Dimension Reduction in Dynamical Systems using Factor Models

Dimension reduction in physical and data sciences

Sayan Mukherjee

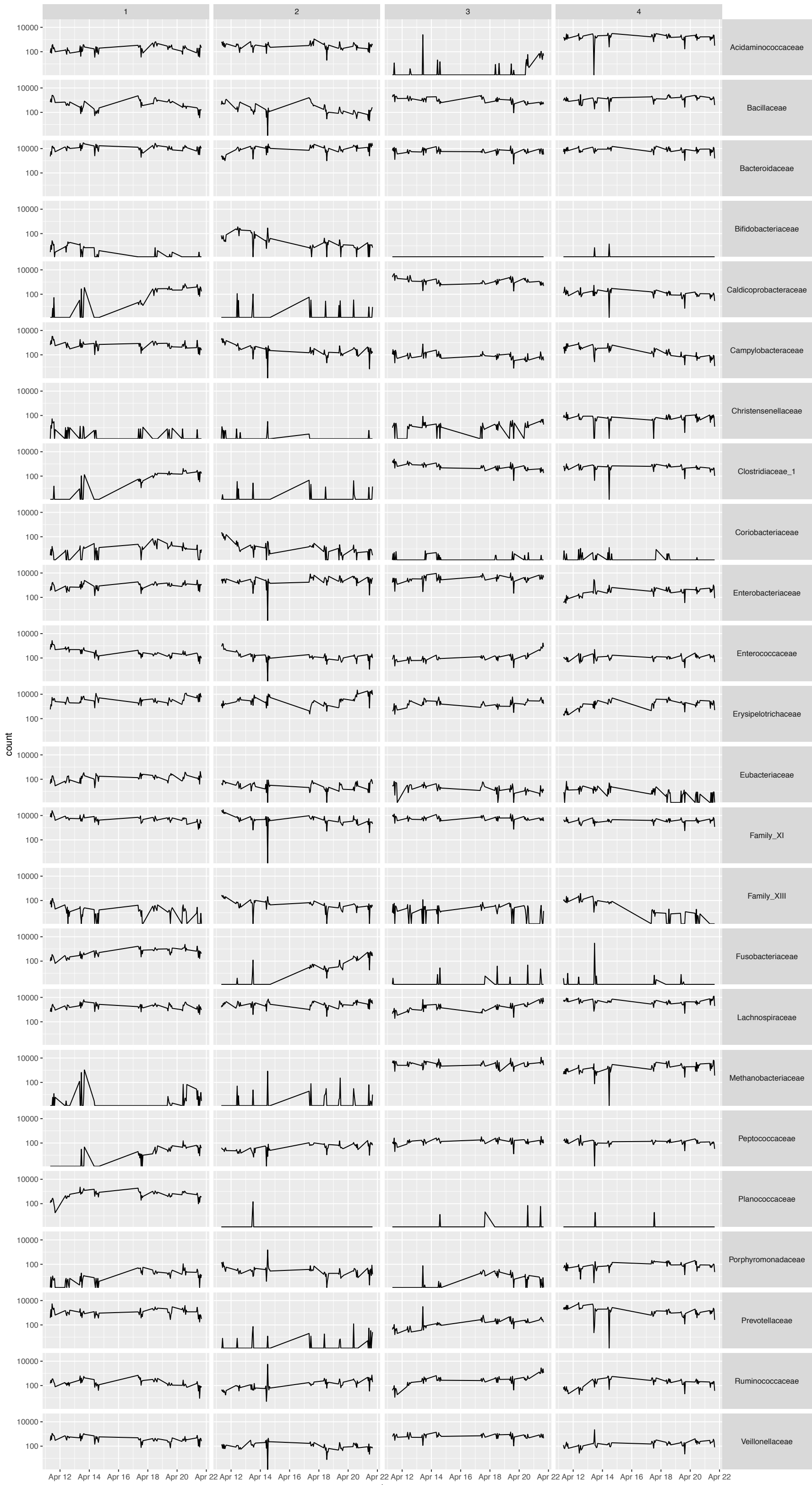
https://sayanmuk.github.io/

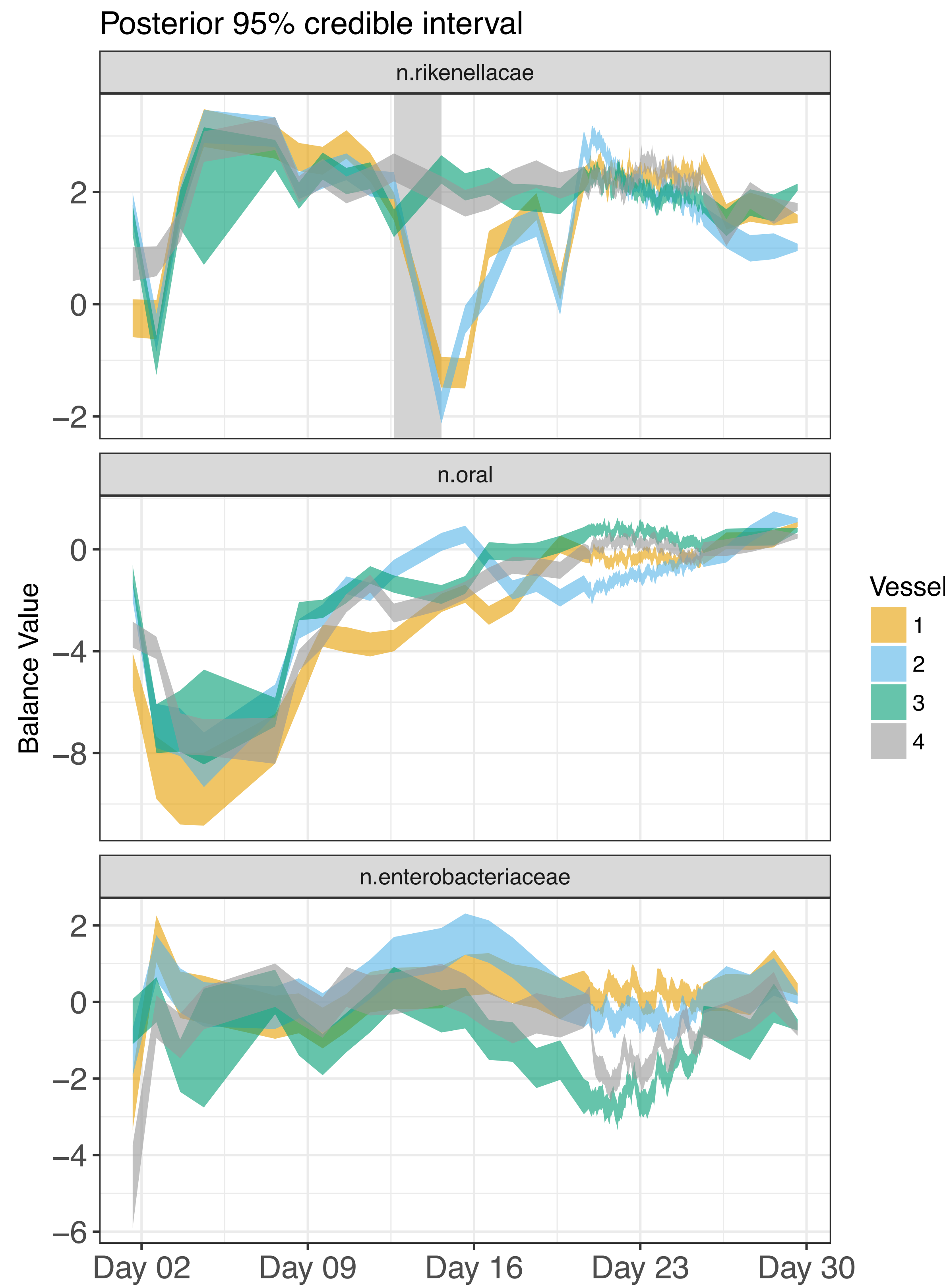
Joint work with:

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

Posterior Consistency for Dynamical Systems

Microbial ecology





General framework

We consider mathematical models of the form:

- X is the “phase space”;
- X_t is the “true state” of bioreactor (your stomach) at time t;
- Y is the “observation space”;
- Y_t is our observation at time t.

General framework

We consider mathematical models of the form:

- X is the “phase space”;
- X_t is the “true state” of bioreactor (your stomach) at time t;
- Y is the “observation space”;
- Y_t is our observation at time t.

We only have access to the observations $\{Y_{t_k}\}_{k=0}^{n}$.

General questions

Given access to the observations $\{Y_{t_k}\}_{k=0}^{n}$, we might want to ask

- what is the "true state" of the bioreactor at time t? (filtering)

General questions

Given access to the observations $\{Y_{t_k}\}_{k=0}^{n}$, 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_{n+1}? (prediction)

General questions

Given access to the observations $\{Y_{t_k}\}_{k=0}^{n}$, 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_{n+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.

Basic assumptions

How are the variables $\{X_{t_k}\}_{k=0}^{n}$ and $\{Y_{t_k}\}_{k=0}^{n}$ related?

We'll assume the process $(X_t, Y_t)_t$ has:

- **stationarity:** the rules governing both the state space and our observations don't change over time.

Basic assumptions

How are the variables $\{X_{t_k}\}_{k=0}^{n}$ and $\{Y_{t_k}\}_{k=0}^{n}$ related?

We'll assume the process $(X_t, Y_t)_t$ has:

- **stationarity:** the rules governing both the state space and our observations don't change over time.
- **Markov property:** given the microbial population today, the microbial population tomorrow is independent of the population yesterday.

Basic assumptions

How are the variables $\{X_{t_k}\}_{k=0}^{n}$ and $\{Y_{t_k}\}_{k=0}^{n}$ related?

We'll assume the process $(X_t, Y_t)_t$ has:

- **stationarity:** the rules governing both the state space and our observations don't change over time.
- **Markov property:** given the microbial population today, the microbial population tomorrow is independent of the population yesterday.
- **conditionally independent observations:** given the state of the population today, today's observation is independent of any other variables.

Such systems are called "hidden Markov models" (HMMs).

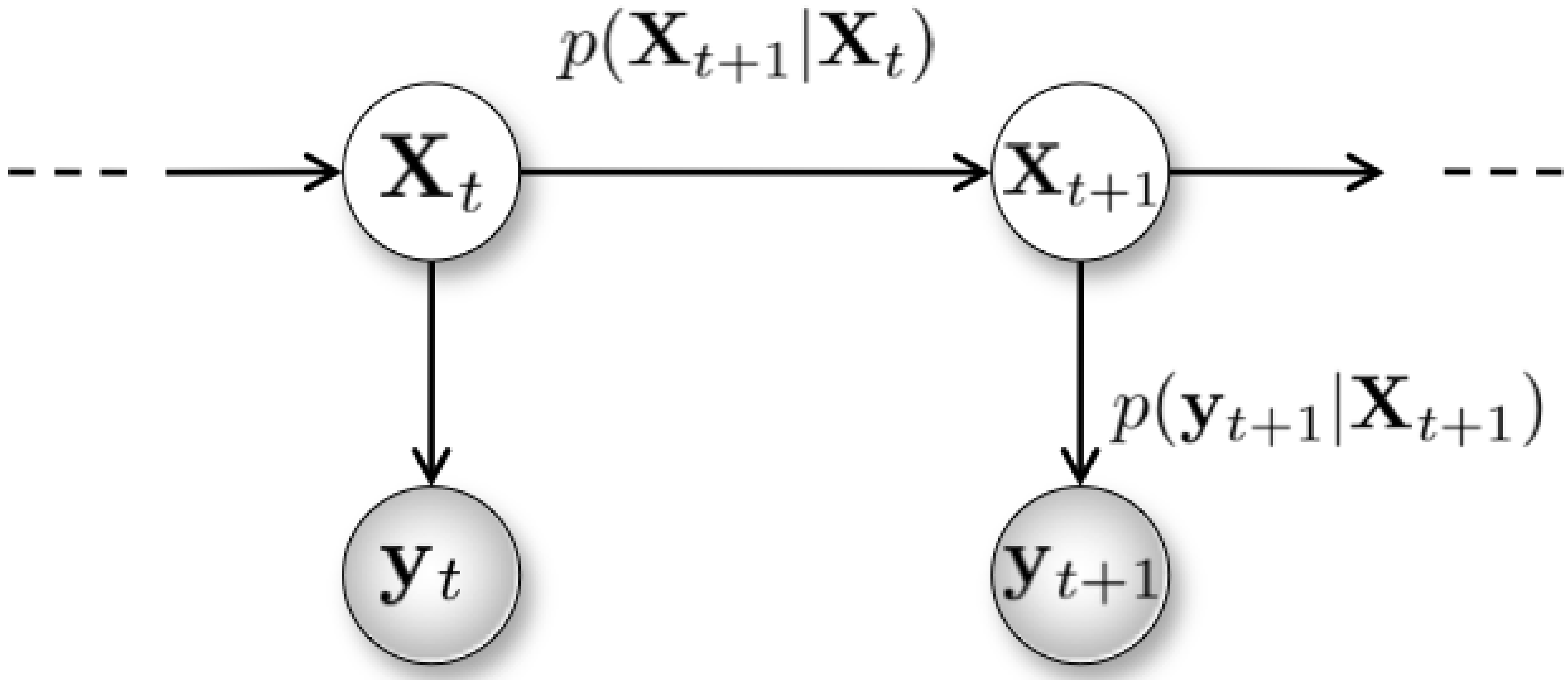
HMMs

The basic model for a HMM is

$$
\begin{aligned}
X_k &= a_\theta(X_{k-1}, W_k), \\
Y_k &= b_\theta(X_k, V_k),
\end{aligned}
$$

where $(V_k)_k$ and $(W_k)_k$ are sequences of iid random variables independent of X_0.

HMMs



$p(\mathbf{X}_{t+1}|\mathbf{X}_t)$
$\mathbf{X}_t$
$\mathbf{X}_{t+1}$
$p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1})$
$\mathbf{y}_t$
$\mathbf{y}_{t+1}$

Stochastic versus deterministic systems

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

Stochastic versus deterministic systems

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

- If the conditional distribution of $X_{t_{k+1}}$ given X_{t_k} has positive variance, then we'll say the process $(X_t)_t$ is stochastic.

Stochastic versus deterministic systems

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

- If the conditional distribution of $X_{t_{k+1}}$ given X_{t_k} 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 θ in Θ (parameter space), we have $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$, where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$

Setting for deterministic dynamics

Suppose that for each θ in Θ (parameter space), we have $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$, where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$
- $T_\theta : \mathsf{X} \to \mathsf{X}$ is a measurable map,

Setting for deterministic dynamics

Suppose that for each θ in Θ (parameter space), we have $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$, where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$
- $T_\theta : \mathsf{X} \to \mathsf{X}$ is a measurable map,
- μ_θ is a probability measure on $(\mathsf{X}, \mathcal{X})$ is T_θ-invariant if $\mu_\theta(T_\theta^{-1}A) = \mu_\theta(A), \quad \forall A \in \mathcal{X}$

Setting for deterministic dynamics

Suppose that for each θ in Θ (parameter space), we have $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$, where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$
- $T_\theta : \mathsf{X} \to \mathsf{X}$ is a measurable map,
- μ_θ is a probability measure on $(\mathsf{X}, \mathcal{X})$ is T_θ-invariant if $\mu_\theta(T_\theta^{-1} A) = \mu_\theta(A), \quad \forall A \in \mathcal{X}$
- the measure preserving system $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$ is ergodic if $T_\theta^{-1} A = A$ implies $\mu(A) = \{0, 1\}$.

Setting for deterministic dynamics

Suppose that for each θ in Θ (parameter space), we have $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$, where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$
- $T_\theta : \mathsf{X} \to \mathsf{X}$ is a measurable map,
- μ_θ is a probability measure on $(\mathsf{X}, \mathcal{X})$ is T_θ-invariant if $\mu_\theta(T_\theta^{-1}A) = \mu_\theta(A), \quad \forall A \in \mathcal{X}$
- the measure preserving system $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$ is ergodic if $T_\theta^{-1}A = A$ implies $\mu(A) = \{0, 1\}$.

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

Observational noise

Conditional likelihood: $g_\theta(y \mid x) = f(Y_t = y \mid x_t = x, \theta)$, with

$$\int g_\theta(y \mid x) d\nu(y) = 1.$$

Also $g : \Theta \times \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$.

Observational noise

Conditional likelihood: $g_\theta(y \mid x) = f(Y_t = y \mid x_t = x, \theta)$, with

$$\int g_\theta(y \mid x) d\nu(y) = 1.$$

Also $g : \Theta \times \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$.

Likelihood for y_0^n in Y^{n+1} conditioned on θ and $X_0 = x$ is

$$p_\theta(y_0^n \mid x) = \prod_{k=0}^{n} g_\theta(y_k \mid T_\theta^k(x)),$$

Observational noise

Conditional likelihood: $g_\theta(y \mid x) = f(Y_t = y \mid x_t = x, \theta)$, with

$$\int g_\theta(y \mid x) d\nu(y) = 1.$$

Also $g : \Theta \times \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$.

Likelihood for y_0^n in Y^{n+1} conditioned on θ and $X_0 = x$ is

$$p_\theta(y_0^n \mid x) = \prod_{k=0}^{n} g_\theta(y_k \mid T_\theta^k(x)),$$

and the (marginal) likelihood of observing y_0^n given θ is

$$p_\theta(y_0^n) = \int p_\theta(y_0^n \mid x)\, d\mu_\theta(x).$$

An example

- $X_0 \sim U[0, 1]$;
- $X_{k+1} = \theta X_k(1 - X_k)$;

An example

- $X_0 \sim U[0, 1]$;
- $X_{k+1} = \theta X_k(1 - X_k)$;
- $Y_k \sim N(X_k, \sigma_\theta^2)$.

Approaches to estimation

There are many approaches to estimation:

- maximum likelihood estimation,
- Bayesian estimation,
- optimization (minimization of a cost function),
- etc.

Approaches to estimation

There are many approaches to estimation:

- maximum likelihood estimation,
- Bayesian estimation,
- optimization (minimization of a cost function),
- etc.

We'll focus on one approaches:

(2) Bayesian inference.

Preliminaries

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

Preliminaries

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

Tracking systems:
Compact metrizable space $\mathcal{X} := \mathsf{X} \times \Theta$ with map $S : \mathcal{X} \to \mathcal{X}$.

$$S : \Theta \times \mathsf{X} \to \mathsf{X}, \quad S_\theta : \mathsf{X} \to \mathsf{X}.$$

Preliminaries

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

Tracking systems:
Compact metrizable space $\mathcal{X} := \mathrm{X} \times \Theta$ with map $S : \mathcal{X} \to \mathcal{X}$.

$$S : \Theta \times \mathrm{X} \to \mathrm{X}, \quad S_\theta : \mathrm{X} \to \mathrm{X}.$$

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

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

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

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 $\mathbb{R}^q$.

Classical Bayesian inference

Likelihood: Lik(data $\mid \theta$)

Classical Bayesian inference

Likelihood: Lik(data $\mid \theta$)

Prior: $\pi(\theta)$

Classical Bayesian inference

Likelihood: $\text{Lik}(\text{data} \mid \theta)$

Prior: $\pi(\theta)$

Marginal likelihood: $\int_\theta \text{Lik}(\text{data} \mid \theta) \times \pi(\theta) d\theta = \text{Pr}(\text{data})$

Classical Bayesian inference

Likelihood: $\text{Lik}(\text{data} \mid \theta)$

Prior: $\pi(\theta)$

Marginal likelihood: $\int_\theta \text{Lik}(\text{data} \mid \theta) \times \pi(\theta) d\theta = \text{Pr}(\text{data})$

Bayes posterior:

$$\pi(\theta \mid \text{data}) = \frac{\text{Lik}(\text{data} \mid \theta) \times \pi(\theta)}{\text{Pr}(\text{data})}.$$

Gibbes:

$$\pi(\theta \mid \text{data}) = \frac{\exp(-\ell(\text{data} \mid \theta)) \times \pi(\theta)}{\int_\theta \exp(-\ell(\text{data} \mid \theta)) \times \pi(\theta) d\theta}.$$

Gibbs posterior

Given observations $(y, Ty, \ldots, T^{n-1}y) \in \mathcal{Y}^n$ and a prior π on $\mathcal{X}$.

Gibbs posterior

Given observations $(y, Ty, \ldots, T^{n-1}y) \in \mathcal{Y}^n$ and a prior π on $\mathcal{X}$.

Consider the probability measure over Borel sets $A \subset \mathcal{X}$

$$\begin{aligned} P_n(A \mid y) &= \frac{\int_A \exp\big(-\ell_n(x, y)\big)\, d\pi(x)}{Z_n(y)}, \quad A \subset \Theta \times \mathrm{X} \\ Z_n(y) &= \int_{\mathcal{X}} \exp\big(-\ell_n(x, y)\big)\, d\pi(x). \end{aligned}$$

Gibbs posterior

Given observations $(y, Ty, \ldots, T^{n-1}y) \in \mathcal{Y}^n$ and a prior π on $\mathcal{X}$.

Consider the probability measure over Borel sets $A \subset \mathcal{X}$

$$
\begin{aligned}
P_n(A \mid y) &= \frac{\int_A \exp\big(-\ell_n(x, y)\big)\, d\pi(x)}{Z_n(y)}, \quad A \subset \Theta \times \mathrm{X} \\
Z_n(y) &= \int_{\mathcal{X}} \exp\big(-\ell_n(x, y)\big)\, d\pi(x).
\end{aligned}
$$

Gibbs posterior

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

Gibbs posterior

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

(2) If ℓ_n is the negative log likelihood then recover standard posterior.

Gibbs posterior

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

(2) If ℓ_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 ℓ_n is the negative log likelihood then recover standard posterior.

(3) Robust to misspecification, robust statistics.

(4) Calibration/violation of likelihood principle

$$P_n(A \mid y) = \frac{\int_A \exp\left(-\psi \ell_n(x,y)\right) \, d\pi(x)}{Z_n(y)}.$$

Mixing shifts of finite type

Alphabet $\mathcal{A}$ and $\Sigma = \mathcal{A}^{\mathbb{Z}}$.

Mixing shifts of finite type

Alphabet $\mathcal{A}$ and $\Sigma = \mathcal{A}^{\mathbb{Z}}$.
Left shift $\sigma(x)_{n+1} = x_n$.

Mixing shifts of finite type

Alphabet $\mathcal{A}$ and $\Sigma = \mathcal{A}^{\mathbb{Z}}$.
Left shift $\sigma(x)_{n+1} = x_n$.
$\mathcal{X}$ ia SFT if $\exists n \geq 0$ and $\mathcal{W} \subset \mathcal{A}^n$ such that $\mathcal{X}$ is exactly the set of sequences in Σ that contain no words from $\mathcal{W}$.

Mixing shifts of finite type

Alphabet $\mathcal{A}$ and $\Sigma = \mathcal{A}^{\mathbb{Z}}$.
Left shift $\sigma(x)_{n+1} = x_n$.
$\mathcal{X}$ ia SFT if $\exists n \geq 0$ and $\mathcal{W} \subset \mathcal{A}^n$ such that $\mathcal{X}$ is exactly the set of sequences in Σ that contain no words from $\mathcal{W}$.

The map $S : \mathcal{X} \to \mathcal{X}$ is the restriction of σ to $\mathcal{X}$.

Mixing shifts of finite type

Alphabet $\mathcal{A}$ and $\Sigma = \mathcal{A}^{\mathbb{Z}}$.
Left shift $\sigma(x)_{n+1} = x_n$.
$\mathcal{X}$ ia SFT if $\exists n \geq 0$ and $\mathcal{W} \subset \mathcal{A}^n$ such that $\mathcal{X}$ is exactly the set of sequences in Σ that contain no words from $\mathcal{W}$.

The map $S : \mathcal{X} \to \mathcal{X}$ is the restriction of σ to $\mathcal{X}$.

A SFT $\mathcal{X}$ is mixing if and only if there exists $N \geq 1$ with matrix

$$A_{uv} = \begin{cases} 1, & \text{if } \exists x \in \mathcal{X} \text{ with } x_{n-1} = u, x_n = v \\ 0 & \text{otherwise,} \end{cases}$$

and A^N contains all positive entries.

Gibbs measure

A Gibbs measure μ is given by a potential function $f : \mathcal{X} \to \mathbb{R}$ if there exists constants $\mathcal{P} \in \mathbb{R}$ and $K > 0$ such that for all $x \in \mathcal{X}$ and $m \geq 1$

$$K^{-1} \leq \frac{\mu(x_0^{m-1})}{\exp\left(-\mathcal{P}m + \sum_{k=1}^{m-1} f(S^k(x))\right)} \leq K.$$

Gibbs measure

A Gibbs measure μ is given by a potential function $f : \mathcal{X} \to \mathbb{R}$ if there exists constants $\mathcal{P} \in \mathbb{R}$ and $K > 0$ such that for all $x \in \mathcal{X}$ and $m \geq 1$

$$K^{-1} \leq \frac{\mu(x_0^{m-1})}{\exp\left(-\mathcal{P}m + \sum_{k=1}^{m-1} f(S^k(x))\right)} \leq K.$$

Under mild conditions unique, ergodic $\mu \in \mathcal{M}(\mathcal{X}, \mathcal{S})$.

The model class

$\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, with Θ compact is a regular family if for all $\theta \in \Theta$, $x \in \mathcal{X}$, and $m \geq 1$

$$K^{-1} \leq \frac{\mu_\theta(x_0^{m-1})}{\exp\left(-\mathcal{P}(f_\theta)m + \sum_{k=1}^{m-1} f_\theta(S^k(x))\right)} \leq K.$$

Hidden SFT models

Let $\mathcal{X}$ be a mixing SFT, $\{f_\theta : \theta \in \Theta\} \subset C^r(\mathcal{X})$ a regular family of Hölder potential functions and $\{\mu_\theta : \theta \in \Theta\}$ the corresponding Gibbs measures with prior Π_o fully supported on Θ. Also let $\psi_\theta(u \mid x)$ be the observation process, with regularity.

Given the prior Π_o and marginal likelihood $p_\theta(u_0^{n-1})$ for $E \in \Theta$

$$\Pi_n(E \mid u_0^{n-1}) = \frac{\int_E p_\theta(u_0^{n-1}) d\Pi_o(\theta)}{\int_\Theta p_\theta(u_0^{n-1}) d\Pi_o(\theta)}$$

Posterior consistency

Theorem (McGoff-M-Nobel)

Let $E \subset \Theta$ be an open neighborhood of $[\Theta^*]$. Then

$$\lim_{n} \Pi_n(\Theta \setminus E \mid U_0^{n-1}) = 0, \quad \mathbb{P}_{\theta^*}^{U} - a.s.$$

Joinings and couplings

Definition (Joining)

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

Joinings and couplings

Definition (Joining)

Let (X, A, μ, T) and (Y, B, ν, S) be two dynamical systems. A joining of T and S is a probability measure λ on $X \times Y$, with marginals μ and ν 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

$\mathcal{J}(\mu, \nu)$ is the set of all joinings of $(\mathcal{X}, S, \mu)$ and $(\mathcal{Y}, T, \nu)$.

Define $\mathcal{J}(S : \nu) = \bigcup_{\mu} \mathcal{J}(\mu, \nu)$, where the union is over all S-invariant Borel probability measures $\mu \in M(\mathcal{X}, S)$.

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

Recall ν is the measure for T and $\lambda \in \mathcal{J}(S : \nu)$

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

Recall ν is the measure for T and $\lambda \in \mathcal{J}(S : \nu)$

Define $\lambda_y \in M(\mathcal{X})$ (λ "projected" onto $d\nu_y$)

$$\lambda = \int_{\mathcal{Y}} \lambda_y \otimes \delta_y \, d\nu(y).$$

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

Recall ν is the measure for T and $\lambda \in \mathcal{J}(S : \nu)$

Define $\lambda_y \in M(\mathcal{X})$ (λ “projected" onto $d\nu_y$)

$$\lambda = \int_{\mathcal{Y}} \lambda_y \otimes \delta_y \, d\nu(y).$$

Limiting average cost

$$\lim_{n \to \infty} \frac{1}{n} \int_{\mathcal{X}} c_n(x, y) \, d\lambda_y(x) = \int c \, d\lambda.$$

Variational formulation of $Z_n(y)$ – entropy term

Given two Borel probability measures π and μ on $\mathcal{X}$ and a finite measurable partition ξ of $\mathcal{X}$.
Denote $\mu \prec_\xi \pi$ as $\pi(C) = 0 \Rightarrow \mu(C) = 0$ for $C \in \xi$.

Variational formulation of $Z_n(y)$ – entropy term

Given two Borel probability measures π and μ on $\mathcal{X}$ and a finite measurable partition ξ of $\mathcal{X}$.
Denote $\mu \prec_\xi \pi$ as $\pi(C) = 0 \Rightarrow \mu(C) = 0$ for $C \in \xi$.

Define

$$L(\mu \,\|\, \pi, \xi) = \begin{cases} \sum_{C \in \xi} \mu(C) \log \pi(C), & \text{if } \mu \prec_\xi \pi \\ -\infty, & \text{otherwise,} \end{cases}$$

with $0 \cdot \log 0 = 0$.

Variational formulation of $Z_n(y)$ – entropy term

Given two Borel probability measures π and μ on $\mathcal{X}$ and a finite measurable partition ξ of $\mathcal{X}$.
Denote $\mu \prec_\xi \pi$ as $\pi(C) = 0 \Rightarrow \mu(C) = 0$ for $C \in \xi$.

Define

$$L(\mu \,\|\, \pi, \xi) = \begin{cases} \sum_{C \in \xi} \mu(C) \log \pi(C), & \text{if } \mu \prec_\xi \pi \\ -\infty, & \text{otherwise,} \end{cases}$$

with $0 \cdot \log 0 = 0$.

In spirit consider all finite measurable partitions ξ

$$F(\mu, \pi) = \sup_\xi L(\mu \,\|\, \pi, \xi).$$

Convergence

Theorem (McGoff-M.-Nobel)

Suppose a GIbbs prior, then for ν almost every y,

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

and the infimum in the above expression is attained.

Convergence

Theorem (McGoff-M.-Nobel)

Suppose a GIbbs prior, then for ν almost every y,

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

and the infimum in the above expression is attained.

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

Bayes as a variational problem

Suppose a GIbbs prior, then for ν almost every y,

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

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

$$
\begin{aligned}
P(\mu_\theta, \nu) &= \inf_{\lambda \in \mathcal{J}(S:\nu)} \left\{ \int c \, d\lambda + F(\lambda, \mu_\theta) \right\} \\
P(\theta : \nu) &= \inf_{\mu \in M(\mathcal{X}_\theta, S_\theta)} P(\mu_\theta, \nu), \\
\theta_* &= \arg\min_{\theta \in \Theta} P(\theta : \nu).
\end{aligned}
$$

Convergence

Proposition (McGoff-M.-Nobel)

Suppose a GIbbs prior and consider the pressure

$$
\begin{aligned}
P(\mu_\theta, \nu) &= \inf_{\lambda \in \mathcal{J}(S:\nu)} \left\{ \int c \, d\lambda + F(\lambda, \mu_\theta) \right\} \\
P(\theta : \nu) &= \inf_{\mu \in M(\mathcal{X}_\theta, S_\theta)} P(\mu_\theta, \nu), \\
\theta_* &= \arg\min_{\theta \in \Theta} P(\theta : \nu).
\end{aligned}
$$

For all $\varepsilon > 0$

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

Contributions

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

Provides general framework and suite of tools from the thermodynamic formalism for analyzing asymptotic behavior of Gibbs posteriors.

Questions

Statistics questions.

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

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?

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?

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?

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?

Dimension reduction

Commutative diagrams

$$
\begin{array}{ccccccc}
Y_t & & Y_{t+1} & & Y_{t+2} & & \\
\uparrow \psi & & \uparrow \psi & & \uparrow \psi & & \\
U_t & \xrightarrow{S} & U_{t+1} & \xrightarrow{S} & U_{t+2} & & \cdots\cdots \\
\downarrow \pi & & \downarrow \pi & & \downarrow \pi & & \\
V_t & \xrightarrow{T} & V_{t+1} & \xrightarrow{T} & V_{t+2} & &
\end{array}
$$

Topological conjugacy

$$
\begin{array}{ccccccc}
U_t & \xrightarrow{S} & U_{t+1} & \xrightarrow{S} & U_{t+2} & & \cdots\cdots \\
\downarrow \pi & & \downarrow \pi & & \downarrow \pi & & \\
V_t & \xrightarrow{T} & V_{t+1} & \xrightarrow{T} & V_{t+2} & & \cdots\cdots
\end{array}
$$

Topological conjugacy

$$
\begin{array}{ccccccc}
U_t & \xrightarrow{S} & U_{t+1} & \xrightarrow{S} & U_{t+2} & & \cdots\cdots \\
\downarrow \pi & & \downarrow \pi & & \downarrow \pi & & \\
V_t & \xrightarrow{T} & V_{t+1} & \xrightarrow{T} & V_{t+2} & & \cdots\cdots
\end{array}
$$

Topological conjugacy: Two functions $S : U \mapsto U$ and $T : V \mapsto V$ are *topologically conjugate* if there exists a homeomorphism $\pi : U \mapsto V$ such that

$$\pi \circ S = T \circ \pi.$$

Factors

$$
\begin{array}{ccccccc}
U_t & \xrightarrow{S} & U_{t+1} & \xrightarrow{S} & U_{t+2} & & \cdots\cdots \\
\downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi} & & \\
V_t & \xrightarrow{T} & V_{t+1} & \xrightarrow{T} & V_{t+2} & & \cdots\cdots
\end{array}
$$

Factors

$$
\begin{array}{ccccccc}
U_t & \xrightarrow{S} & U_{t+1} & \xrightarrow{S} & U_{t+2} & & \cdots\cdots \\
\downarrow \pi & & \downarrow \pi & & \downarrow \pi & & \\
V_t & \xrightarrow{T} & V_{t+1} & \xrightarrow{T} & V_{t+2} & & \cdots\cdots
\end{array}
$$

Factors: Given two dynamical systems $(U, \mathcal{U}, \mu, T)$ and $(V, \mathcal{V}, \nu, S)$ with a map $\pi : U \mapsto V$ if

1. The map π is measurable.
2. For each $V \in \mathcal{V}$, $\mu(\pi^{-1}B) = \nu(B)$.
3. For μ-almost all $u \in U$, $\pi(Tx) = S(\pi x)$,

Then $(V, \mathcal{V}, \nu, S)$ is a factor of $(U, \mathcal{U}, \mu, T)$.

An objective

The factor suggests minimize either the difference in conditional probabilities

$$\min_{\pi^*, T} \text{KL}(U_n \mid U_{n-1} \,\|\, \pi^*(V_n \mid V_{n-1}))$$

or the one step error

$$\min_{\pi^*, T} \mathbb{E}\| U_n - \pi^*(S(V_{n-1})\|.$$

Partial factor model

Consider the the following linear regression setting for the dynamics

$$Y_i = B^T X_i + E_i, \quad E_i \overset{iid}{\sim} \mathrm{N}(0, \Gamma)$$

with $Y_i = X_{i+1}$.

Partial factor model

Consider the the following linear regression setting for the dynamics

$$Y_i = B^T X_i + E_i, \quad E_i \overset{iid}{\sim} \mathrm{N}(0, \Gamma)$$

with $Y_i = X_{i+1}$.
With a factor model on X

$$X_i = Af_i + \nu_i, \quad \nu_i \overset{iid}{\sim} \mathrm{N}(o, \Psi)$$

$$f_i \sim \mathrm{N}(0, I).$$

Partial factor model

Consider the the following linear regression setting for the dynamics

$$Y_i = B^T X_i + E_i, \quad E_i \overset{iid}{\sim} \mathrm{N}(0, \Gamma)$$

with $Y_i = X_{i+1}$.
With a factor model on X

$$X_i = Af_i + \nu_i, \quad \nu_i \overset{iid}{\sim} \mathrm{N}(o, \Psi)$$

$$f_i \sim \mathrm{N}(0, I).$$

The statistical problem is to learn $A := \pi^*$ and $B = T$.

The key idea

Typical joint distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}(0, \Sigma),$$

with

$$\Sigma = \begin{bmatrix} AA^T + \Gamma & BA^T \\ AB^T & \Psi + BB^T \end{bmatrix}.$$

The key idea

Instead model

$$\begin{pmatrix} X \\ f \\ Y \end{pmatrix} \sim \mathrm{N}(0, \Sigma),$$

with

$$\Sigma = \begin{bmatrix} AA^T + \Gamma & A^T & BA^T \\ A & I_k & B^T \\ AB^T & B & \Psi + BB^T \end{bmatrix}.$$

We can apply some standard Bayesian models to infer A, B.

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.

(3) Extension to continuous time dynamics, differential equations.

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.

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.

Acknowledgements

Thanks:
Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger.
Funding:

- Center for Systems Biology at Duke
- NSF DMS, CCF, CISE, DEB
- AFOSR
- DARPA
- NIH