Approximate Bayesian Computation with Path Signatures

Joel Dyer
Mathematical Institute & Institute for New Economic Thinking
University of Oxford

One World ABC Seminar
February 2, 2023

Oxford Mathematics

University of Oxford
Mathematical Institute

Introduction

Aim: To convince you that path signatures can be a useful tool when applying approximate Bayesian computation to time-series simulators of different kinds.

Approximate Bayesian Computation with Path Signatures
Dyer, J.; Cannon, P.; and Schmon, S. M.
arXiv:2106.12555 (2023)

(Just updated!)

Context & notation

Inference for dynamic, stochastic simulation models

Context & notation

Inference for dynamic, stochastic simulation models

- Models that are defined by an underlying computer program

Context & notation

Inference for dynamic, stochastic simulation models

- Models that are defined by an underlying computer program
- Forward simulation easy: $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta})$ for time-series $\mathbf{x}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$

Context & notation

Inference for dynamic, stochastic simulation models

- Models that are defined by an underlying computer program
- Forward simulation easy: $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta})$ for time-series $\mathbf{x}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$
- Evaluating $p(\mathbf{x} \mid \boldsymbol{\theta})$ hard/expensive for any $(\mathbf{x}, \boldsymbol{\theta})$ combination

Context & notation

Inference for dynamic, stochastic simulation models

- Models that are defined by an underlying computer program
- Forward simulation easy: $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta})$ for time-series $\mathbf{x}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$
- Evaluating $p(\mathbf{x} \mid \boldsymbol{\theta})$ hard/expensive for any $(\mathbf{x}, \boldsymbol{\theta})$ combination
- Consequence: inference on $\boldsymbol{\theta}$ is tricky – many inference procedures require access to $p(\mathbf{x} \mid \boldsymbol{\theta})$ e.g. Bayes

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})}{p(\mathbf{x})} \pi(\boldsymbol{\theta}) \tag{1}$$

Context & notation

Inference for dynamic, stochastic simulation models

Mathematical Institute

- Models that are defined by an underlying computer program
- Forward simulation easy: $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta})$ for time-series $\mathbf{x}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$
- Evaluating $p(\mathbf{x} \mid \boldsymbol{\theta})$ hard/expensive for any $(\mathbf{x}, \boldsymbol{\theta})$ combination
- Consequence: inference on $\boldsymbol{\theta}$ is tricky – many inference procedures require access to $p(\mathbf{x} \mid \boldsymbol{\theta})$ e.g. Bayes

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})}{p(\mathbf{x})} \pi(\boldsymbol{\theta}) \tag{1}$$

- Resort to simulation-based inference, such as approximate Bayesian computation (ABC): with kernel function K_ϵ, bandwidth parameter $\epsilon > 0$, and summary statistics $\mathbf{s}$ s.t. $\mathbf{s}_\mathbf{x} := \mathbf{s}(\mathbf{x})$,

$$\pi_{\mathrm{ABC},\epsilon}(\boldsymbol{\theta} \mid \mathbf{s}_\mathbf{x}) \propto \int K_\epsilon(\mathbf{s}_\mathbf{x}, \mathbf{s}_{\tilde{\mathbf{x}}})\, p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\tilde{\mathbf{x}} \tag{2}$$

Challenge

Choosing appropriate distances/summary statistics for time-series data

Challenge

Choosing appropriate distances/summary statistics for time-series data

Mathematical Institute

- Kernel K_ϵ typically uses some distance $\mathcal{D}$ internally, e.g.

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \mathbb{I}\left[\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \leq \epsilon\right], \qquad \text{(Uniform kernel)}$$

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \exp\left\{-\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}})/\epsilon\right\} \qquad \text{(Gaussian kernel)}$$

$$\ldots$$

Challenge

Choosing appropriate distances/summary statistics for time-series data

Mathematical Institute

- Kernel K_ϵ typically uses some distance $\mathcal{D}$ internally, e.g.

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \mathbb{I}\left[\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \leq \epsilon\right], \qquad \text{(Uniform kernel)}$$

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \exp\left\{-\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}})/\epsilon\right\} \qquad \text{(Gaussian kernel)}$$

$$\ldots$$

- What $\mathcal{D}$ and/or $\mathbf{s}$ to use for different time-series data?

Challenge

Choosing appropriate distances/summary statistics for time-series data

- Kernel K_ϵ typically uses some distance $\mathcal{D}$ internally, e.g.

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \mathbb{I}\left[\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \leq \epsilon\right], \qquad \text{(Uniform kernel)}$$

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \exp\left\{-\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}})/\epsilon\right\} \qquad \text{(Gaussian kernel)}$$

$$\ldots$$

- What $\mathcal{D}$ and/or $\mathbf{s}$ to use for different time-series data?

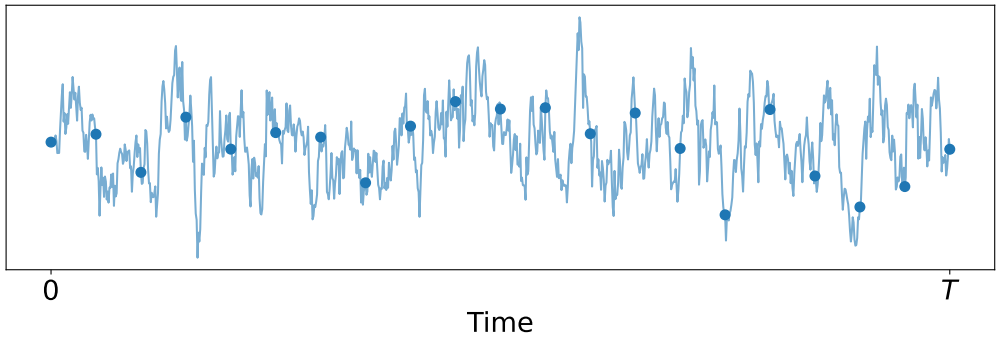


Figure: Univariate time-series

Challenge

Choosing appropriate distances/summary statistics for time-series data

- Kernel K_ϵ typically uses some distance $\mathcal{D}$ internally, e.g.

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \mathbb{I}\left[\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \leq \epsilon\right], \qquad \text{(Uniform kernel)}$$

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \exp\left\{-\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}})/\epsilon\right\} \qquad \text{(Gaussian kernel)}$$

$$\ldots$$

- What $\mathcal{D}$ and/or $\mathbf{s}$ to use for different time-series data?

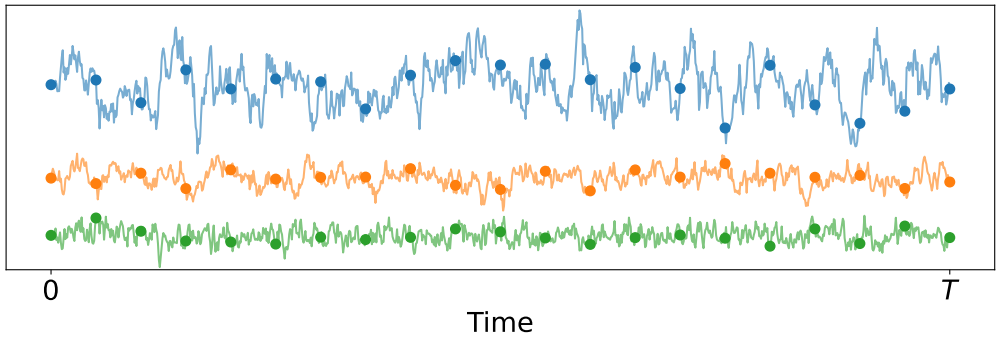


Figure: Multivariate time-series

Challenge

Choosing appropriate distances/summary statistics for time-series data

- Kernel K_ϵ typically uses some distance $\mathcal{D}$ internally, e.g.

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \mathbb{I}\left[\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \leq \epsilon\right], \qquad \text{(Uniform kernel)}$$

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \exp\left\{-\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}})/\epsilon\right\} \qquad \text{(Gaussian kernel)}$$

$$\ldots$$

- What $\mathcal{D}$ and/or $\mathbf{s}$ to use for different time-series data?

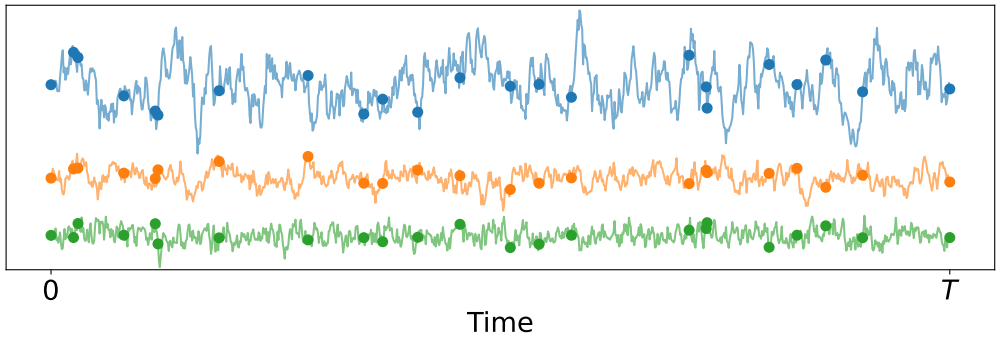


Figure: Irregularly sampled, multivariate time-series

Challenge

Choosing appropriate distances/summary statistics for time-series data

- Kernel K_ϵ typically uses some distance $\mathcal{D}$ internally, e.g.

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \mathbb{I}\left[\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \leq \epsilon\right], \qquad \text{(Uniform kernel)}$$

$$K_\epsilon(\mathbf{s_x}, \mathbf{s_{\tilde{x}}}) \propto \exp\left\{-\mathcal{D}(\mathbf{s_x}, \mathbf{s_{\tilde{x}}})/\epsilon\right\} \qquad \text{(Gaussian kernel)}$$

$$\ldots$$

- What $\mathcal{D}$ and/or $\mathbf{s}$ to use for different time-series data?

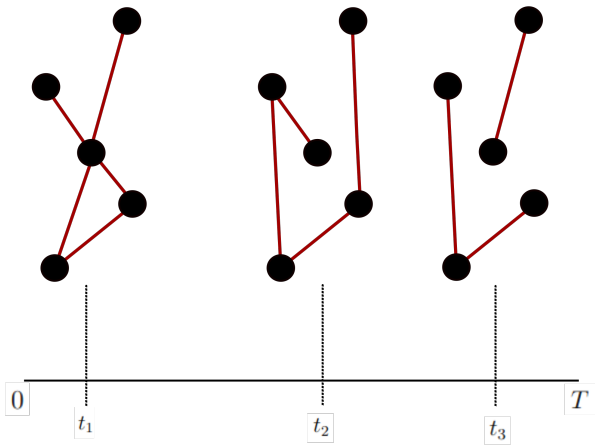


Figure: Irregularly spaced, non-Euclidean time-series

Existing approaches to choosing distance/summary statistics

Mathematical Institute

Prototypical approach: Euclidean distance between hand-crafted summary statistics $\mathbf{s}$.

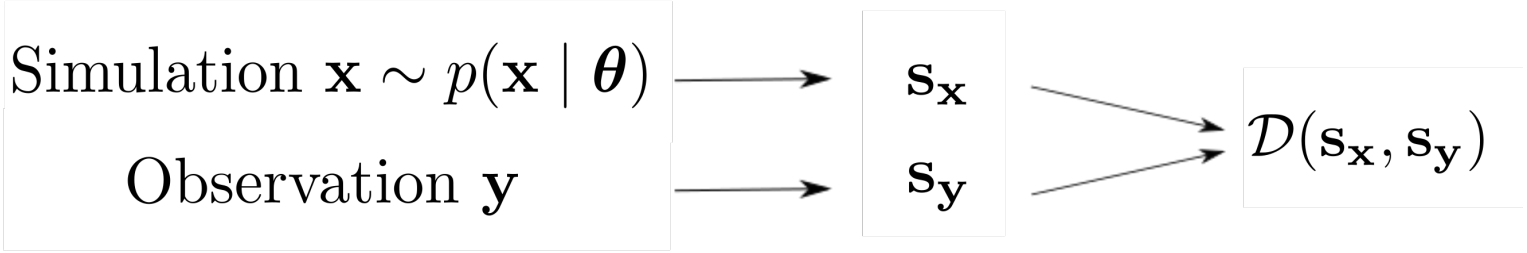


Figure: Schematic of manual method for defining "distance".

Ad-hoc derivation of hand-crafted summary statistics can be time-consuming, expensive, risky...

Existing approaches to choosing distance/summary statistics

Semi-automatic ABC (Fearnhead and Prangle, 2012)

- Use $\mathbf{s_x} = \mathbb{E}_\pi[\boldsymbol{\theta} \mid \mathbf{x}]$ in ABC$\rightarrow$ minimises quadratic loss between "true" parameters $\boldsymbol{\theta}^*$ and

$$\hat{\boldsymbol{\theta}}_{\mathrm{ABC}} = \mathbb{E}_{\pi_{\mathrm{ABC},\epsilon}}[\boldsymbol{\theta} \mid \mathbf{s_y}]$$

- Estimate $\mathbf{s_x}$ by learning nonlinear map from $\mathbf{x}_i$ to $\boldsymbol{\theta}_i$, where $(\mathbf{x}_i, \boldsymbol{\theta}_i)_{i=1}^n \sim p(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is simulated training data. Then use

$$\mathcal{D}(\mathbf{s_x}, \mathbf{s_y}) = \|\mathbf{s_x} - \mathbf{s_y}\|_2$$

- Drawback: need to define suitable space of functions over which to search. What function space is appropriate/useful for time-series of different kinds?

Existing approaches to choosing distance/summary statistics

K2-ABC: ABC with kernel embeddings (Park et al., 2016)

- Compare empirical distribution of simulator and real data via **MMD**
- For sufficiently expressive kernel $k(X, \cdot)$, e.g. RBF, kernel mean embedding encodes all information of the distribution

$$\mu_X := \mathbb{E}[k(X, \cdot)], \qquad \hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \cdot), \qquad \mathcal{D}(\mathbf{x}, \mathbf{y}) = \mathrm{MMD}^2.$$

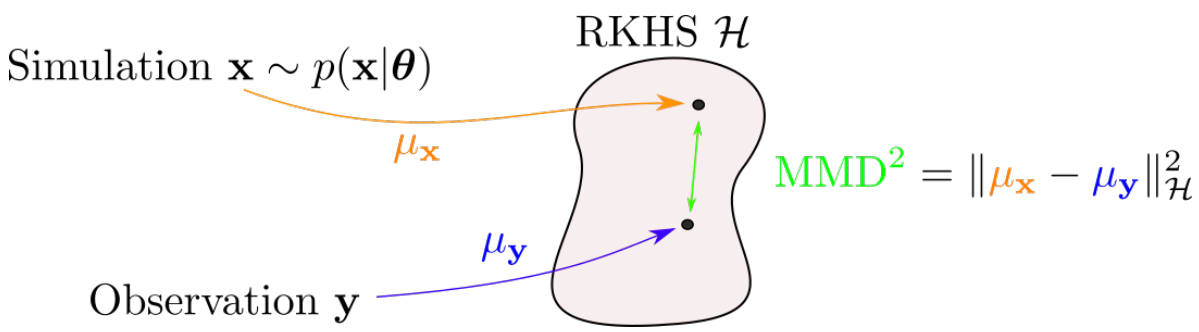


Figure: Schematic of MMD. Treats the $\mathbf{x}_i$ in $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ as exchangeable.

Existing approaches to choosing distance/summary statistics

ABC with the Wasserstein distance (Bernton et al., 2019)

Compare simulated and real time-series, $\mathbf{x}$ and $\mathbf{y}$, via the **Wasserstein distance**:

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = W_p(\mathbf{x}, \mathbf{y}) := \min_{\sigma \in \mathcal{S}_n} \sum_{i=1}^{n} \rho(\mathbf{x}_i, \mathbf{y}_{\sigma(i)})^p,$$

$$\rho(\mathbf{x}_t, \mathbf{y}_s) = \|\mathbf{x}_t - \mathbf{y}_s\| + \lambda|t - s|$$

- $\mathcal{S}_n$: set of permutations of $1, \ldots, n$
- λ: hyperparameter balancing "vertical" and "horizontal" transport

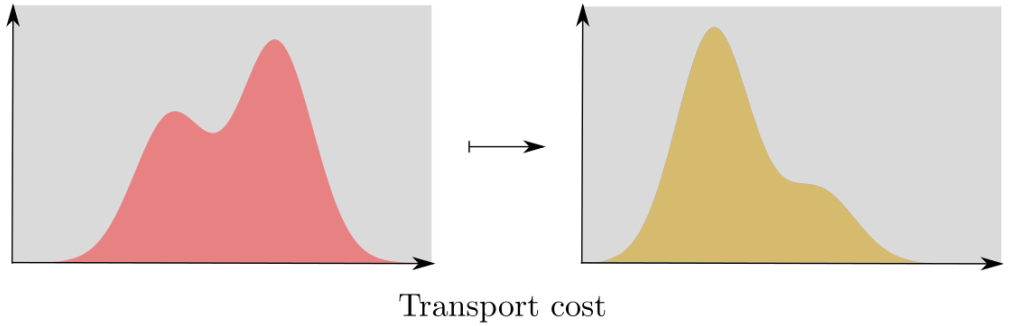


Figure: Schematic of Wasserstein distance. Sort of accounts for dependencies.

Existing approaches to choosing distance/summary statistics

Other approaches in this vein...

- ABC with the Kullback-Liebler divergence (Jiang, 2018)
- ABC with the energy statistic (Nguyen et al., 2020)
- ...

See Drovandi and Frazier (2022) for a recent review and comparison of all of the above.

These approaches are primarily designed for independent – rather than time-series-like – data. How can we deal with time-series?

Our contribution: ABC with path signatures

We propose (semi-)automatic approaches to ABC for time-series models of different kinds using **path signatures**.

Remainder of this talk:

- Overview of path signatures & properties
- Our proposed signature-based methods & some theoretical properties
- Experimental results
- Summary & conclusion

Let $h : [0, T] \to \mathcal{H}$ be a path in Hilbert space $\mathcal{H}$ with bounded variation:

$$\|h\|_{1-\mathrm{var}} := \sup_{\zeta(0,T)} \sum_{i=1}^{n-1} \|h_{t_{i+1}} - h_{t_i}\|_{\mathcal{H}} < \infty.$$

The signature map sends h to an infinite collection of tensors of increasing order:

$$\mathrm{Sig} : h \mapsto (1, S_1(h), S_2(h), \dots) \in \prod_{m \geq 0} \mathcal{H}^{\otimes m}$$

where

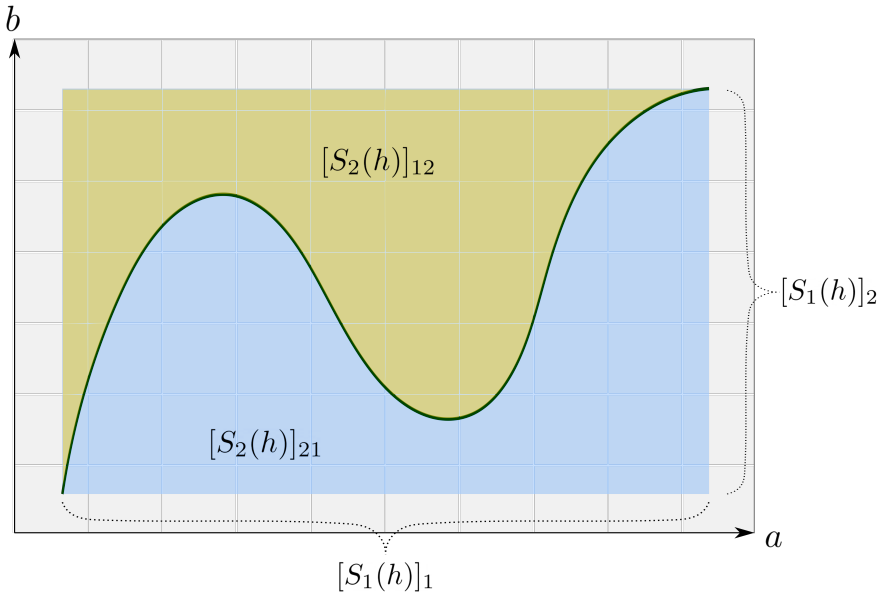
$$S_m(h) = \int_{t_m=0}^{T} \cdots \int_{t_1=0}^{t_2} \mathrm{d}h_{t_1} \otimes \cdots \otimes \mathrm{d}h_{t_m} \tag{3}$$

Path signatures

Example 2.3, Király and Oberhauser (2019)

Consider $h_t = (a_t, b_t) \in \mathbb{R}^2$.

$$S_1(h) = \begin{bmatrix} \int_0^T \mathrm{d}a_t \\ \int_0^T \mathrm{d}b_t \end{bmatrix} \quad \text{and} \quad S_2(h) = \begin{bmatrix} \int_0^T \int_0^{t_2} \mathrm{d}a_{t_1} \mathrm{d}a_{t_2} & \int_0^T \int_0^{t_2} \mathrm{d}a_{t_1} \mathrm{d}b_{t_2} \\ \int_0^T \int_0^{t_2} \mathrm{d}b_{t_1} \mathrm{d}a_{t_2} & \int_0^T \int_0^{t_2} \mathrm{d}b_{t_1} \mathrm{d}b_{t_2} \end{bmatrix}.$$



Path signatures

Where have they come from?

Path signatures

Where have they come from?

Mathematical Institute

They appear in solutions to controlled/stochastic differential equations

Path signatures

Where have they come from?

Mathematical Institute

They appear in solutions to controlled/stochastic differential equations, e.g. for Brownian motion B_t and linear operator A,

$$\mathrm{d}Y_t = A(Y_t) \circ \mathrm{d}B_t, \quad Y_0 = y$$

has a solution of the form (Lyons et al., 2007)

$$Y_t = \sum_{m \geq 0} A^{\otimes m} S_{m,[0,t]}(B)\, y.$$

They appear in solutions to controlled/stochastic differential equations, e.g. for Brownian motion B_t and linear operator A,

$$\mathrm{d}Y_t = A(Y_t) \circ \mathrm{d}B_t, \quad Y_0 = y$$

has a solution of the form (Lyons et al., 2007)

$$Y_t = \sum_{m \geq 0} A^{\otimes m} S_{m,[0,t]}(B)\, y.$$

Lyons's *rough path theory* (see e.g. Lyons et al., 2002, 2007; Friz and Hairer, 2020) built around the closely related notion of a multiplicative functional.

Path signatures

Key properties

Mathematical Institute

Uniqueness: For injective paths h, g with common origin, $\mathrm{Sig}(h) = \mathrm{Sig}(g)$ iff $h = g$. (Enforcing these conditions always possible.)

Path signatures

Key properties

Uniqueness: For injective paths h, g with common origin, $\mathrm{Sig}(h) = \mathrm{Sig}(g)$ iff $h = g$. (Enforcing these conditions always possible.)

Universal nonlinearity: Linear functionals on signatures are dense in the space of continuous, real-valued functions on compact sets of paths.

Path signatures

Key properties

Uniqueness: For injective paths h, g with common origin, $\mathrm{Sig}(h) = \mathrm{Sig}(g)$ iff $h = g$. (Enforcing these conditions always possible.)

Universal nonlinearity: Linear functionals on signatures are dense in the space of continuous, real-valued functions on compact sets of paths.

Kernelisation: The *signature kernel* for bounded variation paths $h, g : [0, T] \to \mathcal{H}$ is

$$k(h, g) = \langle \mathrm{Sig}(h), \mathrm{Sig}(g) \rangle := \sum_{m \geq 0} \langle S_m(h), S_m(g) \rangle_{\mathcal{H}^{\otimes m}} .$$

Király and Oberhauser (2019): $k(h, g)$ can be evaluated using only inner products on points in the path. Király and Oberhauser (2019) and Salvi et al. (2020) provide efficient methods for evaluating k.

Path signatures

From sequences in data space to paths in Hilbert space

Real scenario: observe sequence $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}^n$ at times $0 = t_1 < \cdots < t_n = T$. How to map to path in $\mathcal{H}$?

1. **Embed data in a Hilbert space**: Through canonical feature map for kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with corresponding RKHS $\mathcal{H}$.
2. **Obtain a $\mathcal{H}$-valued path**: Through interpolation (e.g. linear).

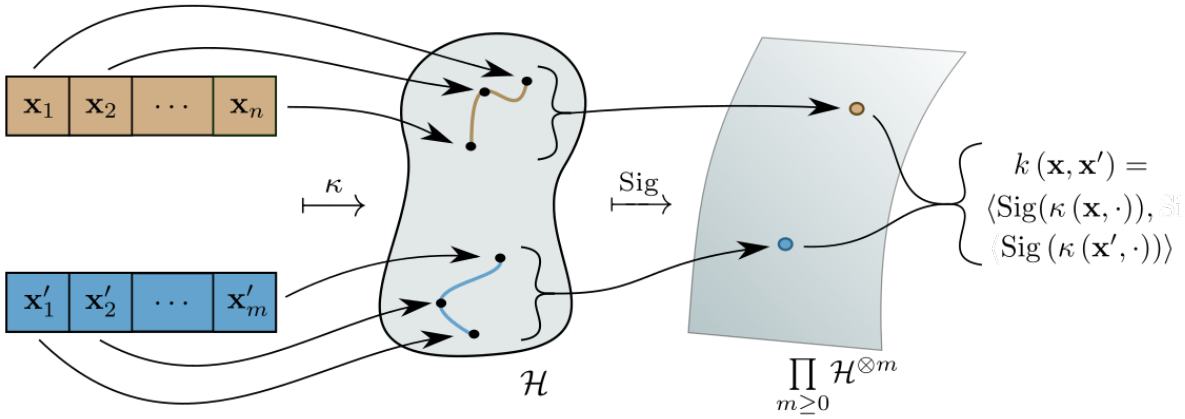


Figure: Embedding sequences with the signature kernel (Dyer et al., 2022).

Mathematical Institute

Interpret signature as summary statistic: $\mathbf{s}(\mathbf{x}) = \mathrm{Sig}(\mathbf{x}) \Rightarrow$ a sufficient statistic!

Mathematical Institute

Interpret signature as summary statistic: $\mathbf{s}(\mathbf{x}) = \mathrm{Sig}(\mathbf{x}) \Rightarrow$ a sufficient statistic!

Distance computation can be kernelised with signature kernel k:

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) := \|\mathrm{Sig}(\mathbf{x}) - \mathrm{Sig}(\mathbf{y})\|^2 = k(\mathbf{x}, \mathbf{x}) + k(\mathbf{y}, \mathbf{y}) - 2k(\mathbf{x}, \mathbf{y}) \tag{4}$$

Mathematical Institute

Interpret signature as summary statistic: $\mathbf{s}(\mathbf{x}) = \mathrm{Sig}(\mathbf{x}) \Rightarrow$ a sufficient statistic!

Distance computation can be kernelised with signature kernel k:

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) := \|\mathrm{Sig}(\mathbf{x}) - \mathrm{Sig}(\mathbf{y})\|^2 = k(\mathbf{x}, \mathbf{x}) + k(\mathbf{y}, \mathbf{y}) - 2k(\mathbf{x}, \mathbf{y}) \tag{4}$$

Convergence of Signature ABC posterior

Let $\mathcal{X} = \mathbb{R}^m$, $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$ be length-n sequences in $\mathcal{X}$. Under the regularity conditions on $p(\mathbf{x} \mid \boldsymbol{\theta})$ assumed in Bernton et al. (2019), and with $\mathcal{D}$ as in Equation (4) + some additional benign conditions on κ, the ABC posterior

$$\pi_{\mathrm{ABC},\epsilon}(\boldsymbol{\theta} \mid \mathrm{Sig}(\mathbf{y})) \to \pi(\boldsymbol{\theta} \mid \mathbf{y}) \quad \text{as} \quad \epsilon \to 0.$$

Mathematical Institute

Interpret signature as summary statistic: $\mathbf{s}(\mathbf{x}) = \mathrm{Sig}(\mathbf{x}) \Rightarrow$ a sufficient statistic!

Distance computation can be kernelised with signature kernel k:

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) := \|\mathrm{Sig}(\mathbf{x}) - \mathrm{Sig}(\mathbf{y})\|^2 = k(\mathbf{x}, \mathbf{x}) + k(\mathbf{y}, \mathbf{y}) - 2k(\mathbf{x}, \mathbf{y}) \tag{4}$$

Convergence of Signature ABC posterior

Let $\mathcal{X} = \mathbb{R}^m$, $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$ be length-n sequences in $\mathcal{X}$. Under the regularity conditions on $p(\mathbf{x} \mid \boldsymbol{\theta})$ assumed in Bernton et al. (2019), and with $\mathcal{D}$ as in Equation (4) + some additional benign conditions on κ, the ABC posterior

$$\pi_{\mathrm{ABC},\epsilon}(\boldsymbol{\theta} \mid \mathrm{Sig}(\mathbf{y})) \to \pi(\boldsymbol{\theta} \mid \mathbf{y}) \quad \text{as} \quad \epsilon \to 0.$$

Behaviour as $n \to \infty$ for fixed ϵ

For a continuous-time process observed over fixed time interval $[0, T]$,

$$\pi_{\mathrm{ABC},\epsilon}(\boldsymbol{\theta} \mid \mathrm{Sig}(\mathbf{y})) \rightharpoonup \pi_{\mathrm{ABC},\epsilon}(\boldsymbol{\theta} \mid \mathrm{Sig}(h)) \quad \text{as} \quad n \to \infty.$$

Exploit "universal nonlinearity" of Sig to perform **semi-automatic** ABC for time-series simulators.

What space of functions to optimise over? Linear functionals on signatures – signature kernel is universal for sequences (Király and Oberhauser, 2019).

Exploit "universal nonlinearity" of Sig to perform **semi-automatic** ABC for time-series simulators.

What space of functions to optimise over? Linear functionals on signatures – signature kernel is universal for sequences (Király and Oberhauser, 2019).

E.g. kernel ridge regression with signature kernel k, s.t.

$$\hat{\boldsymbol{\theta}}(\mathbf{x}) := \left(\sum_{r=1}^{R} a_{1,r} k(\mathbf{x}, \mathbf{x}^{(r)}), \ldots, \sum_{r=1}^{R} a_{d,r} k(\mathbf{x}, \mathbf{x}^{(r)}) \right)'$$

and

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \|\hat{\boldsymbol{\theta}}(\mathbf{x}) - \hat{\boldsymbol{\theta}}(\mathbf{y})\|_2$$

Experimental results

Ricker model (Wood, 2010)

$$\log N_{t+1} = \log r + \log N_t - N_t + \sigma\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0,1)$$
$$\mathbf{y}_t \sim \mathrm{Po}\left(\phi N_t\right).$$

Aim: approximate posterior for $\boldsymbol{\theta} = (\log r, \phi, \sigma)$ with pseudo-data $\mathbf{y} \sim p(\mathbf{y} \mid \boldsymbol{\theta}^*)$, $\boldsymbol{\theta}^* = (4, 10, 0.3)$ and priors

$$\log r \sim \mathcal{U}(3,8), \qquad \phi \sim \mathcal{U}(0,20), \qquad \sigma \sim \mathcal{U}(0,0.6).$$

Experimental results

Ricker model (Wood, 2010)

$$\text{“Delay”} : (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \mapsto ((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_2, \mathbf{x}_3), \ldots, (\mathbf{x}_{n-1}, \mathbf{x}_n))$$

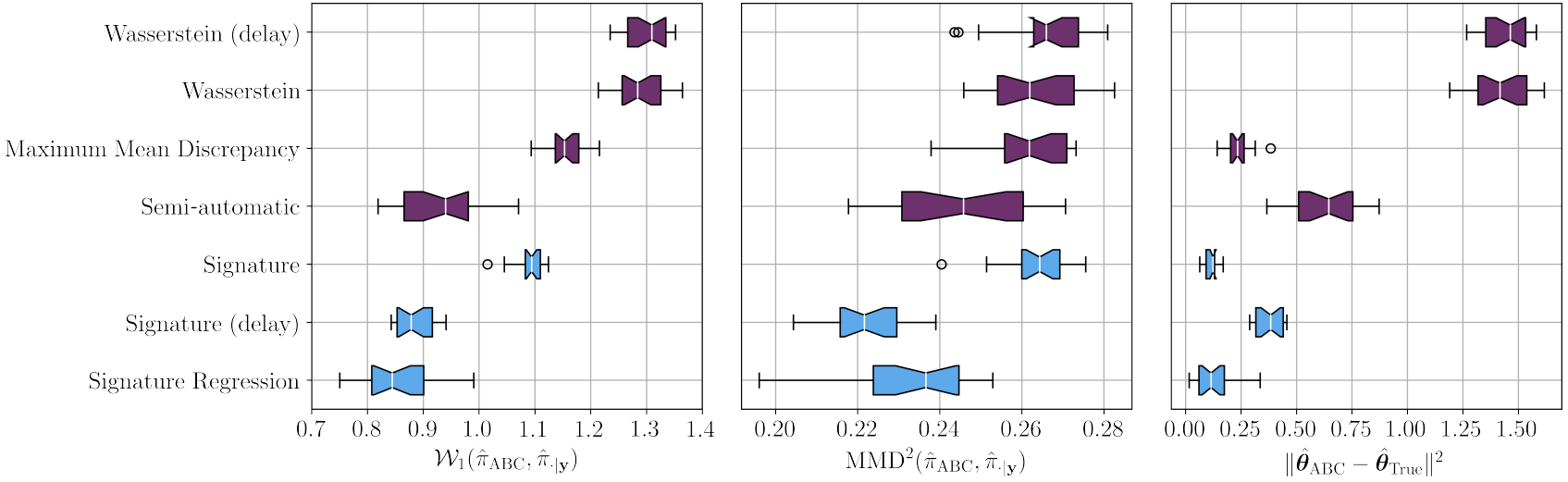


Figure: (Ricker model) Wasserstein distance (**left**) and MMD (**middle**) between ABC and ground-truth posteriors. **Right**: Euclidean distances between ABC and ground truth posterior means.

Experimental results

Results for multivariate, irregularly spaced sequences of variable length: Kypraios (2007)

Mathematical Institute

- Stochastic model of infections by and recovery from virus in population, determine by parameters β and γ
- Event-based simulation: events occur with time-dependent rate in discrete population over time interval $[0, T]$
- Output: multivariate time-series $\mathbf{y}$ of variable length and with irregularly spaced observations counting number of infected and susceptible individuals at each time
- Posterior available in closed form for Gamma priors on β and γ
- Perform inference on $\boldsymbol{\theta} = (\beta, \gamma)$ given pseudo-true data $\mathbf{y}$

Experimental results

Results for multivariate, irregularly spaced sequences of variable length: Kypraios (2007)

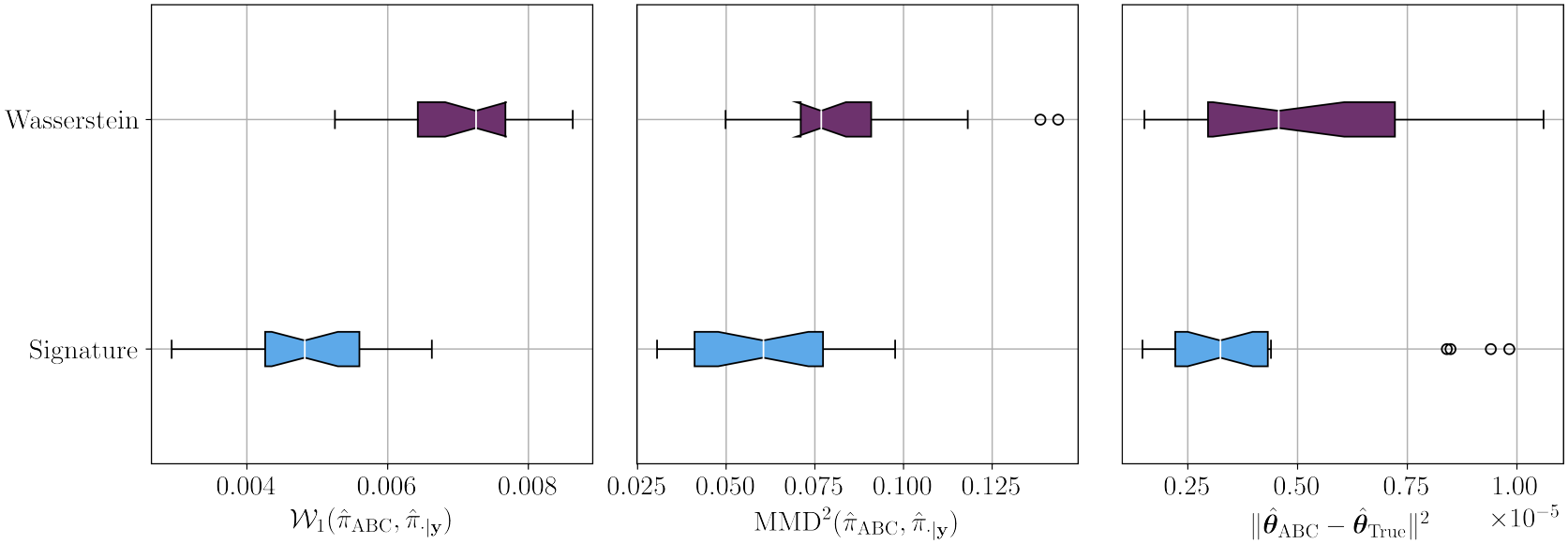


Figure: (Generalised stochastic epidemic model) **Left**: Wasserstein distance between ABC and ground-truth posteriors. **Middle**: MMD between ABC and ground-truth posteriors. **Right**: Euclidean distances between ABC and ground truth posterior means.

Kernelisation lets us operate on sequences of arbitrary objects. Example: **graphs** (e.g. simulating social networks).

Signature/Signature Regression ABC performs the following steps implicitly:

1. Send observations to RKHS $\mathcal{H}$ of graph kernel κ (e.g. Weisfeiler-Lehman)
2. Compute signature kernel for these $\mathcal{H}$-valued, piecewise-linear paths

Experimental results

Results for dynamic graph model: Zhang et al. (2017)

Model:

- If an edge absent at time $t-1$, they appear (resp. remain absent) with probability φ (resp. $1-\varphi$);
- If edge present at time $t-1$, they disappear (resp. remain present) with probability τ (resp. $1-\tau$).

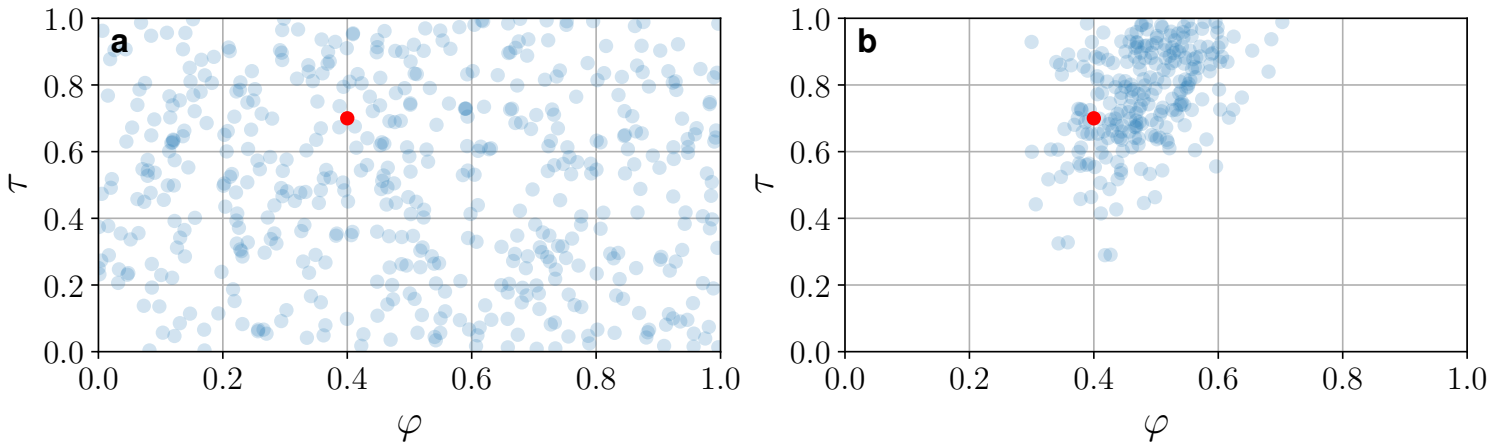


Figure: **a**: Prior. **b**: Signature ABC posterior. **Red point**: true parameters.

Computational complexity & cost

Mathematical Institute

Complexity:

- Signature kernel: $\mathcal{O}(n^2)$
- MMD: $\mathcal{O}(n^2)$
- Wasserstein: $\mathcal{O}(n^3)$ in multivariate settings with Hungarian algorithm

Observed cost: signature methods generally more expensive in our experiments (with some exceptions)

- Possible to reduce cost with e.g. GPUs, truncated signature kernels etc.
- Possible improvements to implementation

Summary & conclusion

- ABC for dynamic, stochastic simulation models can be challenging: difficult to construct summary statistics or distance measure
- Argued for the use of path signatures in ABC for generic time-series simulators
- Demonstrated empirical performance on a range of dynamic, stochastic simulation models

Thank you!

Twitter: `@joelnmdyer`
Website: https://joelnmdyer.github.io
Email: `joel.dyer@maths.ox.ac.uk`

References I

Espen Bernton, Pierre E. Jacob, Mathieu Gerber, and Christian P. Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 81(2):235–269, 2019. ISSN 14679868. doi: 10.1111/rssb.12312.

Christopher Drovandi and David T Frazier. A comparison of likelihood-free methods with and without summary statistics. *Statistics and Computing*, 32(3):42, 2022.

References II

Joel Dyer, Patrick W. Cannon, and Sebastian M. Schmon. Amortised likelihood-free inference for expensive time-series simulators with signatured ratio estimation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 11131–11144. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/dyer22a.html.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 74(3):419–474, 2012. ISSN 13697412. doi: 10.1111/j.1467-9868.2011.01010.x.

Peter K Friz and Martin Hairer. *A course on rough paths*. Springer, 2020.

References III

Bai Jiang. Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International conference on artificial intelligence and statistics*, pages 1711–1721. PMLR, 2018.

Franz J Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31): 1–45, 2019.

Theo Kypraios. Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models. 2007.

T. J Lyons, Michael Caruana, and Thierry Lévy. *Differential equations driven by rough paths : École d'été de probabilités de Saint-Flour XXXIV-2004 [electronic resource]*. Lecture notes in mathematics (Springer-Verlag) ; 1908. Springer, Berlin ; New York, 2007. ISBN 9783540712855.

References IV

Terry Lyons, Zhongmin Qian, et al. *System control and rough paths*. Oxford University Press, 2002.

Hien Duy Nguyen, Julyan Arbel, Hongliang Lü, and Florence Forbes. Approximate Bayesian computation via the energy statistic. *IEEE Access*, 8:131683–131698, 2020.

Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, 41:398–407, 2016.

Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The signature kernel is the solution of a goursat pde. *arXiv preprint arXiv:2006.14794*, 2020.

Simon N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010. ISSN 00280836. doi: 10.1038/nature09319.

References V

Xiao Zhang, Cristopher Moore, and Mark EJ Newman. Random graph models for dynamic networks. *The European Physical Journal B*, 90(10):1–14, 2017.