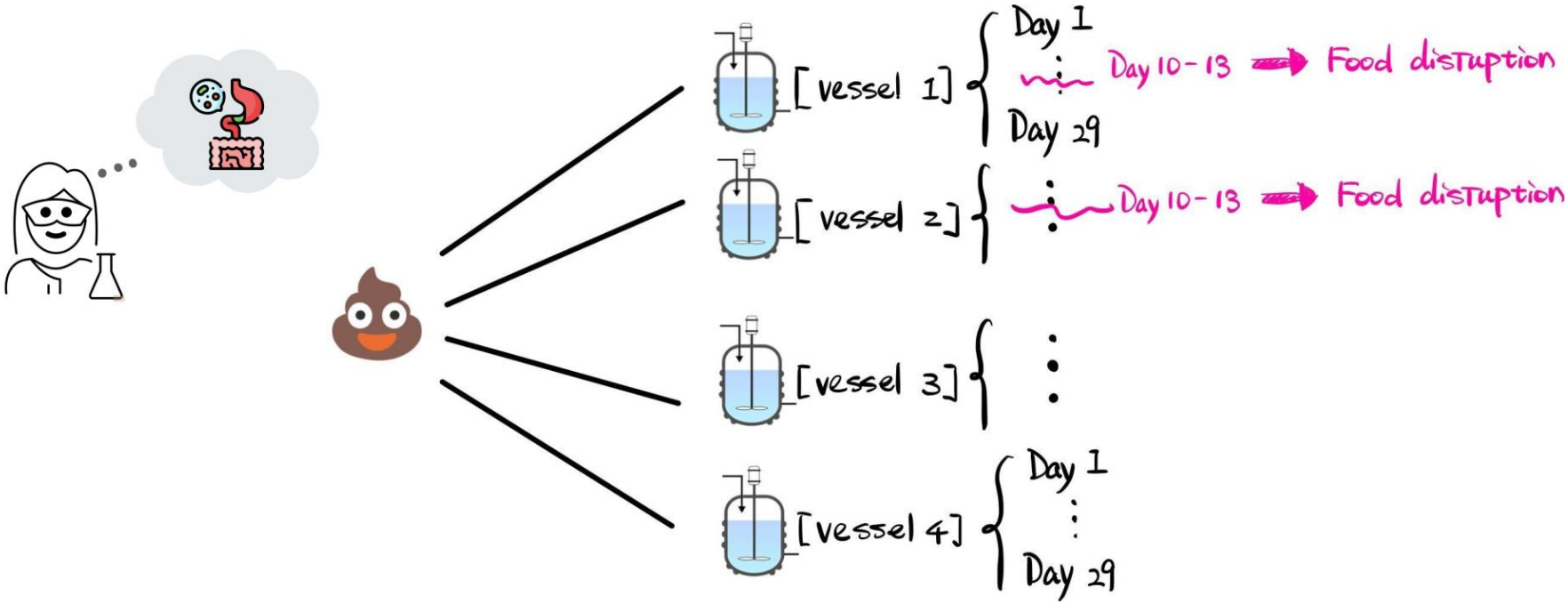
Motivated Example 2: Artificial Gut Data



Motivated Example 2: Artificial Gut Data



Motivated Example 2: Artificial Gut Data



The data is published in Silverman et al 2018.

Motivated Example 2: Artificial Gut Data

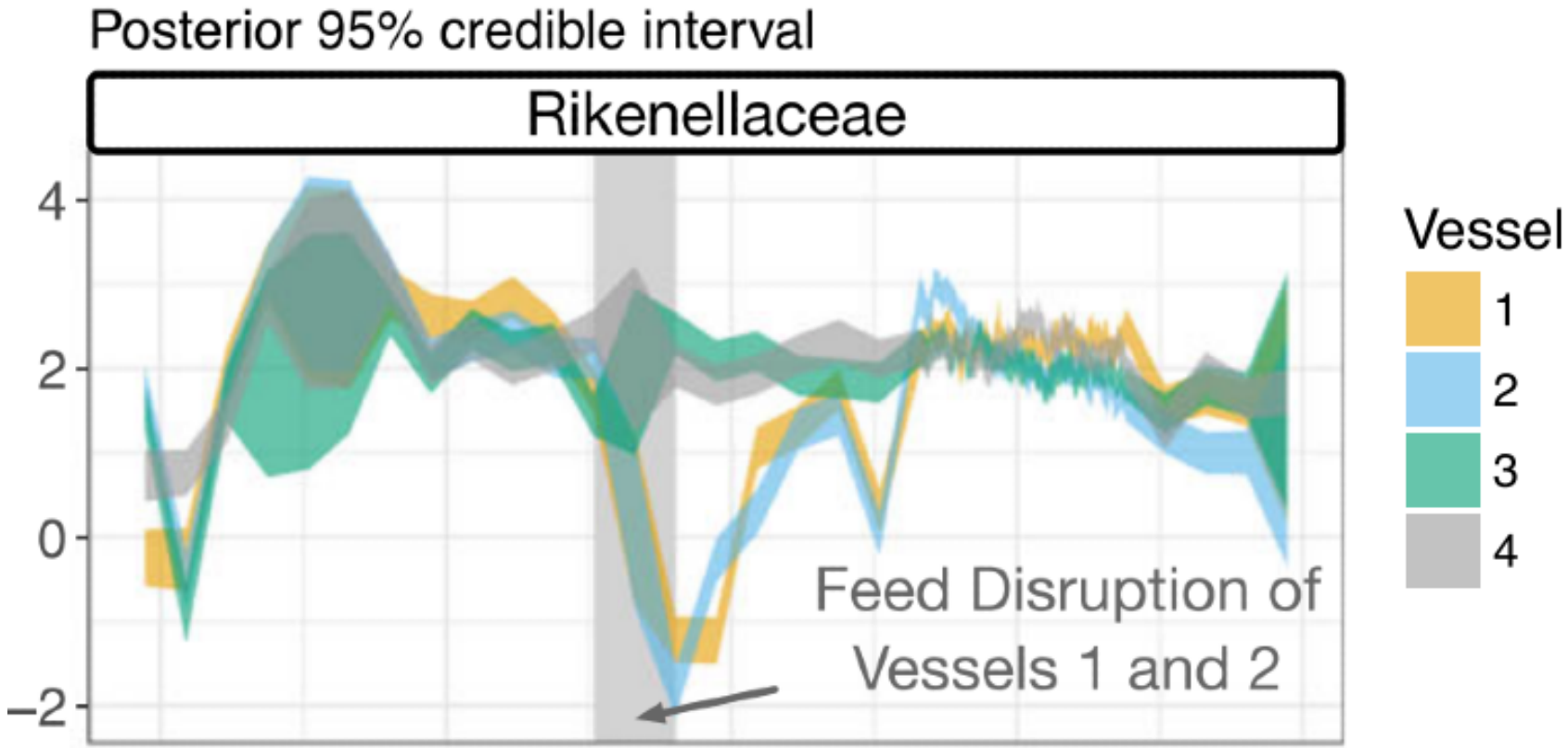


Figure 1: Silverman et al. 2018

Motivated Example 2: Artificial Gut Data

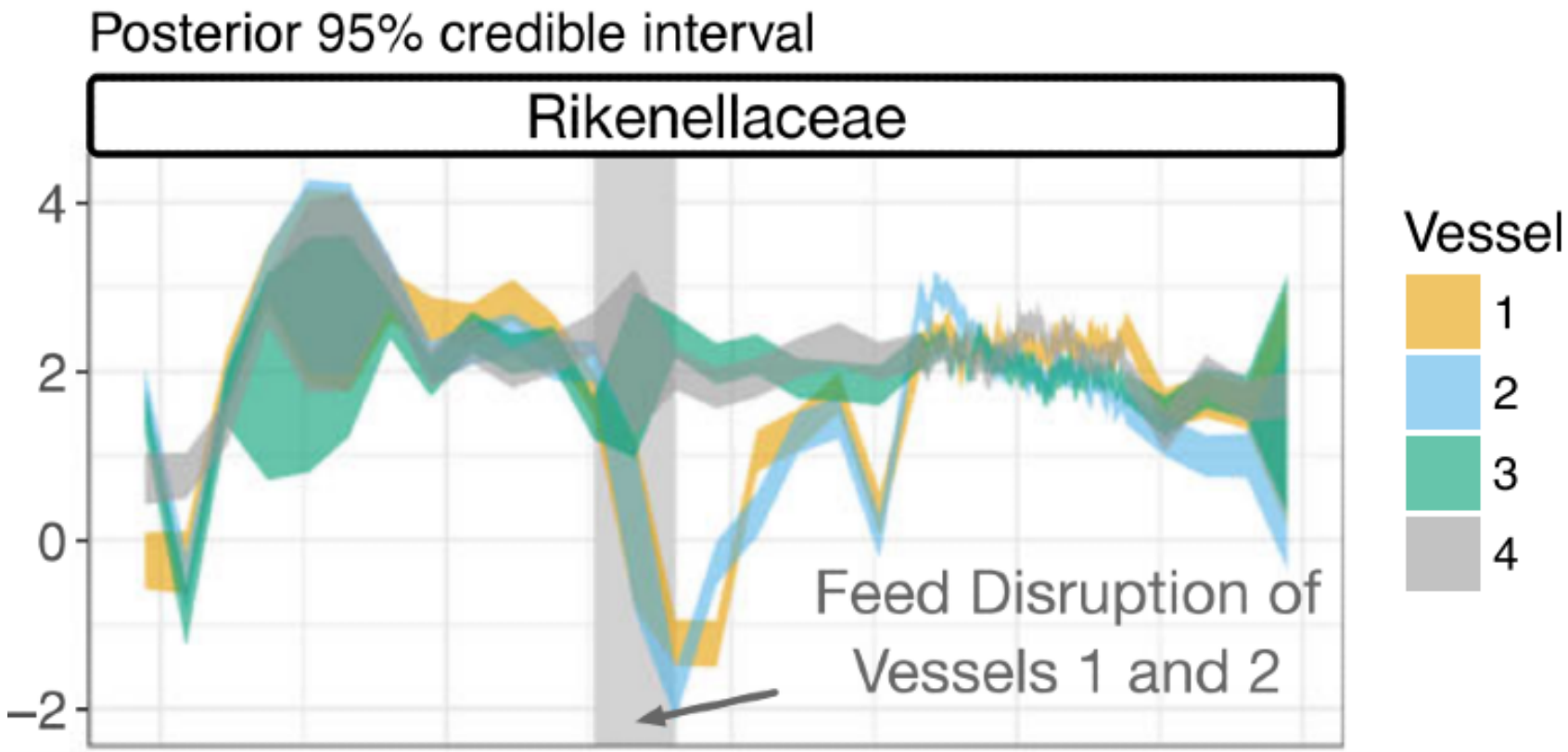


Figure 1: Silverman et al. 2018

$$Y = f_{\text{base}}(t) + f_{\text{disruption}}(t)$$

Motivated Example 2: Artificial Gut Data

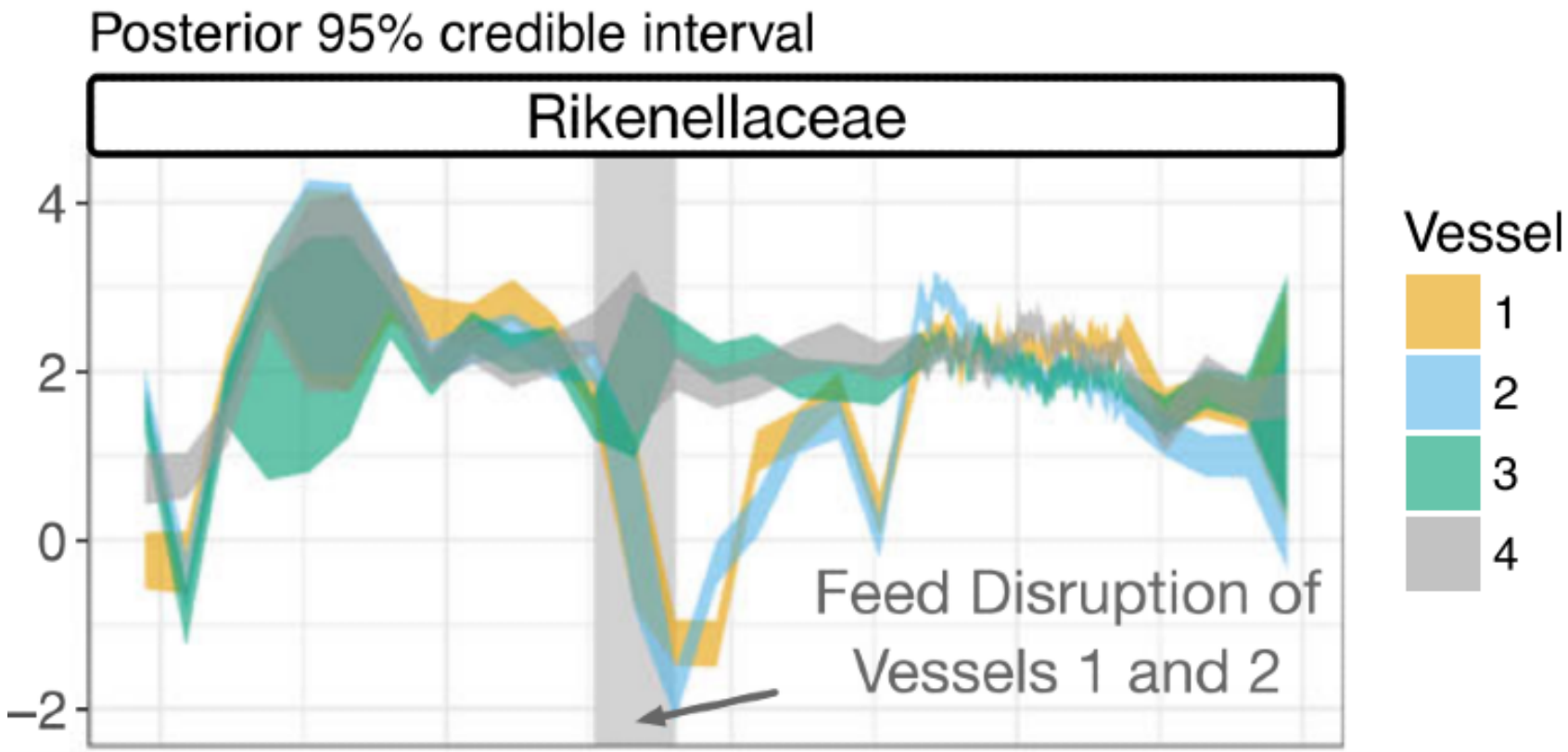
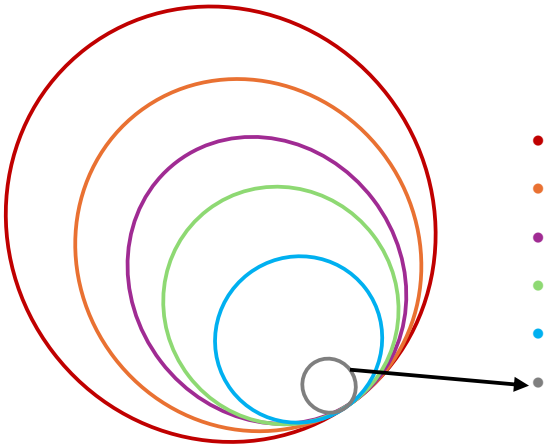


Figure 1: Silverman et al. 2018

$$Y = f_{\text{base}}(t) + f_{\text{disruption}}(t)$$

However, analyzing microbiome data is a challenge.

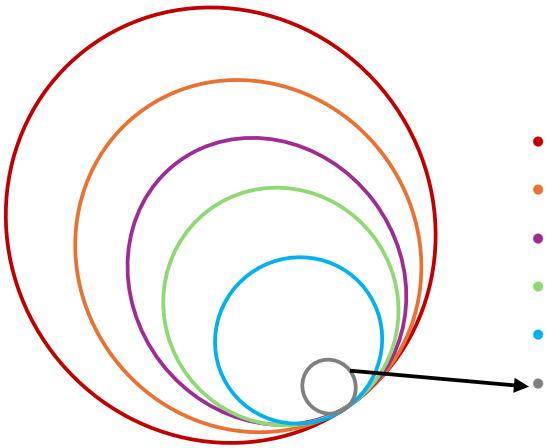
Problem: data generating process



- Total bacterial population
- Sampled portion
- DNA/RNA sample
- Amplified product
- Library for sequencing
- Reads from the sequencer

	Subject 1	Subject 2
Microbe 1	50	60
Microbe 2	80	75
Microbe 3	45	55
...	...	...
Sequence Depth	500	520

Problems: data generating process



- Total bacterial population
- Sampled portion
- DNA/RNA sample
- Amplified product
- Library for sequencing
- Reads from the sequencer

	Subject 1	Subject 2
Microbe 1	50	60
Microbe 2	80	75
Microbe 3	45	55
...	...	...
Sequence Depth	500	520

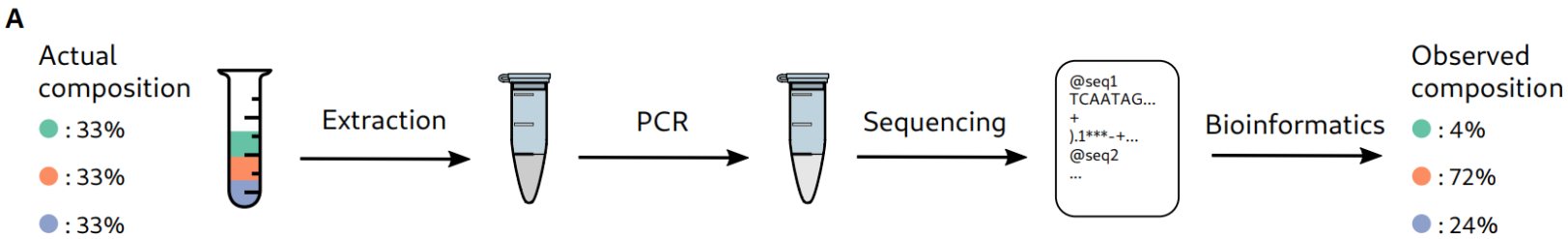
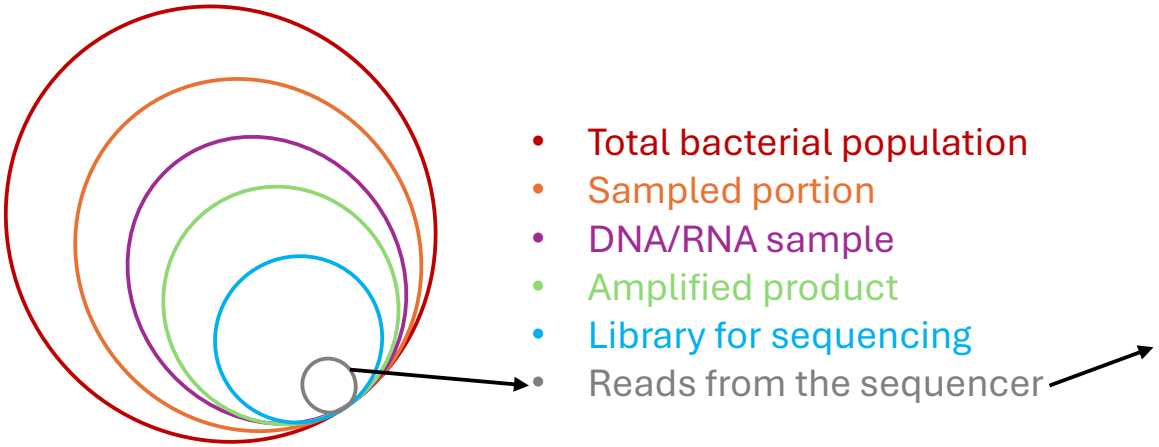


Figure 2: McLaren et al. 2019

Problem: data generating process



- Total bacterial population
- Sampled portion
- DNA/RNA sample
- Amplified product
- Library for sequencing
- Reads from the sequencer

	Subject 1	Subject 2
Microbe 1	50	60
Microbe 2	80	75
Microbe 3	45	55
...	...	...
Sequence Depth	500	520

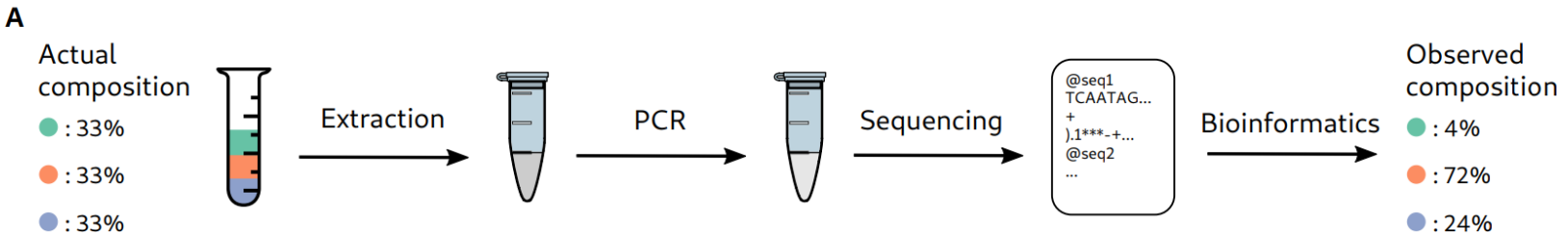


Figure 2: McLaren et al. 2019

Two sources of uncertainty:

- Uncertainty due to counting
- Uncertainty due to the measurement process

Multinomial Logistic Normal (MLN)

$$Y_{.n} \sim \text{Multinomial}(\pi_{.n})$$
$$\pi_{.n} = \text{Log-ratio}^{-1}(\eta_{.n})$$
$$\eta_{.n} \sim \text{MN}(F, \Sigma)$$

Multinomial Logistic Normal (MLN)

$$Y_{.n} \sim \text{Multinomial}(\pi_{.n}) \quad \leftarrow \text{counting uncertainty}$$

$$\pi_{.n} = \text{Log-ratio}^{-1}(\eta_{.n})$$

$$\eta_{.n} \sim \text{MN}(F, \Sigma)$$

Multinomial Logistic Normal (MLN)

$$Y_{.n} \sim \text{Multinomial}(\pi_{.n})$$

$$\pi_{.n} = \text{Log-ratio}^{-1}(\eta_{.n})$$

$$\eta_{.n} \sim \text{MN}(F, \Sigma) \quad \leftarrow \text{measurement process}$$

Remember, our goal is:

$$Y = BX + \sum_{k=1}^{K} f^{(k)}(Z^{(k)})$$

Multinomial Logistic Normal Additive Gaussian Process Regression (MultiAddGPs)

$$
\begin{aligned}
\mathrm{Y}_{.n} &\sim \text{Multinomial}\,(\Pi_{.n}) \\
\Pi_{.n} &= \phi^{-1}(\mathrm{H}_{.n}) \\
\mathrm{H}_{.n} &\sim \mathrm{N}\,(\mathrm{F}_{.n}\,, \Sigma) \\
\mathrm{F} &= \mathrm{BX} + \sum_{k=1}^{K} \mathrm{f}^{(k)}(\mathrm{Z}^{(k)}) \\
\mathrm{B} &\sim \mathrm{N}(\Theta^{(0)}, \Sigma\,, \Gamma^{(0)}) \\
\mathrm{f}^{(k)} &\sim \mathrm{GP}(\,\Theta^{(k)}, \Sigma\,, \Gamma^{(k)}) \\
\Sigma &\sim \text{InvWishart}(\Xi\,, \nu\,)
\end{aligned}
$$

MultiAddGPs

Multinomial Logistic Normal Additive Gaussian Process Regression (MultiAddGPs)

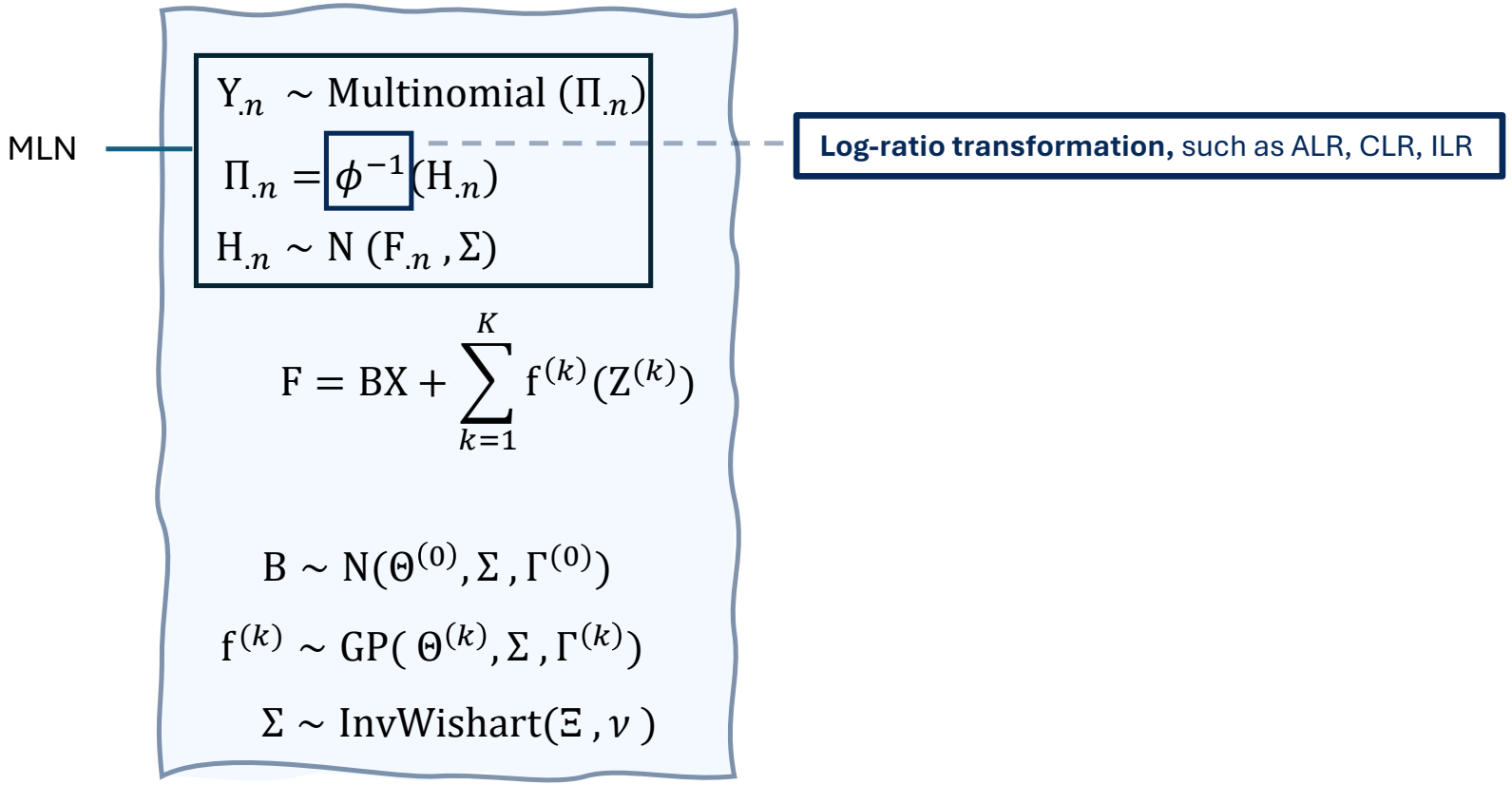
MLN

$$Y_{.n} \sim \text{Multinomial}(\Pi_{.n})$$
$$\Pi_{.n} = \phi^{-1}(H_{.n})$$
$$H_{.n} \sim N(F_{.n}, \Sigma)$$

$$F = BX + \sum_{k=1}^{K} f^{(k)}(Z^{(k)})$$

$$B \sim N(\Theta^{(0)}, \Sigma, \Gamma^{(0)})$$
$$f^{(k)} \sim GP(\Theta^{(k)}, \Sigma, \Gamma^{(k)})$$
$$\Sigma \sim \text{InvWishart}(\Xi, \nu)$$

Multinomial Logistic Normal Additive Gaussian Process Regression (MultiAddGPs)



Multinomial Logistic Normal Additive Gaussian Process Regression (MultiAddGPs)

MLN

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

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

$$H_{.n} \sim N(F_{.n}, \Sigma)$$

Log-ratio transformation, such as ALR, CLR, ILR

Latent Biological Composition Variation

$$F = BX + \sum_{k=1}^{K} f^{(k)}(Z^{(k)})$$

linear effect

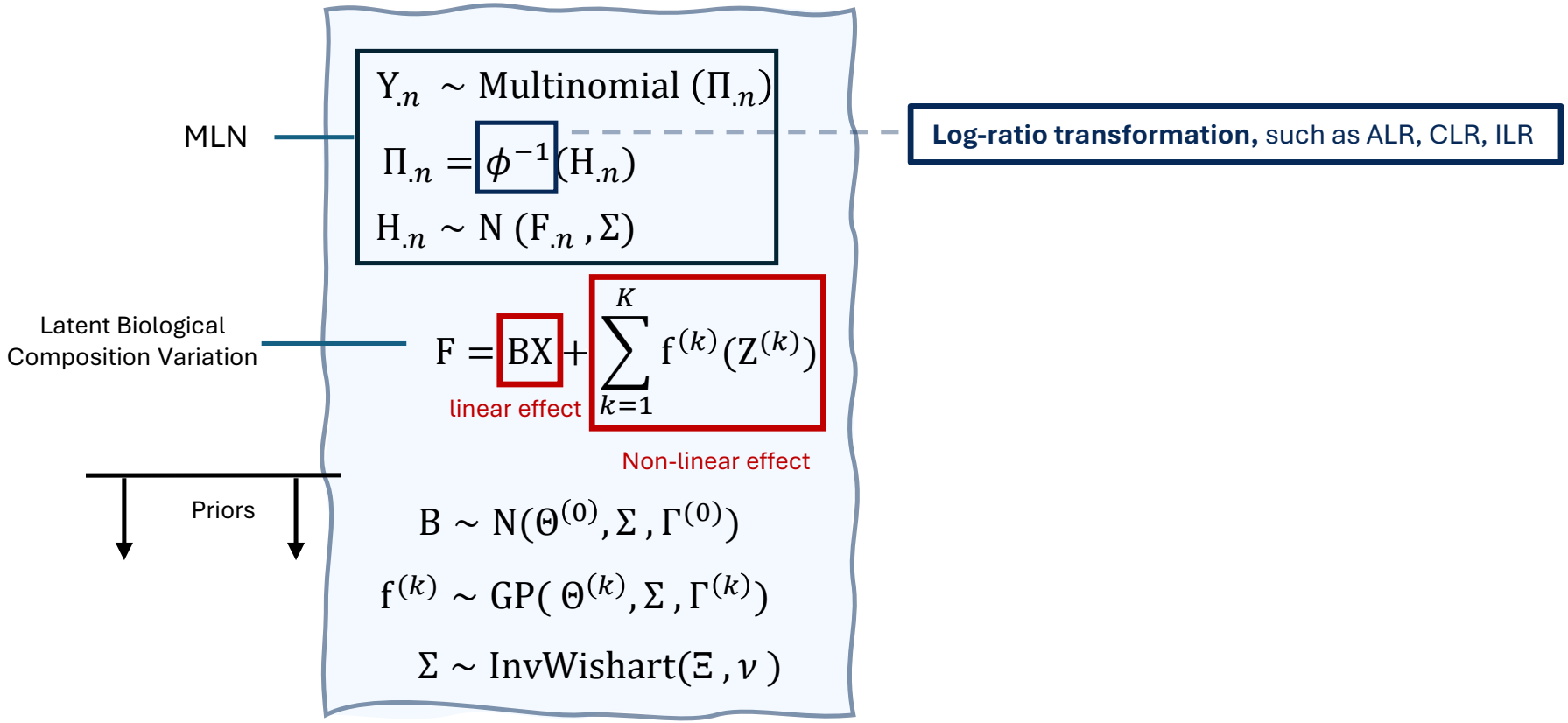
Non-linear effect

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

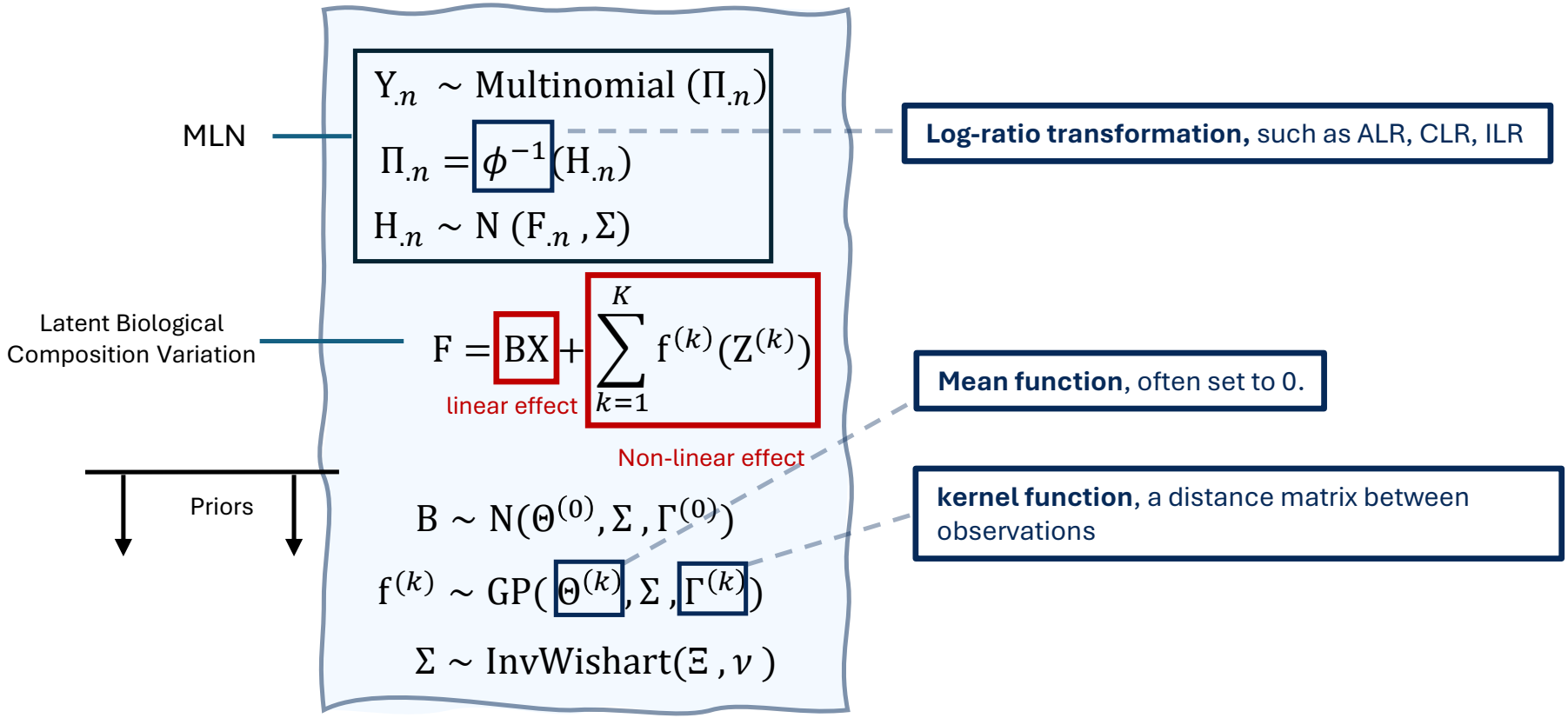
$$f^{(k)} \sim GP(\Theta^{(k)}, \Sigma, \Gamma^{(k)})$$

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

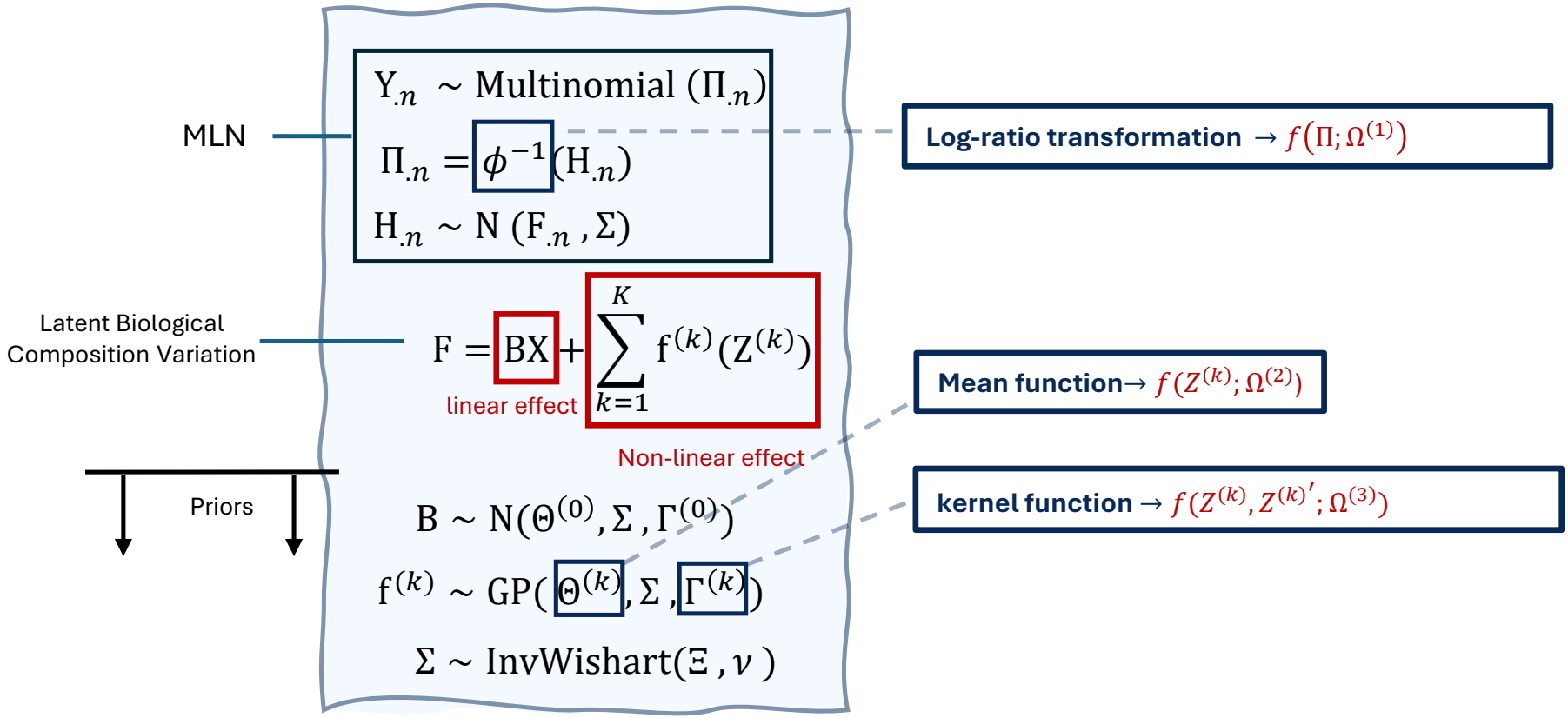
Multinomial Logistic Normal Additive Gaussian Process Regression (MultiAddGPs)



Multinomial Logistic Normal Additive Gaussian Process Regression (MultiAddGPs)



Multinomial Logistic Normal Additive Gaussian Process Regression (MultiAddGPs)



Recall:Artificial Gut Example

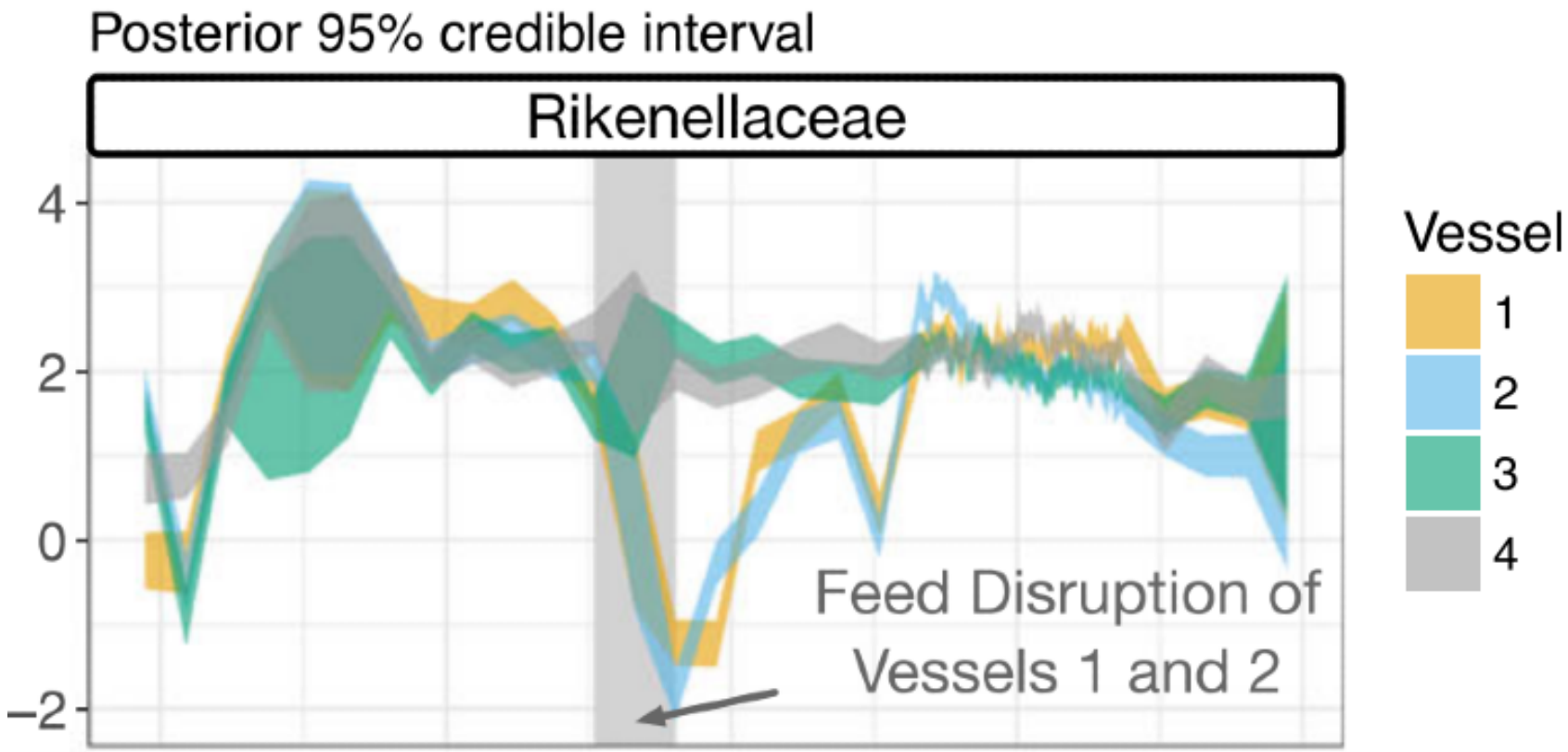


Figure 1: Silverman et al. 2018

$$Y = f_{\text{base}}(t) + f_{\text{disruption}}(t)$$

Recall: Artificial Gut Example

$$Y_{\cdot n} \sim \text{MLN}(F_{\cdot n}, \Sigma)$$

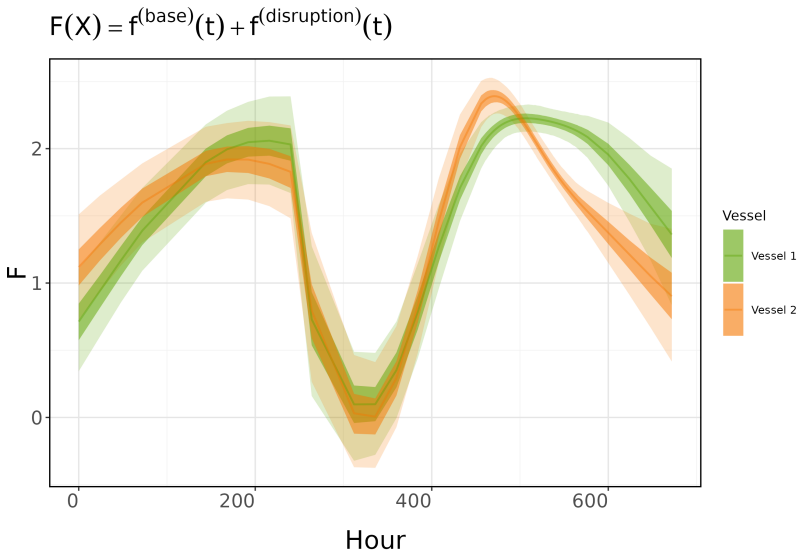
$$F = \sum_{v=1}^{4} f^{(\text{base},v)}(t_n) + I[v \in \{1,2\}] f^{(\text{disrupt},v)}(t)$$

$$f^{(\text{base},v)} \sim \text{GP}(\Theta^{(\text{base},v)}, \Sigma, \Gamma^{(v)} \odot \Gamma^{(\text{base})}), v \in \{1,2,3,4\}$$

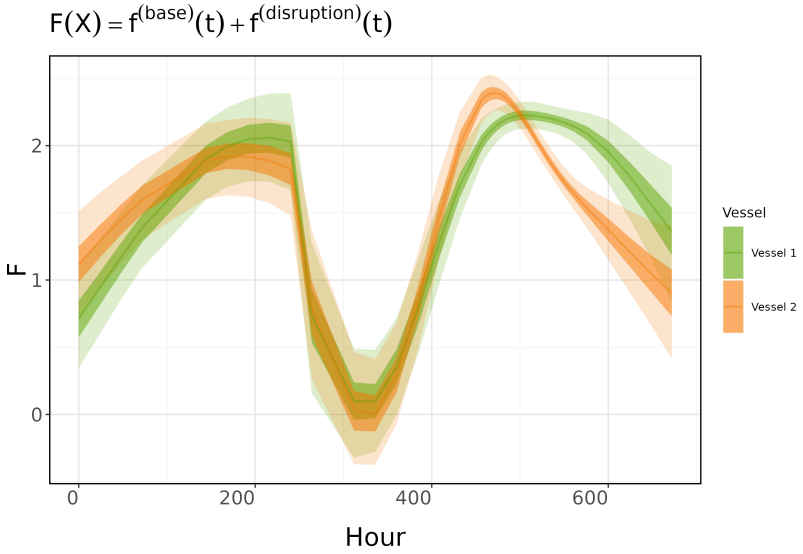
$$f^{(\text{disrupt},v)} \sim \text{GP}(\Theta^{(\text{disrupt},v)}, \Sigma, \Gamma^{(v)} \odot \Gamma^{(\text{disrupt})}), v \in \{1,2\}$$

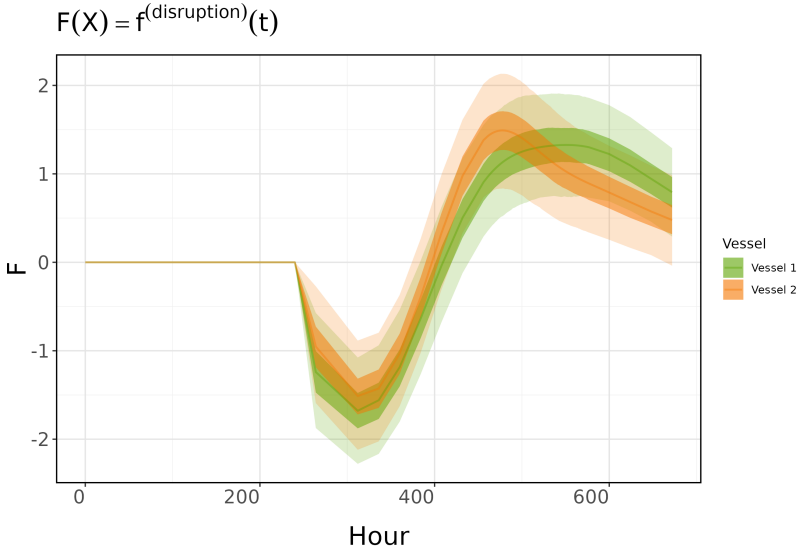
$$\Sigma \sim \text{InvWishart}(\Xi, \zeta)$$

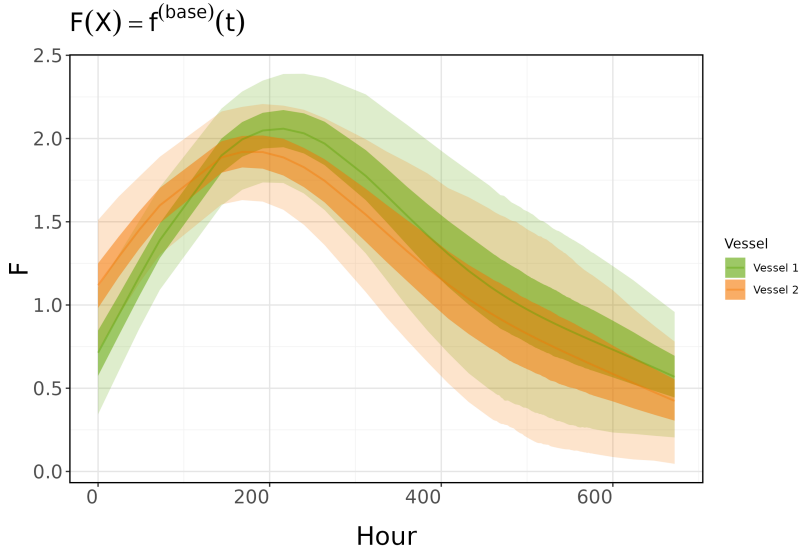
Result: Artificial Gut Example



Result: Artificial Gut Example







R package: Fido

fido 1.1.0 Reference Articles Changelog

R-CMD-check passing license GPL (>= 2)

fido (formerly stray)

Multinomial Logistic-Normal Models (really fast)
*its a little **tar**-ball of joy*

Citation

Silverman, JD, Roche, K, Holmes, ZC, David, LA, and Mukherjee, S. *Journal of Machine Learning Research*. 23(7), 2022:1–42.

License

All source code freely availale under GPL-3 License.

A walkthrough example is now available at:

`https://tinghua-chen.github.io/blog/MultiAddGPs/`

Something that I did not cover here:

- The posterior sampling method we developed
- Hyperparameters selection via maximum marginal likelihood

Acknowledgment

- Dr. Justin Silverman
- Dr. Michelle Nixion
- Kyle McGovern
- Andrew Sugarman
- Manan Saxena
- Won Gu
- Maxwell Konnaris
- Paul Yu
- Ziang Shi

The project is funded by NIH R01 GM148972-01