

# Bayesian Inference and the Thermodynamic Formalism

SIAM Conference on Applications of Dynamical Systems  
(DS19)

Sayan Mukherjee

Duke University  
<https://sayanmuk.github.io/>

Joint work with:

— K. McGoff (UNC Ch) | A. Nobel (UNC CH)

# Bayesian Inference and the Thermodynamic Formalism

SIAM Conference on Applications of Dynamical Systems  
(DS19)

Sayan Mukherjee

Duke University  
<https://sayanmuk.github.io/>

Joint work with:

— K. McGoff (UNC Ch) | A. Nobel (UNC CH)

# Bayesian Inference and the Thermodynamic Formalism

SIAM Conference on Applications of Dynamical Systems  
(DS19)

Sayan Mukherjee

Duke University  
<https://sayanmuk.github.io/>

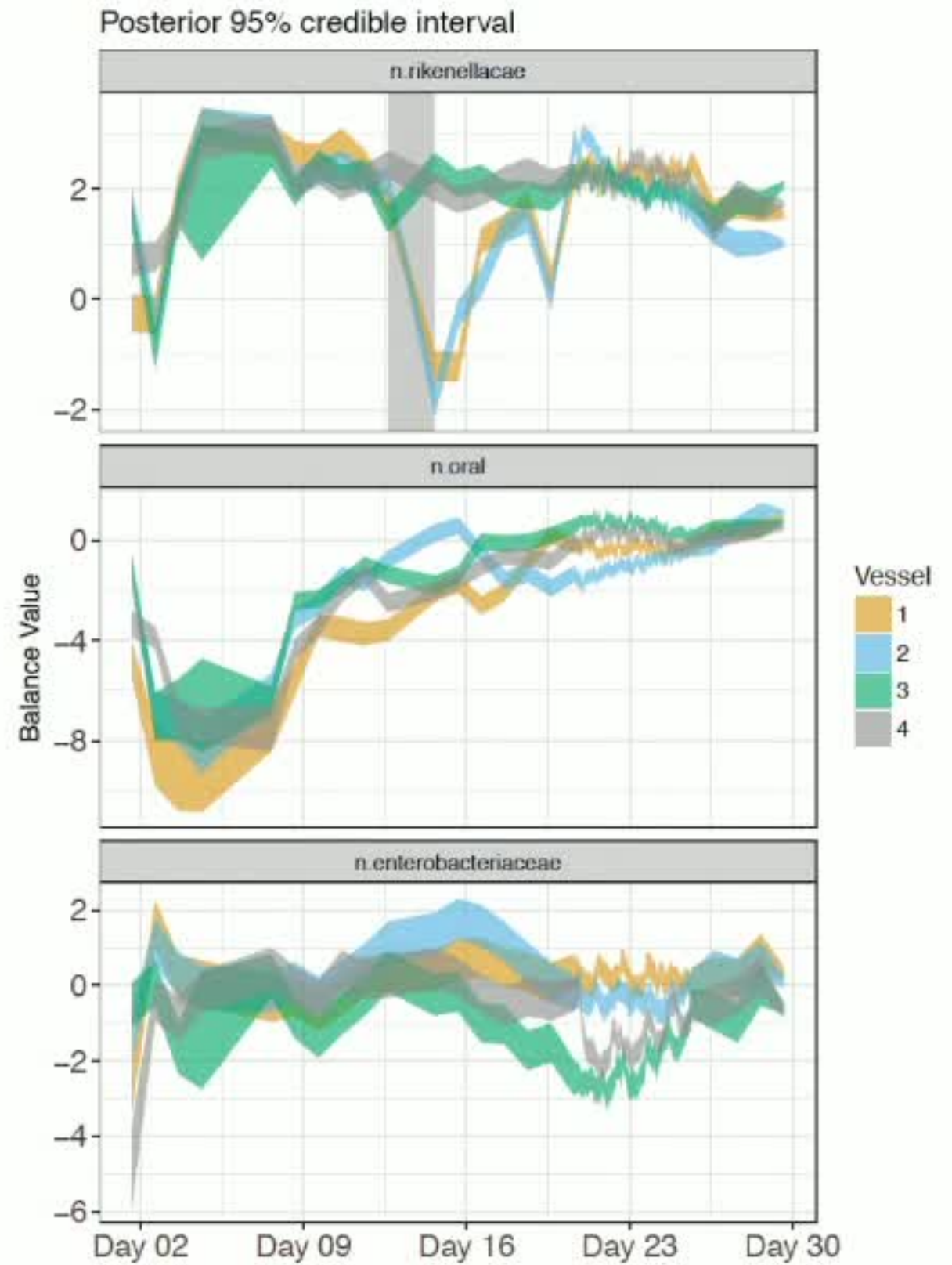
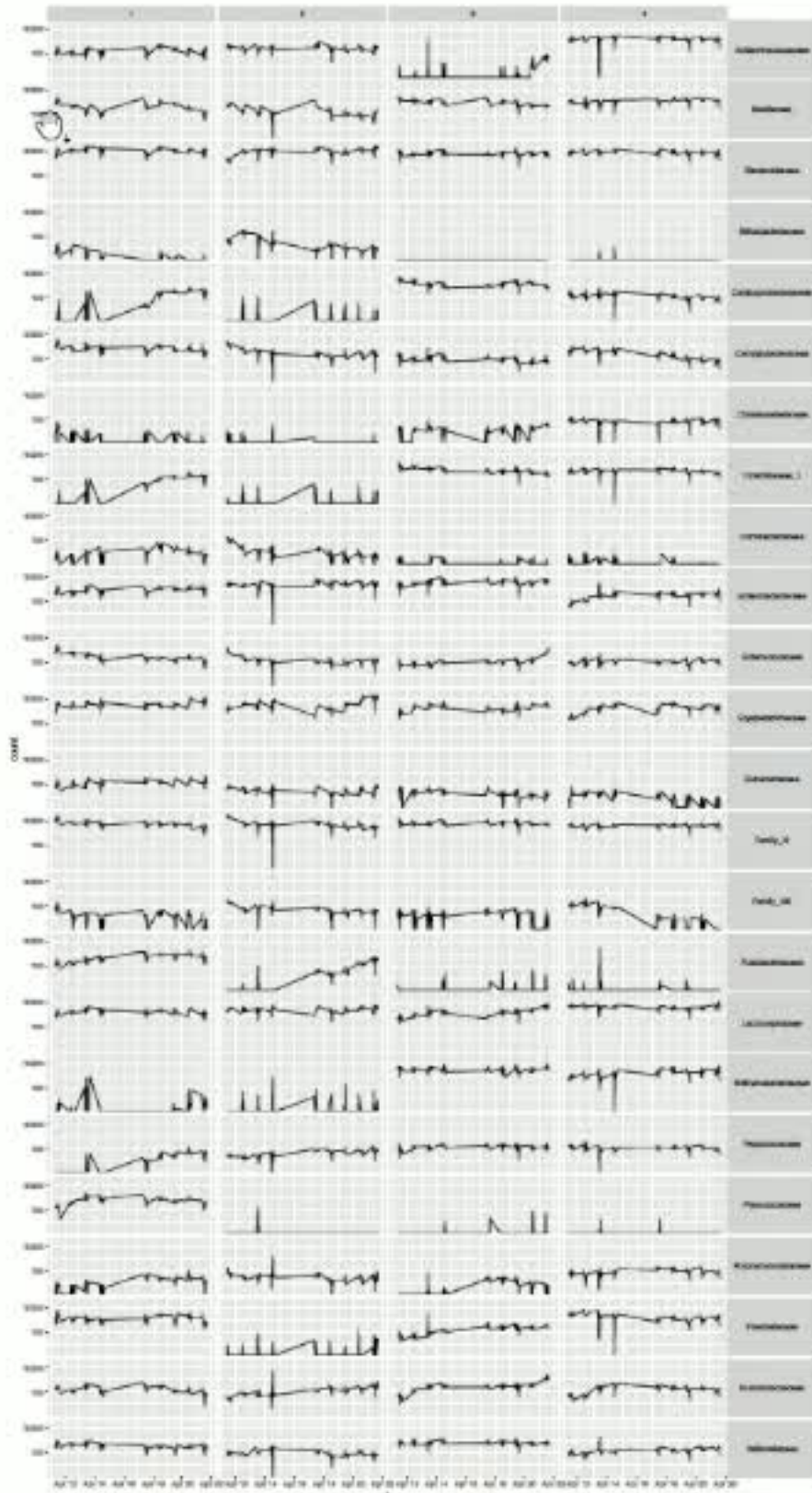
Joint work with:

— K. McGoff (UNC Ch) | A. Nobel (UNC CH)

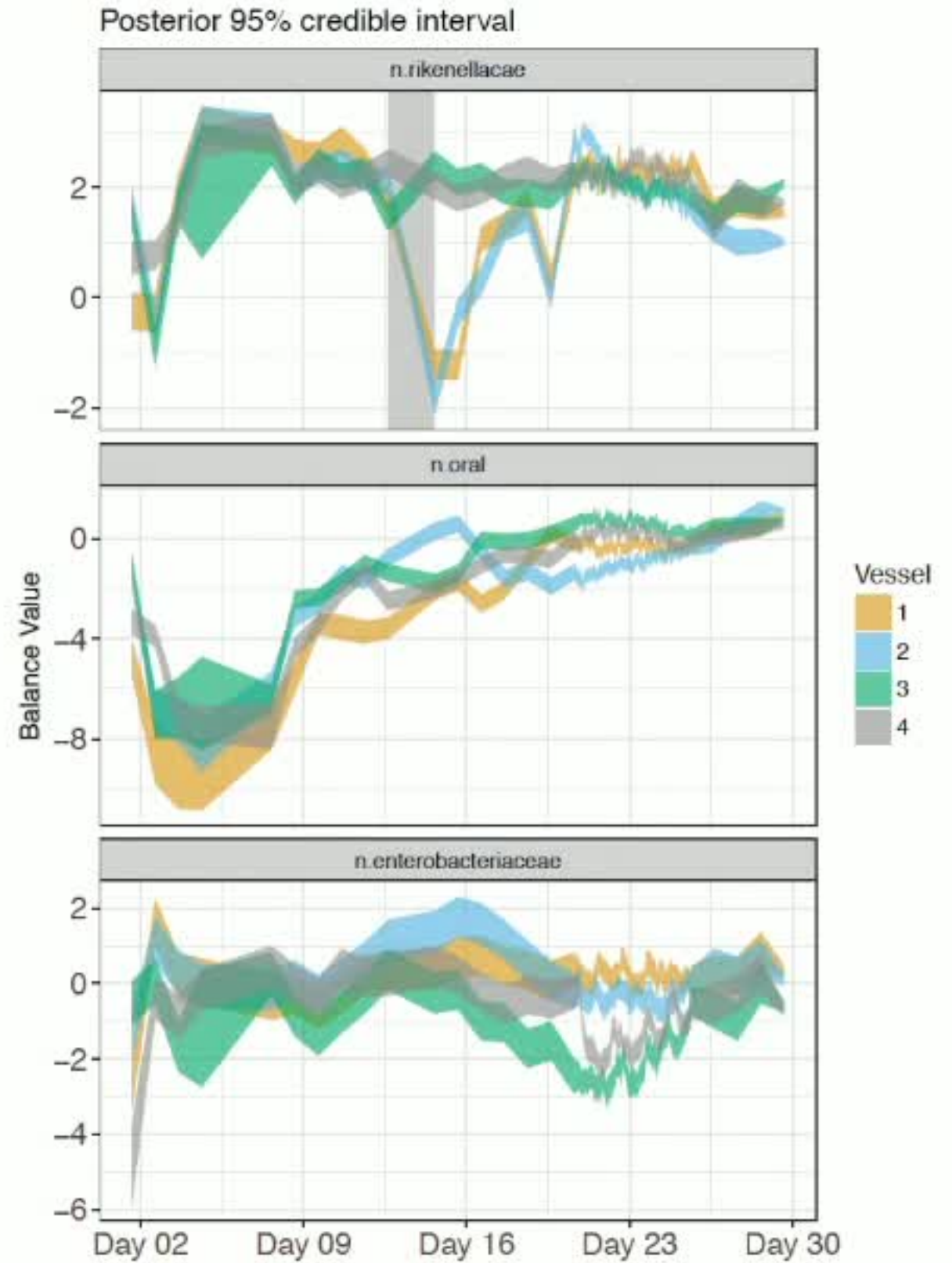
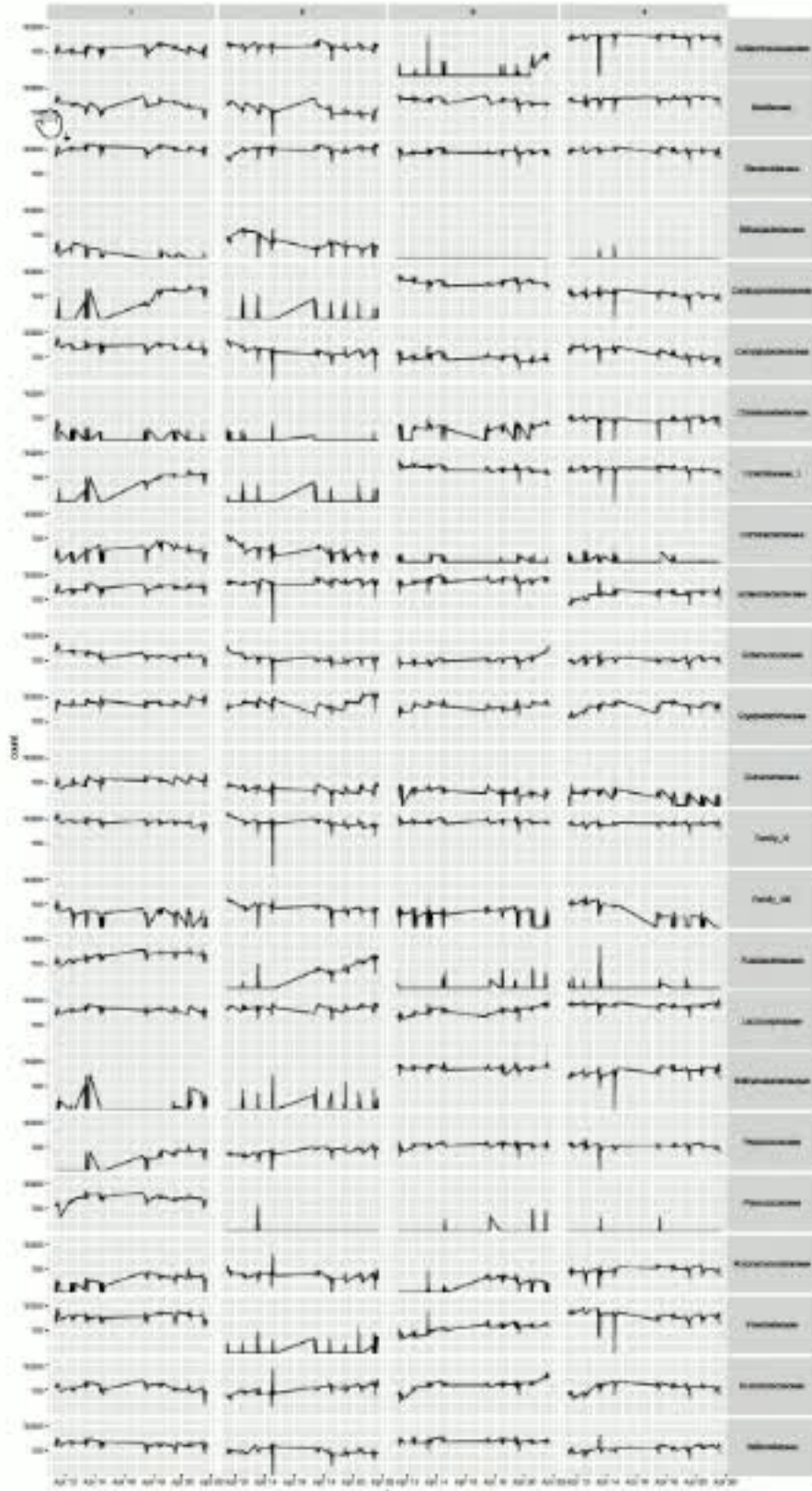




# Microbial ecology

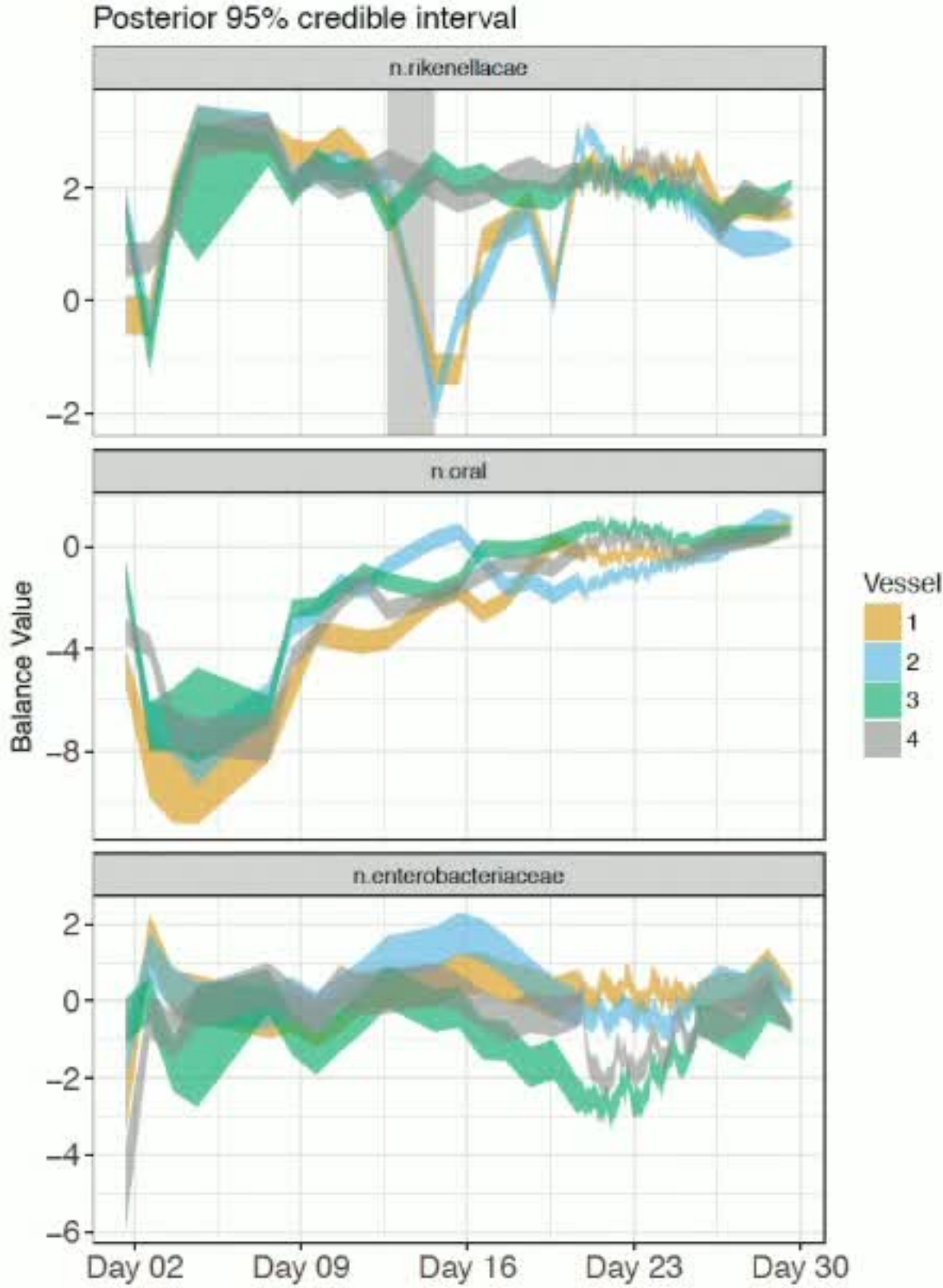
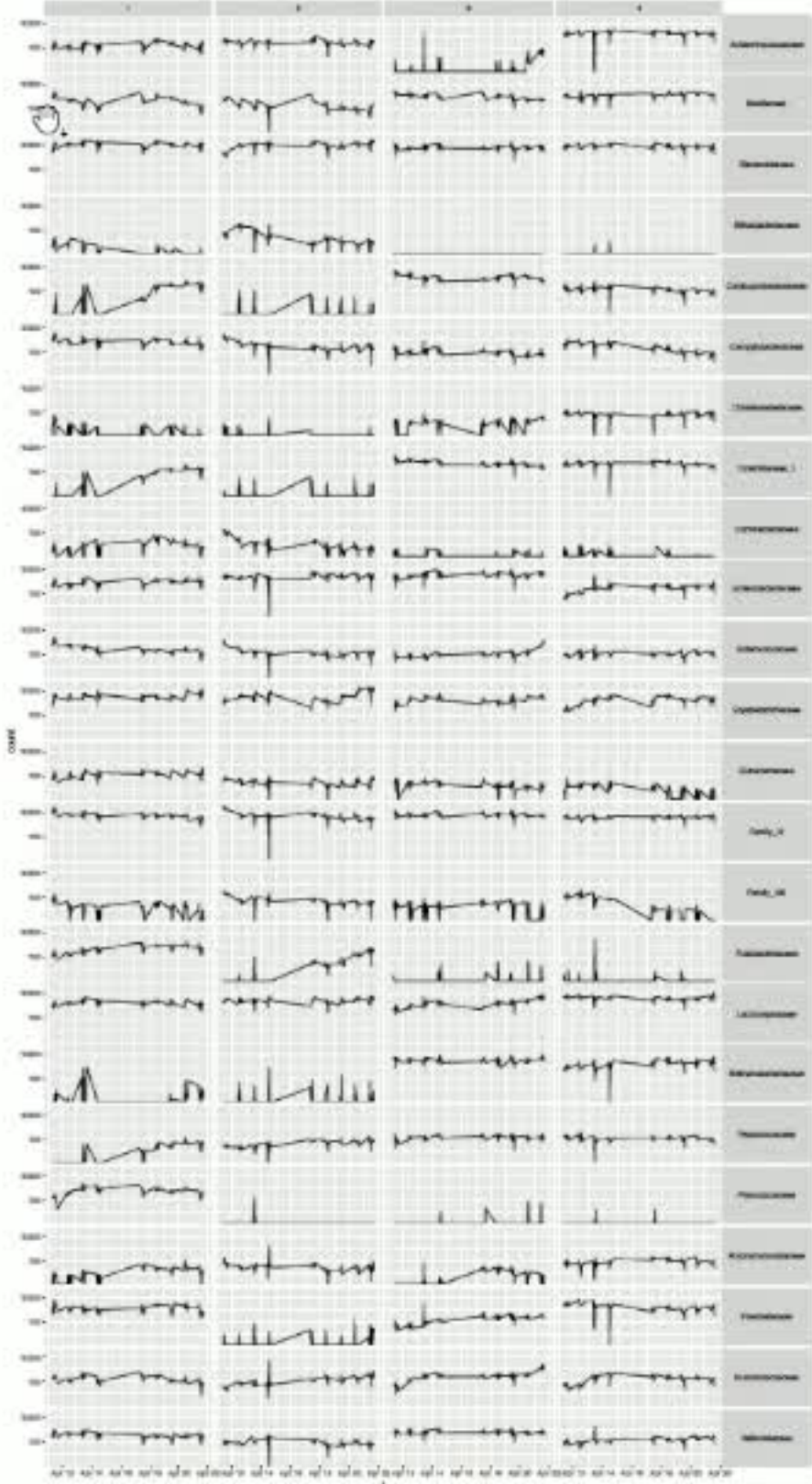


# Microbial ecology

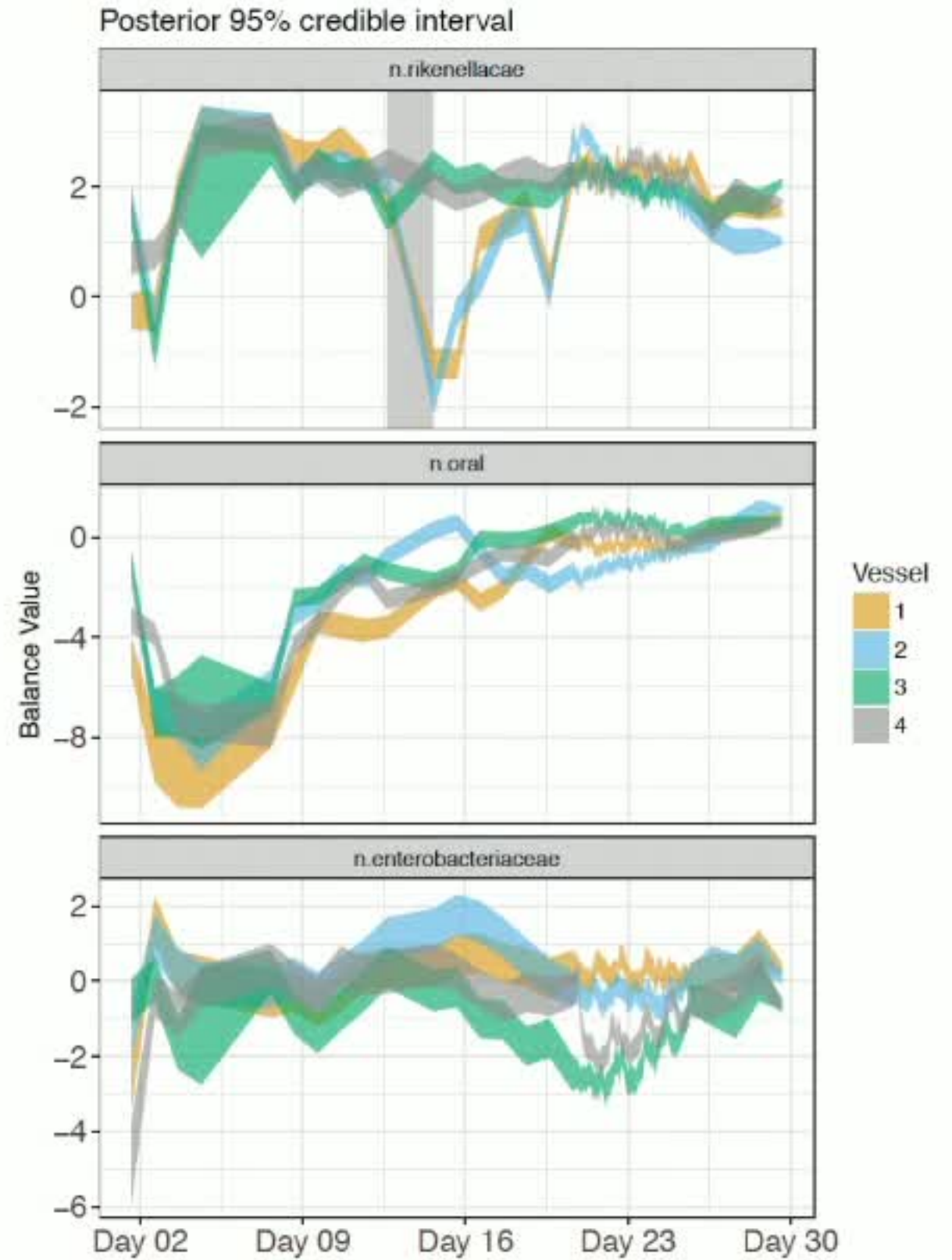
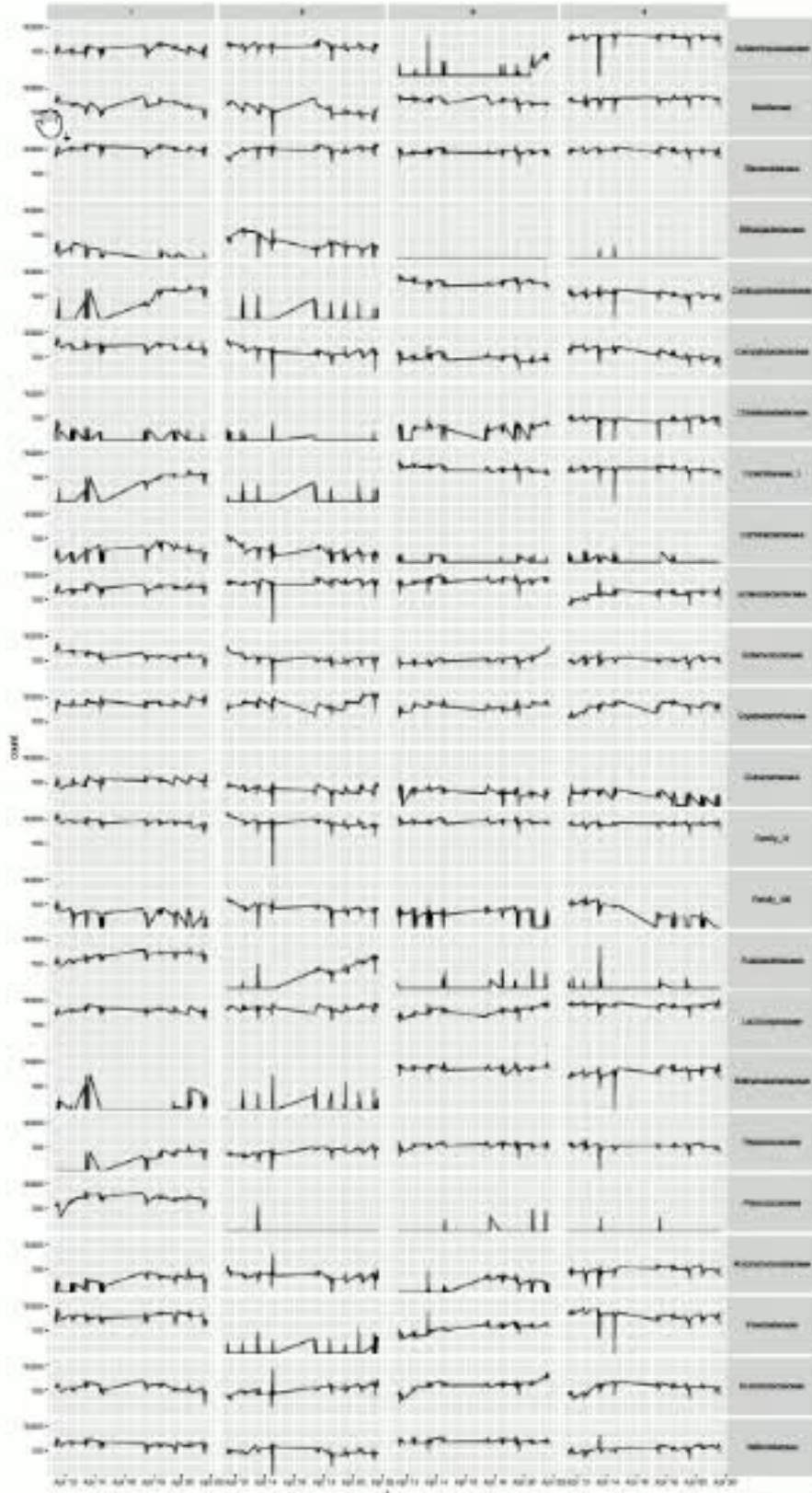




# Microbial ecology

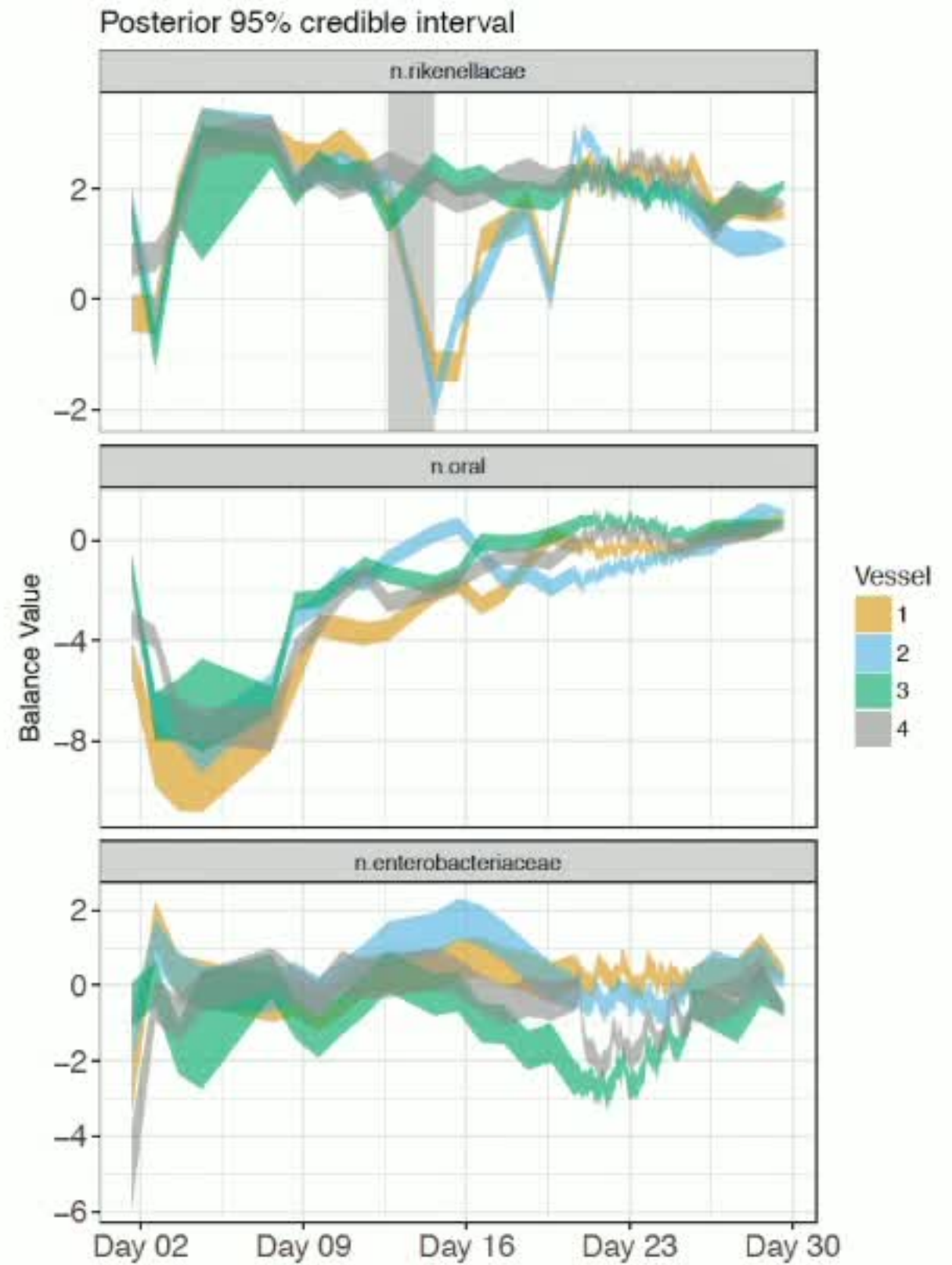
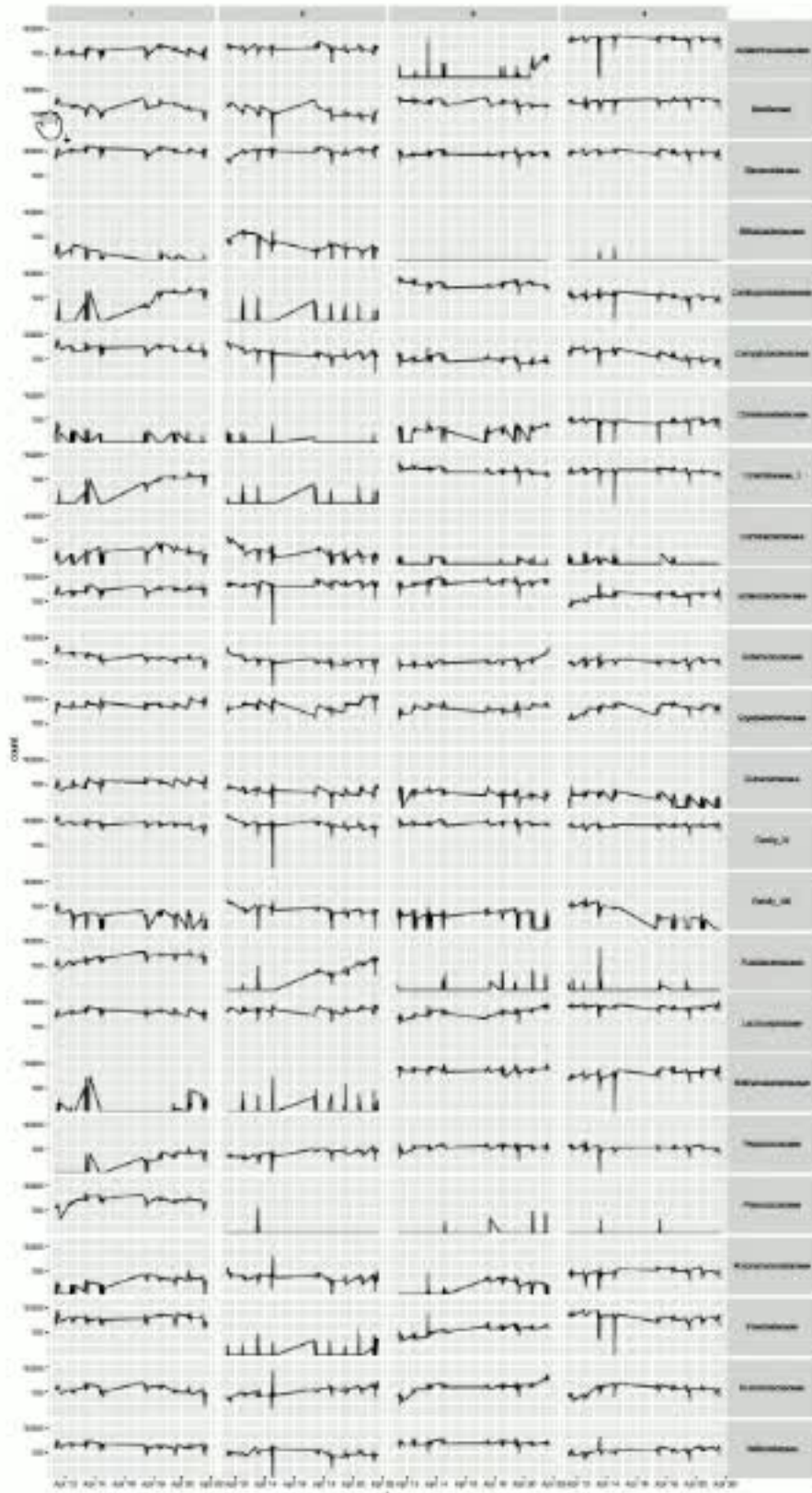


# Microbial ecology





# Microbial ecology



# Posterior consistency

Q.

Does  $\lim_{n \rightarrow \infty} \Pi_n(\theta \mid y_1^n)$  concentrate around an open neighborhood of  $\theta^*$  ?

# Classical setting

Consider

$\mathcal{Y}$  : a complete metric space endowed with its Borel  $\sigma$ -algebra;

$\{Y_n\}_{n \geq 0}$ : observations as a  $\mathcal{Y}$ -valued process;

$(\Theta, \{p_\theta : \theta \in \Theta\})$ : a parameter space and a collection of Borel probability densities on  $\mathcal{Y}$  (with respect to a common measure);

$\pi(\theta)$ : the prior, a Borel probability distribution on  $\Theta$ .





# Classical setting

• Consider

$\mathcal{Y}$  : a complete metric space endowed with its Borel  $\sigma$ -algebra;

$\{Y_n\}_{n \geq 0}$ : observations as a  $\mathcal{Y}$ -valued process;

$(\Theta, \{p_\theta : \theta \in \Theta\})$ : a parameter space and a collection of Borel probability densities on  $\mathcal{Y}$  (with respect to a common measure);

$\pi(\theta)$ : the prior, a Borel probability distribution on  $\Theta$ .

# Posterior consistency

•

We say that  $(\theta_0, \pi)$  is consistent if for all open neighborhoods  $U$  of  $\theta_0$ ,

$$\Pi_n(\Theta \setminus U \mid Y_0^{n-1}) \rightarrow 0, \quad P_{\theta_0}^\infty - a.s.$$

## Theorem (Doob, 1949)

For  $\pi$ -almost every  $\theta$  in  $\Theta$ , the pair  $(\theta, \pi)$  is consistent.

What about for *every*  $\theta$  in  $\Theta$ ?



# Schwartz conditions

## • Theorem (Schwartz, 1965)

Let  $\theta_0 \in \Theta$ . Suppose that

1. for each neighborhood  $U$  of  $\theta_0$ , there exist constants  $\beta > 0$  and  $C > 0$  and measurable functions  $\varphi_n : \mathcal{Y}^n \rightarrow [0, 1]$  such that

a)  $\mathbb{E}_{\theta_0}[\varphi_n(Y_0^{n-1})] \leq Ce^{-\beta n}$ , and

b)  $\sup_{\theta \notin U} \mathbb{E}_{\theta}[1 - \varphi_n(Y_0^{n-1})] \leq Ce^{-\beta n}$ .

2. for each  $\epsilon > 0$ ,

$$\pi \left( \theta : \mathbb{E}_{\theta_0}[-\log(p_{\theta}/p_{\theta_0})] < \epsilon \right) > 0.$$

Then  $(\theta_0, \pi)$  is consistent.

# More recent work



1990's: Inconsistency results for nonparametric models ( $\Theta$  is infinite dimensional) by Diaconis and Freedman.

2000-2010: Extensive results for nonparametric models, Ghosal and van der Vaart [2017]

2000-2019: Rates of convergence

# Dependence

- We would like to consider posterior consistency for stationary processes.

Suppose that  $\{Y_n\}_{n \geq 0}$  is stationary (not necessarily i.i.d.).

$\Theta$  parametrizes a collection of stationary stochastic processes, serving as models of  $\{Y_n\}_n$ .

Given a prior distribution  $\pi$ , we'll define a posterior distribution  $\Pi_n(\cdot | Y_0^{n-1})$ .

**Question:** What happens to  $\Pi_n(\cdot | Y_0^{n-1})$  as  $n$  tends to infinity?



# Hidden Markov Models

• Markov model:

$$x_{t+1} = f(x_t; \theta), \quad \text{state process}$$

Hidden Markov model:

$$\begin{aligned} x_{t+1} &= f(x_t; \theta_1) && \text{hidden state process} \\ y_{t+1} &= g(x_{t+1}; \theta_2) && \text{observation process.} \end{aligned}$$

# General questions



Given access to the observations  $Y_0^t$ , we might want to ask

what is the “true state” of the bioreactor at time  $t$ ? (filtering)

what are we likely to observe at time  $t + 1$ ? (prediction)

what are the rules governing the evolution of the system?  
(model selection / parameter estimation)

We'll focus on the last type of question.

# Stochastic versus deterministic systems



Should the process  $(X_t)_t$  be stochastic or deterministic?

- ▶ If the conditional distribution of  $X_{t+1}$  given  $X_t$  has positive variance, then we'll say the process  $(X_t)_t$  is stochastic.
- ▶ Otherwise, we'll say the process  $(X_t)_t$  is deterministic.

In ecology both types of systems are commonly used.



# Setting for deterministic dynamics

- Suppose that for each  $\theta$  in  $\Theta$  (parameter space), we have  $(X, \mathcal{X}, T_\theta, \mu_\theta)$ , where
- ▶  $X$  is a complete separable metric space with Borel  $\sigma$ -algebra  $\mathcal{X}$
  - ▶  $T_\theta : X \rightarrow X$  is a measurable map,
  - ▶  $\mu_\theta$  is a probability measure on  $(X, \mathcal{X})$  is  $T_\theta$ -invariant if  $\mu_\theta(T_\theta^{-1}A) = \mu_\theta(A), \quad \forall A \in \mathcal{X}$
  - ▶ the measure preserving system  $(X, \mathcal{X}, T_\theta, \mu_\theta)$  is ergodic if  $T_\theta^{-1}A = A$  implies  $\mu(A) = \{0, 1\}$ .

Family of systems  $(X, \mathcal{X}, T_\theta, \mu_\theta)_{\theta \in \Theta} \equiv (T_\theta, \mu_\theta)_{\theta \in \Theta}$ .

# Dynamic linear models

$$\begin{aligned}x_{t+1} &= A_{t+1}x_t \\ y_t &= B_t x_t + v_t,\end{aligned}$$

Here:

$y_t$  is an observation in  $\mathbb{R}^p$ ;

$x_t$  is a hidden state in  $\mathbb{R}^q$ ;

$A_t$  is a  $p \times p$  state transition matrix;

$B_t$  is a  $q \times p$  observation matrix;

$v_t$  is a zero-mean vector in  $\mathbb{R}^q$ .

# Preliminaries

Observation system  $(\mathcal{Y}, T, \nu)$  with  $T : \mathcal{Y} \rightarrow \mathcal{Y}$

Tracking systems:

Compact metrizable space  $\mathcal{X} := X \times \Theta$  with map  $S : \mathcal{X} \rightarrow \mathcal{X}$ .

$$S : \Theta \times X \rightarrow X, \quad S_\theta : X \rightarrow X.$$

Loss or regret:  $\ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ . Cost of

$$\ell_n(x, y; \theta) := \ell_n(x_0^{n-1}, y_0^{n-1}) = \sum_{k=0}^{n-1} \ell(x_k, y_k),$$

$$x_0^{n-1} = (x, S_\theta x, \dots, S_\theta^{n-1} x) \text{ and } y_0^{n-1} = (y, Ty, \dots, T^{n-1} y).$$



# Gibbs posterior

- (1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.
- (2) If  $\ell_n$  is the negative log likelihood then recover standard posterior.
- (3) Robust to misspecification, robust statistics.

# Gibbs posterior

- (1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.
- (2) If  $\ell_n$  is the negative log likelihood then recover standard posterior.
- (3) Robust to misspecification, robust statistics.

# Gibbs posterior

- (1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.
- (2) If  $\ell_n$  is the negative log likelihood then recover standard posterior.
- (3) Robust to misspecification, robust statistics.
- (4) Calibration/violation of likelihood principle

$$\Pi_n(A | y) = \frac{\int_A \exp(-\psi \ell_n(x, y; \theta)) d\pi(x)}{Z_n(y)}.$$



# Gibbs measures

Given  $\mathcal{X}$ , the map  $S$ , a potential function  $f$ , and a measure  $\mu_0$

$$G_n(x; \mu_0, f) = \frac{\exp(\sum_{k=1}^n f(S^k x))}{\int_{\mathcal{X}} \exp(\sum_{k=1}^n f(S^k x)) d\mu_0}.$$

The Gibbs measure is the limit point of the sequence  $G_n(x; \mu_0, f)$  and the Gibbs measure is denoted as  $\mu_0(f)$ .

# Gibbs posterior

- (1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.
- (2) If  $\ell_n$  is the negative log likelihood then recover standard posterior.
- (3) Robust to misspecification, robust statistics.
- (4) Calibration/violation of likelihood principle

$$\Pi_n(A | y) = \frac{\int_A \exp(-\psi \ell_n(x, y; \theta)) d\pi(x)}{Z_n(y)}.$$

# Gibbs measures

Given  $\mathcal{X}$ , the map  $S$ , a potential function  $f$ , and a measure  $\mu_0$

$$G_n(x; \mu_0, f) = \frac{\exp(\sum_{k=1}^n f(S^k x))}{\int_{\mathcal{X}} \exp(\sum_{k=1}^n f(S^k x)) d\mu_0}.$$

The Gibbs measure is the limit point of the sequence  $G_n(x; \mu_0, f)$  and the Gibbs measure is denoted as  $\mu_0(f)$ .



# Gibbs posterior

- (1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.

# Gibbs posterior

Given observations  $y$  and a prior  $\pi$  on  $\mathcal{X}$ .

The Gibbs posterior is

$$\Pi_n(A | y) = \frac{\int_A \exp(-\ell_n(x, y; \theta)) d\pi(x)}{Z_n(y)}, \quad A \subset \Theta \times X$$

$$Z_n(y) = \int_{\mathcal{X}} \exp(-\ell_n(x, y; \theta)) d\pi(x).$$

Two questions

- (1) Is  $\lim_{n \rightarrow \infty} \Pi_n(\cdot | y)$  unique.
- (2) Does  $\lim_{n \rightarrow \infty} \Pi_n(\cdot | y)$  concentrate around  $T$ .

# Gibbs posterior

- (1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.
- (2) If  $\ell_n$  is the negative log likelihood then recover standard posterior.
- (3) Robust to misspecification, robust statistics.
- (4) Calibration/violation of likelihood principle

$$\Pi_n(A | y) = \frac{\int_A \exp(-\psi \ell_n(x, y; \theta)) d\pi(x)}{Z_n(y)}$$



# Gibbs measures

Given  $\mathcal{X}$ , the map  $S$ , a potential function  $f$ , and a measure  $\mu_0$

$$G_n(x; \mu_0, f) = \frac{\exp(\sum_{k=1}^n f(S^k x))}{\int_{\mathcal{X}} \exp(\sum_{k=1}^n f(S^k x)) d\mu_0}.$$

The Gibbs measure is the limit point of the sequence  $G_n(x; \mu_0, f)$  and the Gibbs measure is denoted as  $\mu_0(f)$ .

Recall the Gibbs posterior

$$\Pi_n(x | y) = \frac{\exp(-\sum_{k=1}^n \ell(S^k x, T^k y))}{\int_{\mathcal{X}} \exp(-\sum_{k=1}^n \ell(S^k x, T^k y)) d\pi(x)}.$$

# Gibbs measures

Given  $\mathcal{X}$ , the map  $S$ , a potential function  $f$ , and a measure  $\mu_0$

$$G_n(x; \mu_0, f) = \frac{\exp(\sum_{k=1}^n f(S^k x))}{\int_{\mathcal{X}} \exp(\sum_{k=1}^n f(S^k x)) d\mu_0}.$$

The Gibbs measure is the limit point of the sequence  $G_n(x; \mu_0, f)$  and the Gibbs measure is denoted as  $\mu_0(f)$ .

Recall the Gibbs posterior

$$\Pi_n(x | y) = \frac{\exp(-\sum_{k=1}^n \ell(S^k x, T^k y))}{\int_{\mathcal{X}} \exp(-\sum_{k=1}^n \ell(S^k x, T^k y)) d\pi(x)}.$$

# Sequence space model

Alphabet  $\mathcal{A}$  is a finite set ( $|\mathcal{A}| = N$ ) and  $\Sigma = \mathcal{A}^{\mathbb{Z}}$ .



# Gibbs measures

Given  $\mathcal{X}$ , the map  $S$ , a potential function  $f$ , and a measure  $\mu_0$

$$G_n(x; \mu_0, f) = \frac{\exp(\sum_{k=1}^n f(S^k x))}{\int_{\mathcal{X}} \exp(\sum_{k=1}^n f(S^k x)) d\mu_0}.$$

The Gibbs measure is the limit point of the sequence  $G_n(x; \mu_0, f)$  and the Gibbs measure is denoted as  $\mu_0(f)$ .

Recall the Gibbs posterior

$$\Pi_n(x | y) = \frac{\exp(-\sum_{k=1}^n \ell(S^k x, T^k y))}{\int_{\mathcal{X}} \exp(-\sum_{k=1}^n \ell(S^k x, T^k y)) d\pi(x)}.$$

# Sequence space model

Alphabet  $\mathcal{A}$  is a finite set ( $|\mathcal{A}| = N$ ) and  $\Sigma = \mathcal{A}^{\mathbb{Z}}$ .

Left shift operator  $\sigma : \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{A}^{\mathbb{Z}}$  with  $(\sigma x)_i = x_{i+1}$ .

# Sequence space model

Alphabet  $\mathcal{A}$  is a finite set ( $|\mathcal{A}| = N$ ) and  $\Sigma = \mathcal{A}^{\mathbb{Z}}$ .

Left shift operator  $\sigma : \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{A}^{\mathbb{Z}}$  with  $(\sigma x)_i = x_{i+1}$ .

# Sequence space model

Alphabet  $\mathcal{A}$  is a finite set ( $|\mathcal{A}| = N$ ) and  $\Sigma = \mathcal{A}^{\mathbb{Z}}$ .

Left shift operator  $\sigma : \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{A}^{\mathbb{Z}}$  with  $(\sigma x)_i = x_{i+1}$ .

The set obtained by forbidding a finite number of words  $\mathcal{F}$

$$\Sigma_{\mathcal{F}} = \{x \in \mathcal{A}^{\mathbb{Z}} \mid x_{[i,j]} \neq u \forall i, j \in \mathbb{Z}, u \in \mathcal{F}\}$$

is a shift of finite type (SFT)




# Sequence space model

Alphabet  $\mathcal{A}$  is a finite set ( $|\mathcal{A}| = N$ ) and  $\Sigma = \mathcal{A}^{\mathbb{Z}}$ .

Left shift operator  $\sigma : \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{A}^{\mathbb{Z}}$  with  $(\sigma x)_i = x_{i+1}$ .

The set obtained by forbidding a finite number of words  $\mathcal{F}$

$$\Sigma_{\mathcal{F}} = \{x \in \mathcal{A}^{\mathbb{Z}} \mid x_{[i,j]} \neq u \forall i, j \in \mathbb{Z}, u \in \mathcal{F}\}$$

is a shift of finite type (SFT) 

# Sequence space model

The restriction of the shift maps encoded by matrix  $A$

$$\Sigma_A = \{(a_i)_{i=-\infty}^{\infty} \in \Sigma_{\mathcal{F}}, \quad A_{a_i, a_{i+1}} = 1 \quad \forall i \in \mathbb{Z}\}$$

are called a topological Markov chain or a 1-step SFT.  
One can similarly define  $m$ -step SFT.

# Sequence space model

The restriction of the shift maps encoded by matrix  $A$

$$\Sigma_A = \{(a_i)_{i=-\infty}^{\infty} \in \Sigma_{\mathcal{F}}, \quad A_{a_i, a_{i+1}} = 1 \quad \forall i \in \mathbb{Z}\}$$

are called a topological Markov chain or a 1-step SFT.

One can similarly define  $m$ -step SFT.

$\Sigma_A$  is mixing if and only if there exists  $n \geq 1$  such that  $A^n$  contains all positive entries.

# Sequence space model

The restriction of the shift maps encoded by matrix  $A$

$$\Sigma_A = \{(a_i)_{i=-\infty}^{\infty} \in \Sigma_{\mathcal{F}}, \quad A_{a_i, a_{i+1}} = 1 \quad \forall i \in \mathbb{Z}\}$$

are called a topological Markov chain or a 1-step SFT.

One can similarly define  $m$ -step SFT.

$\Sigma_A$  is mixing if and only if there exists  $n \geq 1$  such that  $A^n$  contains all positive entries.



# Sequence space model

The restriction of the shift maps encoded by matrix  $A$

$$\Sigma_A = \{(a_i)_{i=-\infty}^{\infty} \in \Sigma_{\mathcal{F}}, \quad A_{a_i, a_{i+1}} = 1 \quad \forall i \in \mathbb{Z}\}$$

are called a topological Markov chain or a 1-step SFT. One can similarly define  $m$ -step SFT.

$\Sigma_A$  is mixing if and only if there exists  $n \geq 1$  such that  $A^n$  contains all positive entries.

For  $x \in \Sigma_A$ , let  $x[i, j] = \{y \in \Sigma_A : x_i^j = y_i^j\}$ .

⊙.

# Sequence space model

The restriction of the shift maps encoded by matrix  $A$

$$\Sigma_A = \{(a_i)_{i=-\infty}^{\infty} \in \Sigma_{\mathcal{F}}, \quad A_{a_i, a_{i+1}} = 1 \quad \forall i \in \mathbb{Z}\}$$

are called a topological Markov chain or a 1-step SFT. One can similarly define  $m$ -step SFT.

$\Sigma_A$  is mixing if and only if there exists  $n \geq 1$  such that  $A^n$  contains all positive entries.

For  $x \in \Sigma_A$ , let  $x[i, j] = \{y \in \Sigma_A : x_i^j = y_i^j\}$ .







# Gibbs measure

## Definition

Let  $f : \Sigma_{\mathcal{F}} \rightarrow \mathbb{R}$  be continuous. A measure  $\mu$  on  $\Sigma_{\mathcal{F}}$  has the Gibbs property for  $f$  if there exists  $K > 1$  and  $\mathcal{P} \in \mathbb{R}$  such that for all  $x \in \mathcal{A}^{\mathbb{Z}}$  and  $m \geq 1$ ,

$$K^{-1} \leq \frac{\mu(x[0, m-1])}{\exp(-\mathcal{P}m + \sum_{k=0}^{m-1} f(\sigma^k(x)))} \leq K.$$

## Theorem (Bowen)

If  $\Sigma_{\mathcal{F}}$  is a mixing SFT, and  $f : \Sigma_{\mathcal{F}} \rightarrow \mathbb{R}$  is Hölder continuous, then there exists a unique Gibbs measure for  $f$  on  $\Sigma_{\mathcal{F}}$ .

# Gibbs measure

## Definition

Let  $f : \Sigma_{\mathcal{F}} \rightarrow \mathbb{R}$  be continuous. A measure  $\mu$  on  $\Sigma_{\mathcal{F}}$  has the Gibbs property for  $f$  if there exists  $K > 1$  and  $\mathcal{P} \in \mathbb{R}$  such that for all  $x \in \mathcal{A}^{\mathbb{Z}}$  and  $m \geq 1$ ,

$$K^{-1} \leq \frac{\mu(x[0, m-1])}{\exp(-\mathcal{P}m + \sum_{k=0}^{m-1} f(\sigma^k(x)))} \leq K.$$

## Theorem (Bowen)

If  $\Sigma_{\mathcal{F}}$  is a mixing SFT, and  $f : \Sigma_{\mathcal{F}} \rightarrow \mathbb{R}$  is Hölder continuous, then there exists a unique Gibbs measure for  $f$  on  $\Sigma_{\mathcal{F}}$ .

$f : \Sigma_{\mathcal{F}} \rightarrow \mathbb{R}$  is called a potential, and  $\mathcal{P} = \mathcal{P}(f)$  is its pressure.

## The model class

We consider families of dependent processes as follows.

Let  $\Theta$  be a compact metric space.

# The model class

We consider families of dependent processes as follows.

Let  $\Theta$  be a compact metric space.

Let  $\{f_\theta : \theta \in \Theta\}$  be a continuously parametrized family of Hölder continuous potential functions.

Let  $\{\mu_\theta : \theta \in \Theta\}$  be the corresponding family of Gibbs measures.

Markov chains of all orders are included in these model classes.





# Observation densities

We consider a general observational model as follows.

Let  $\lambda$  be a Borel measure on  $\mathcal{Y}$

Let  $g : \Theta \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, \infty)$  be a measurable function such that for all  $\theta \in \Theta$  and  $x \in \mathcal{X}$ ,

$$\int g(\theta, x, y) \lambda(dy) = 1.$$

# Observation densities

We consider a general observational model as follows.

Let  $\lambda$  be a Borel measure on  $\mathcal{Y}$

Let  $g : \Theta \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, \infty)$  be a measurable function such that for all  $\theta \in \Theta$  and  $x \in \mathcal{X}$ ,

$$\int g(\theta, x, y) \lambda(dy) = 1.$$

- ▶ We write  $g_\theta(\cdot | x)$  instead of  $g(\theta, x, \cdot)$ , and we interpret it as a conditional density on  $\mathcal{Y}$  given  $\theta$  and  $x$ .
- ▶ We require several integrability and regularity conditions on  $g$ .

# Hidden Gibbs processes

Given  $\theta \in \Theta$ , the marginal likelihood of  $y_0^{n-1}$  is

$$p_\theta(y_0^{n-1}) = \int \prod_{k=0}^{n-1} g_\theta(y_k | \sigma^k(x)) \mu_\theta(dx).$$

Equivalently, we have

$$\begin{aligned} X_0 &\sim \mu_\theta \\ X_{n+1} &= \sigma(X_n) \\ Y_n &\sim g_\theta(y | X_n) \lambda(dy). \end{aligned}$$

Let  $\mathbb{P}_\theta^Y$  denote the distribution of the process  $\{Y_n\}_{n \geq 0}$  under  $\theta$ .



# Hidden Gibbs processes

Given  $\theta \in \Theta$ , the marginal likelihood of  $y_0^{n-1}$  is

$$p_\theta(y_0^{n-1}) = \int \prod_{k=0}^{n-1} g_\theta(y_k \mid \sigma^k(x)) \mu_\theta(dx).$$

Equivalently, we have

$$\begin{aligned} X_0 &\sim \mu_\theta \\ X_{n+1} &= \sigma(X_n) \\ Y_n &\sim g_\theta(y \mid X_n) \lambda(dy). \end{aligned}$$

Let  $\mathbb{P}_\theta^Y$  denote the distribution of the process  $\{Y_n\}_{n \geq 0}$  under  $\theta$ .

# Hidden Gibbs processes

Given  $\theta \in \Theta$ , the marginal likelihood of  $y_0^{n-1}$  is

$$p_\theta(y_0^{n-1}) = \int \prod_{k=0}^{n-1} g_\theta(y_k \mid \sigma^k(x)) \mu_\theta(dx).$$

Equivalently, we have

$$\begin{aligned} X_0 &\sim \mu_\theta \\ X_{n+1} &= \sigma(X_n) \\ Y_n &\sim g_\theta(y \mid X_n) \lambda(dy). \end{aligned}$$

Let  $\mathbb{P}_\theta^Y$  denote the distribution of the process  $\{Y_n\}_{n \geq 0}$  under  $\theta$ .

# Hidden Gibbs processes

Given  $\theta \in \Theta$ , the marginal likelihood of  $y_0^{n-1}$  is

$$p_\theta(y_0^{n-1}) = \int \prod_{k=0}^{n-1} g_\theta(y_k \mid \sigma^k(x)) \mu_\theta(dx).$$

Equivalently, we have

$$\begin{aligned} X_0 &\sim \mu_\theta \\ X_{n+1} &= \sigma(X_n) \\ Y_n &\sim g_\theta(y \mid X_n) \lambda(dy). \end{aligned}$$

Let  $\mathbb{P}_\theta^Y$  denote the distribution of the process  $\{Y_n\}_{n \geq 0}$  under  $\theta$ .





# Bayesian inference

Given observations  $Y_0^{n-1}$ , the posterior

$$\Pi_n(E \mid Y_0^{n-1}) = \frac{\int_E p_\theta(Y_0^{n-1}) \pi(d\theta)}{\int_\Theta p_\theta(Y_0^{n-1}) \pi(d\theta)}, \quad E \subset \Theta.$$

# Posterior consistency

For  $\theta \in \Theta$ , let  $[\theta] = \{\theta' \in \Theta : \mathbb{P}_{\theta}^Y = \mathbb{P}_{\theta'}^Y\}$ .

# Posterior consistency

For  $\theta \in \Theta$ , let  $[\theta] = \{\theta' \in \Theta : \mathbb{P}_{\theta}^Y = \mathbb{P}_{\theta'}^Y\}$ .

# Posterior consistency

For  $\theta \in \Theta$ , let  $[\theta] = \{\theta' \in \Theta : \mathbb{P}_\theta^Y = \mathbb{P}_{\theta'}^Y\}$ .

## Theorem (McGoff-M-Nobel)

Suppose  $\pi$  is fully supported on  $\Theta$ , and let  $\theta_0 \in \Theta$ . Then for any neighborhood  $U$  of  $[\theta_0]$ ,

$$\Pi_n(\Theta \setminus U \mid Y_0^{n-1}) \xrightarrow{\mathbb{P}_{\theta_0}^Y} 0, \quad \mathbb{P}_{\theta_0}^Y - a.s.$$



# Posterior consistency

For  $\theta \in \Theta$ , let  $[\theta] = \{\theta' \in \Theta : \mathbb{P}_\theta^Y = \mathbb{P}_{\theta'}^Y\}$ .

## Theorem (McGoff-M-Nobel)

Suppose  $\pi$  is fully supported on  $\Theta$ , and let  $\theta_0 \in \Theta$ . Then for any neighborhood  $U$  of  $[\theta_0]$ ,

$$\Pi_n(\Theta \setminus U \mid Y_0^{n-1}) \rightarrow 0, \quad \mathbb{P}_{\theta_0}^Y \text{ a.s.}$$

## More general setting

We consider

$\Theta$  as before;

$\mathcal{X}$  and  $\{\mu_\theta : \theta \in \Theta\}$  as before;

$\ell : \Theta \times \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$  a continuous loss function;

$\{Y_n\}_{n \geq 0}$  an arbitrary stationary ergodic process.

$y_0^{n-1} := (y_0, \dots, y_{n-1}) \in \mathcal{Y}^n$ .

The loss incurred by parameter  $\theta$  and initial condition  $x$

$$\ell(\theta, x; y_0^{n-1}) = \sum_{k=0}^{n-1} \ell(\theta, \sigma^k(x), y_k).$$

## More general setting

We consider

$\Theta$  as before;

$\mathcal{X}$  and  $\{\mu_\theta : \theta \in \Theta\}$  as before;

$\ell : \Theta \times \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$  a continuous loss function;

$\{Y_n\}_{n \geq 0}$  an arbitrary stationary ergodic process.

$y_0^{n-1} := (y_0, \dots, y_{n-1}) \in \mathcal{Y}^n$ .

The loss incurred by parameter  $\theta$  and initial condition  $x$

$$\ell(\theta, x; y_0^{n-1}) = \sum_{k=0}^{n-1} \ell(\theta, \sigma^k(x), y_k).$$

# Posterior consistency

For  $\theta \in \Theta$ , let  $[\theta] = \{\theta' \in \Theta : \mathbb{P}_\theta^Y = \mathbb{P}_{\theta'}^Y\}$ .

## Theorem (McGoff-M-Nobel)

Suppose  $\pi$  is fully supported on  $\Theta$ , and let  $\theta_0 \in \Theta$ . Then for any neighborhood  $U$  of  $[\theta_0]$ ,

$$\Pi_n(\Theta \setminus U \mid Y_0^{n-1}) \rightarrow 0, \quad \mathbb{P}_{\theta_0}^Y - a.s.$$



## More general setting

We consider

$\Theta$  as before;

$\mathcal{X}$  and  $\{\mu_\theta : \theta \in \Theta\}$  as before;

$\ell : \Theta \times \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$  a continuous loss function;

$\{Y_n\}_{n \geq 0}$  an arbitrary stationary ergodic process.

$y_0^{n-1} := (y_0, \dots, y_{n-1}) \in \mathcal{Y}^n$ .

The loss incurred by parameter  $\theta$  and initial condition  $x$

$$\ell(\theta, x; y_0^{n-1}) = \sum_{k=0}^{n-1} \ell(\theta, \sigma^k(x), y_k).$$



## More general setting

We consider

$\Theta$  as before;

$\mathcal{X}$  and  $\{\mu_\theta : \theta \in \Theta\}$  as before;

$\ell : \Theta \times \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$  a continuous loss function;

$\{Y_n\}_{n \geq 0}$  an arbitrary stationary ergodic process.

$y_0^{n-1} := (y_0, \dots, y_{n-1}) \in \mathcal{Y}^n$ .

The loss incurred by parameter  $\theta$  and initial condition  $x$

$$\ell(\theta, x; y_0^{n-1}) = \sum_{k=0}^{n-1} \ell(\theta, \sigma^k(x), y_k).$$

## More general setting

We consider

$\Theta$  as before;

$\mathcal{X}$  and  $\{\mu_\theta : \theta \in \Theta\}$  as before;

$\ell : \Theta \times \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$  a continuous loss function;

$\{Y_n\}_{n \geq 0}$  an arbitrary stationary ergodic process.

$y_0^{n-1} := (y_0, \dots, y_{n-1}) \in \mathcal{Y}^n$ .

The loss incurred by parameter  $\theta$  and initial condition  $x$

$$\ell(\theta, x; y_0^{n-1}) = \sum_{k=0}^{n-1} \ell(\theta, \sigma^k(x), y_k).$$







# Gibbs posterior distribution

Prior  $\pi$  and same  $\{\mu_\theta : \theta \in \Theta\}$  as before.

$P_0$  on  $\Theta \times \mathcal{X}$  is

$$P_0(A \times B) = \int_A \mu_\theta(B) \pi(d\theta).$$

The Gibbs posterior is

$$\Pi_n(A | y_0^{n-1}) = \frac{\int_A \exp\left(-\ell(\theta, x; y_0^{n-1})\right) P_0(d\theta, dx)}{Z_n(y_0^{n-1})}, A \subset \Theta \times \mathcal{X}$$

where  $Z_n(y_0^{n-1})$  is a normalization constant.

# Questions

1. Does the following limit exist with  $\mathbb{P}^Y$ -probability 1,

$$\lim_n \frac{1}{n} \log Z_n(Y_0^{n-1}),$$

and if so, what is it?



# Gibbs posterior distribution

Prior  $\pi$  and same  $\{\mu_\theta : \theta \in \Theta\}$  as before.

$P_0$  on  $\Theta \times \mathcal{X}$  is

$$P_0(A \times B) = \int_A \mu_\theta(B) \pi(d\theta).$$

The Gibbs posterior is

$$\Pi_n(A | y_0^{n-1}) = \frac{\int_A \exp\left(-\ell(\theta, x; y_0^{n-1})\right) P_0(d\theta, dx)}{Z_n(y_0^{n-1})}, A \subset \Theta \times \mathcal{X}$$

where  $Z_n(y_0^{n-1})$  is a normalization constant.

# Questions

1. Does the following limit exist with  $\mathbb{P}^Y$ -probability 1,

$$\lim_n \frac{1}{n} \log Z_n(Y_0^{n-1}),$$

and if so, what is it?

2. What can be said about the convergence of the posterior distributions  $\{\Pi_n\}_n$ ?

# Joinings

## Definition (Joining)

Let  $(X, \mathcal{A}, \mu, T)$  and  $(Y, \mathcal{B}, \nu, S)$  be two dynamical systems. A joining of  $T$  and  $S$  is a probability measure  $\lambda$  on  $X \times Y$ , with marginals  $\mu$  and  $\nu$  respectively, and invariant to the product map  $T \times S$ .

# Questions

1. Does the following limit exist with  $\mathbb{P}^Y$ -probability 1,

$$\lim_n \frac{1}{n} \log Z_n(Y_0^{n-1}),$$

and if so, what is it?

2. What can be said about the convergence of the posterior distributions  $\{\Pi_n\}_n$ ?



# Joinings

## Definition (Joining)

Let  $(X, A, \mu, T)$  and  $(Y, B, \nu, S)$  be two dynamical systems. A joining of  $T$  and  $S$  is a probability measure  $\lambda$  on  $X \times Y$ , with marginals  $\mu$  and  $\nu$  respectively, and invariant to the product map  $T \times S$ .



# Joinings

## Definition (Joining)

Let  $(X, \mathcal{A}, \mu, T)$  and  $(Y, \mathcal{B}, \nu, S)$  be two dynamical systems. A joining of  $T$  and  $S$  is a probability measure  $\lambda$  on  $X \times Y$ , with marginals  $\mu$  and  $\nu$  respectively, and invariant to the product map  $T \times S$ .

# Joinings

## Definition (Joining)

Let  $(X, A, \mu, T)$  and  $(Y, B, \nu, S)$  be two dynamical systems. A joining of  $T$  and  $S$  is a probability measure  $\lambda$  on  $X \times Y$ , with marginals  $\mu$  and  $\nu$  respectively, and invariant to the product map  $T \times S$ .

## Definition (Coupling)

A coupling of two random variable  $X$  and  $X'$  taking values in  $(E, \mathcal{E})$  is any pair of random variables  $(Y, Y')$  taking values in  $(E \times E, \mathcal{E} \times \mathcal{E})$  whose marginals have the same distribution as  $X$  and  $X'$ ,  $X \stackrel{D}{=} Y$  and  $X' \stackrel{D}{=} Y'$ .



# Joinings

## Definition (Joining)

Let  $(X, A, \mu, T)$  and  $(Y, B, \nu, S)$  be two dynamical systems. A joining of  $T$  and  $S$  is a probability measure  $\lambda$  on  $X \times Y$ , with marginals  $\mu$  and  $\nu$  respectively, and invariant to the product map  $T \times S$ .

## Definition (Coupling)

A coupling of two random variable  $X$  and  $X'$  taking values in  $(E, \mathcal{E})$  is any pair of random variables  $(Y, Y')$  taking values in  $(E \times E, \mathcal{E} \times \mathcal{E})$  whose marginals have the same distribution as  $X$  and  $X'$ ,  $X \stackrel{D}{=} Y$  and  $X' \stackrel{D}{=} Y'$ .





# Joinings

## Definition (Joining)

Let  $(X, A, \mu, T)$  and  $(Y, B, \nu, S)$  be two dynamical systems. A joining of  $T$  and  $S$  is a probability measure  $\lambda$  on  $X \times Y$ , with marginals  $\mu$  and  $\nu$  respectively, and invariant to the product map  $T \times S$ .

## Definition (Coupling)

A coupling of two random variable  $X$  and  $X'$  taking values in  $(E, \mathcal{E})$  is any pair of random variables  $(Y, Y')$  taking values in  $(E \times E, \mathcal{E} \times \mathcal{E})$  whose marginals have the same distribution as  $X$  and  $X'$ ,  $X \stackrel{D}{=} Y$  and  $X' \stackrel{D}{=} Y'$ .

# Joinings

A stationary  $\mathcal{X}$ -valued process  $\{X_n\}_{n \geq 0}$  is in  $\mathcal{P}(\mathcal{X}, \sigma)$  if

$$X_{n+1} = \sigma(X_n), \quad \forall n, \text{ wp } 1.$$

A joining of  $(\mathcal{X}, \sigma)$  with  $\{Y_n\}_{n \geq 0}$  is a stationary bi-variate process  $(\mathbf{U}, \mathbf{V}) = \{(U_n, V_n)\}_{n \geq 0}$  on  $\mathcal{X} \times \mathcal{Y}$  such that

$\mathbf{U} = \{U_n\}_{n \geq 0}$  is in  $\mathcal{P}(\mathcal{X}, \sigma)$ , and

$\mathbf{V} = \{V_n\}_{n \geq 0}$  is equal to  $\{Y_n\}_{n \geq 0}$  in distribution.

The set of joinings of  $(\mathcal{X}, \sigma)$  with  $\{Y_n\}_{n \geq 0}$  is denoted by  $\mathcal{J}$ .



# Convergence theorem

## Theorem (McGoff-M-Nobel)

Suppose  $\pi$  is fully supported and  $\ell$  satisfies appropriate regularity and integrability conditions. Then there exists a lower semicontinuous function  $\phi : \Theta \rightarrow \mathbb{R}$  such that with probability 1,

$$\lim_n -\frac{1}{n} \log Z_n(y) = \inf_{\theta \in \Theta} \phi(\theta).$$

The above is the rate function in the large deviation sense.



# Variational formulation of $Z_n(y)$ – average cost

Limiting average cost

$$\lim_{n \rightarrow \infty} \frac{1}{n} \int_{\mathcal{X}} \ell_n(x, y) d\lambda_y(x) = \int \ell d\lambda.$$

# Convergence

## Theorem (McGoff-M.-Nobel)

Suppose a Gibbs prior, then for  $\nu$  almost every  $y$ ,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log Z_n(y) = \inf_{\lambda \in \mathcal{J}} \left\{ \int \ell \, d\lambda + F(\lambda, \mu_\theta) \right\},$$

and the infimum in the above expression is attained.



# Convergence

## Theorem (McGoff-M.-Nobel)

Suppose a Gibbs prior, then for  $\nu$  almost every  $y$ ,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log Z_n(y) = \inf_{\lambda \in \mathcal{J}} \left\{ \int \ell \, d\lambda + F(\lambda, \mu_\theta) \right\},$$

and the infimum in the above expression is attained.

# Convergence

## Theorem (McGoff-M.-Nobel)

Suppose a Gibbs prior, then for  $\nu$  almost every  $y$ ,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log Z_n(y) = \inf_{\lambda \in \mathcal{J}} \left\{ \int \ell \, d\lambda + F(\lambda, \mu_\theta) \right\},$$

and the infimum in the above expression is attained.



# Bayes as a variational problem

Suppose a Gibbs prior, then for  $\nu$  almost every  $y$ ,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log Z_n(y) = \inf_{\lambda \in \mathcal{J}} \left\{ \int \ell \, d\lambda + F(\lambda, \mu_\theta) \right\},$$

A way to write Bayes rule

$$\Pi(\theta \mid x) = \arg \min_{\mu} \left\{ \int_{\theta} \ell(\theta, x) \, d\mu(\theta) + d_{KL}(\mu, \pi) \right\}$$



# Convergence

## Proposition (McGoff-M.-Nobel)

Suppose a Gibbs prior and consider the pressure

$$\mathcal{P} = \inf_{\lambda \in \mathcal{J}} \left\{ \int \ell d\lambda + F(\lambda, \mu_\theta) \right\}$$
$$\theta_* = \arg \min_{\theta \in \Theta} \mathcal{P}.$$

For all  $\varepsilon > 0$

$$P(d(S_{\theta_*}, T) < \varepsilon) \rightarrow 1 \text{ a.s as } n \rightarrow \infty.$$

# Convergence

## Proposition (McGoff-M.-Nobel)

Suppose a Gibbs prior and consider the pressure

$$\mathcal{P} = \inf_{\lambda \in \mathcal{J}} \left\{ \int \ell d\lambda + F(\lambda, \mu_\theta) \right\}$$
$$\theta_* = \arg \min_{\theta \in \Theta} \mathcal{P}.$$

For all  $\varepsilon > 0$

$$P(d(S_{\theta_*}, T) < \varepsilon) \rightarrow 1 \text{ a.s as } n \rightarrow \infty.$$

# Ideas used in proofs

The main technical tools include:

- (1) The thermodynamic formalism from dynamical systems (as developed by Sinai, Ruelle, Bowen, and others);
- (2) The theory of joinings, introduced by Furstenberg;
- (3) Aspects of the “random” thermodynamic formalism of Kifer.



# Ideas used in proofs

The main technical tools include:

- (1) The thermodynamic formalism from dynamical systems (as developed by Sinai, Ruelle, Bowen, and others);
- (2) The theory of joinings, introduced by Furstenberg;
- (3) Aspects of the “random” thermodynamic formalism of Kifer.

# Ideas used in proofs

The main technical tools include:

- (1) The thermodynamic formalism from dynamical systems (as developed by Sinai, Ruelle, Bowen, and others);
- (2) The theory of joinings, introduced by Furstenberg;
- (3) Aspects of the “random” thermodynamic formalism of Kifer.

# Contributions

**Reframes** Gibbs posterior consistency as two-stage process: first find the limiting variational problem, and then analyze this problem to address consistency.

# Contributions

**Reframes** Gibbs posterior consistency as two-stage process: first find the limiting variational problem, and then analyze this problem to address consistency.





# Questions

## **Statistics questions.**

What types of observations and models are amenable to this analysis?

For which combinations of observations and models can one establish posterior consistency?

## **Dynamics questions.**

How far can the thermodynamic formalism be pushed?

Under what conditions is there a limiting variational characterization?

Under what conditions is there a unique equilibrium joining?

# Open problems

- (1) Rates of convergence for a family of dynamical systems  $\mathcal{F}$ .

# Open problems

- (1) Rates of convergence for a family of dynamical systems  $\mathcal{F}$ .
- (2) General conditions for learnability in dynamical systems.

# Open problems

- (1) Rates of convergence for a family of dynamical systems  $\mathcal{F}$ .
- (2) General conditions for learnability in dynamical systems.

# Open problems

- (1) Rates of convergence for a family of dynamical systems  $\mathcal{F}$ .
- (2) General conditions for learnability in dynamical systems.



# Open problems

- (1) Rates of convergence for a family of dynamical systems  $\mathcal{F}$ .
- (2) General conditions for learnability in dynamical systems.
- (3) Extension to continuous time dynamics, differential equations.

# Acknowledgements

## Thanks:

Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Natesh Pillai.

# Open problems

- (1) Rates of convergence for a family of dynamical systems  $\mathcal{F}$ .
- (2) General conditions for learnability in dynamical systems.
- (3) Extension to continuous time dynamics, differential equations.
- (4) Computational issues.
- (5) Integration of ideas from statistical models of time series and dynamical systems theory.

# Acknowledgements

## Thanks:

Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Natesh Pillai.

## Funding:

- ▶ NSF DMS, CCF, CISE, DEB
- ▶ AFOSR
- ▶ DARPA
- ▶ NIH

# Acknowledgements

## Thanks:

Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Natesh Pillai.

## Funding:

- ▶ NSF DMS, CCF, CISE, DEB
- ▶ AFOSR
- ▶ DARPA
- ▶ NIH





# Acknowledgements

## Thanks:

Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Natesh Pillai.

## Funding:

- ▶ NSF DMS, CCF, CISE, DEB
- ▶ AFOSR
- ▶ DARPA
- ▶ NIH



# Acknowledgements

## Thanks:

Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Natesh Pillai.

## Funding:

- ▶ NSF DMS, CCF, CISE, DEB
- ▶ AFOSR
- ▶ DARPA
- ▶ NIH

# Acknowledgements

## Thanks:

Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Natesh Pillai.

## Funding:

- ▶ NSF DMS, CCF, CISE, DEB
- ▶ AFOSR
- ▶ DARPA
- ▶ NIH





# Acknowledgements

## Thanks:

Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Natesh Pillai.

## Funding:

- ▶ NSF DMS, CCF, CISE, DEB
- ▶ AFOSR
- ▶ DARPA
- ▶ NIH

# Acknowledgements

## Thanks:

Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Natesh Pillai.

## Funding:

- ▶ NSF DMS, CCF, CISE, DEB
- ▶ AFOSR
- ▶ DARPA
- ▶ NIH

# Acknowledgements

## Thanks:

Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Natesh Pillai.

## Funding:

- ▶ NSF DMS, CCF, CISE, DEB
- ▶ AFOSR
- ▶ DARPA
- ▶ NIH