Approximate Bayesian Computation for Panel Data with Signature Maximum Mean Discrepancies

Joel Dyer[1,2,3] **John Fitzgerald**[2] **Bastian Rieck**[4,5,6] **Sebastian M. Schmon**[1,7]

[1]Improbable
[2]Mathematical Institute, University of Oxford
[3]Alan Turing Institute
[4]AIDOS Lab, Institute of AI for Health, Helmholtz Munich
[5]Helmholtz Pioneer Campus, Helmholtz Munich
[6]Technical University Munich
[7]Durham University

Abstract

Simulation models are becoming a staple tool across application domains from economics to biology. When such models are stochastic, evaluating their likelihood functions in a reasonable time is typically infeasible or even impossible. In these settings, simulation-based inference (SBI) procedures are a convenient means to approximating conventional parameter calibration procedures. A popular example is approximate Bayesian computation (ABC), in which the observed data is compared to the simulation output at different parameter values through some distance function. While many such methods exist, few are compatible with panel data of various kinds, as might appear in medical settings, for example; many methods instead assume *iid* observations in both the simulated and observed data. We seek to address this gap through the use of signature maximum mean discrepancies as distance measures in ABC. Through experiments with a dynamical model of functional brain networks, we demonstrate that such an approach can flexibly operate on panel data of various kinds, for example dynamic graph data arising from multiple patients/subjects in fMRI settings.

1 Introduction

Computer simulations are a key component in many areas of scientific research and in numerous real-world decision-making processes. Such programs are implemented to model the behaviour of some real-world system, and often entail the use of random number generators internally to capture stochastic influences. Performing a forward simulation of the model to generate data $\mathbf{x}$ may thus be represented as a draw from the model's likelihood function – $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta})$ – and the problem of *model calibration* – discovering values for the model's free parameters $\boldsymbol{\theta}$ that result in model behaviour that closely matches real-world data $\mathbf{y}$ – can be appropriately performed using likelihood-based parameter inference procedures.

In general, however, the likelihood for arbitrary simulation models is not available in closed form, and it is also typically not possible to evaluate the likelihood function numerically within a reasonable time. This exacerbates the application of classical likelihood-based parameter inference procedures and motivates the development of alternative techniques, from which a class of *simulation-based inference (*SBI*)* procedures that approximate standard, exact inference methods have emerged (see e.g. 4, for a recent survey).

NeurIPS 2022 Temporal Graph Learning Workshop held in New Orleans, United States

One prevailing approach to SBI is approximate Bayesian computation (ABC) (1; 13; 17). Given a prior density $\pi(\boldsymbol{\theta})$ on parameters, ABC approximates the exact parameter posterior given by Bayes' theorem,

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{1}$$

by, broadly speaking, assigning higher posterior density to parameter values that generate simulations $\mathbf{x}$ sufficiently close to $\mathbf{y}$ under the topology induced by some distance measure $\mathcal{D}$. A key step in the application of ABC then is the choice of distance employed in the algorithm, which has led to significant research effort investigating the appropriateness of a large collection of ABC distances under different problem settings (see e.g. 2; 5; 12).

However, many such approaches make limiting assumptions on the form of the data, such as assuming that the data $\mathbf{x}$ consists of *iid* vector-valued observations, and few are compatible with more structured settings such as sequential data consisting of potentially non-Euclidean observations. Such settings arise in, for example, biomedical applications – where the data may consist of multiple instances of longitudinal data recorded for different patients or subjects, who may reasonably be treated as *iid* instances from some underlying distribution – or in the natural or behavioural sciences – where multiple independent experiments may be run in which the behaviour of some system or a collection of individuals is recorded over time. Developing SBI methods that can readily handle data consisting of multiple *iid* sequences is thus an important step in bridging the gap between explanatory, mechanistic models of these systems and data generated by the real-world counterparts to these models.

To address this challenge, we propose to build on previous work on the use of the *signature method* in computational statistics and machine learning (3; 5; 7; 9; 10; 11) by introducing ABC based on the signature maximum mean discrepancy (MMD). Our approach embodies a principled approach to automatically building distances $\mathcal{D}$ between simulated sequences and *iid* sequences from the real-world, such as medical panel data, for use in ABC procedures. Furthermore, we illustrate such an approach can be readily applied to complex, structured data encountered in such settings by demonstrating its use on a model of dynamic brain networks that generates sequences of graphs describing the evolving functional connectivity of simulated patients' brains. Further, this extends previous work on calibrating parameters for dynamic graph simulators using neural SBI methods (6).

2 Background

We provide some background on path signatures. Let $\mathcal{H}$ be a Hilbert space and $h : [0, T] \to \mathcal{H}$ a continuous $\mathcal{H}$-valued path on interval $[0, T]$. Assume further that h is of bounded variation, i.e.

$$\|h\|_{1-\text{var}} := \sup_{\mathcal{P} \in \zeta(0,T)} \sum_{j=1}^{|\mathcal{P}|-1} \left\| h_{t_{j+1}} - h_{t_j} \right\|_{\mathcal{H}} < \infty, \tag{2}$$

where $\zeta(0, T)$ is the set of all partitions of the interval $[0, T]$ and $|\mathcal{P}|$ is the size of the partition. The *path signature* (10) of h is the infinite set of tensors obtained through the following iterated integrals:

$$\text{Sig}(h) := \left(1, \int_0^T \mathrm{d}h, \int_0^T \int_0^t \mathrm{d}h \otimes \mathrm{d}h_t, \ldots, \int_0^T \int_0^t \mathrm{d}h^{\otimes(m-1)} \otimes \mathrm{d}h_t, \ldots\right) \tag{3}$$

Such objects appear in the solutions to controlled and stochastic differential equations, and provide an efficient summary of the responses of a controlled differential equation to its driving path (10). Of particular relevance to us is the *signature kernel* (9; 14). Defining the inner product

$$\langle A, B \rangle := \sum_{m \geq 0} \langle a_m, b_m \rangle_{\mathcal{H}^{\otimes m}} \tag{4}$$

for $C = (c_0, c_1, c_2, \ldots) \in \mathbb{R} \oplus \mathcal{H} \oplus \mathcal{H}^{\otimes 2} \oplus \ldots$, we arrive at the definition of the signature kernel:

Definition 1 (Signature kernel, Kiraly and Oberhauser (9)) *Let h, g be two bounded variation paths on $[0, T]$ in $\mathcal{H}$. With an inner product as in Equation 4, the* signature kernel *is defined to be*

$$k(h, g) = \langle \mathit{Sig}(h), \mathit{Sig}(g) \rangle. \tag{5}$$

It can be shown that the signature kernel computations can be completely kernelised and expressed through inner products on points in the path, permitting the sequentialisation of arbitrary "static"

kernels κ on the data space $\mathcal{X}$. This enables the signature kernel to be applied to data evolving in arbitrary topological spaces provided a suitable static kernel κ is available.

A further property of signatures is that the *expected* signature characterises the law of stochastic processes (3). An additional application of the signature kernel then is to construct maximum mean discrepancies for stochastic processes, giving a distance between distributions on paths (3; 14) as

$$\mathcal{D}_{\text{M}}(P, Q) := \left\| \mathbb{E}_{h\sim P}[\text{Sig}(h)] - \mathbb{E}_{g\sim Q}[\text{Sig}(g)] \right\|^2, \tag{6}$$

an unbiased estimate of which may be obtained as

$$\widehat{\mathcal{D}_{\text{M}}}(P, Q) = \frac{1}{n(n-1)} \sum_{i\neq j} k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + \frac{1}{m(m-1)} \sum_{i\neq j} k(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) - \frac{2}{nm} \sum_{i,j} k(\mathbf{x}^{(i)}, \mathbf{y}^{(j)}) \tag{7}$$

where $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)}$ are *iid* sequences (discretised paths) drawn from P and Q, respectively.

3 Method

In this paper, we consider a generalisation of the method presented in (5) by considering that the signature MMD, Equations (6)-(7), can be used in ABC algorithms in cases where multiple *iid* sequences are obtained from real-world data. In such cases, the discrepancy measure used in (5), namely the signature MMD between Dirac measures on real data $\mathbf{y}$ and simulated data $\mathbf{x}$, calculated as

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) := \left\| \mathbb{E}_{\mathbf{x}\sim\delta_{\mathbf{x}}}[\text{Sig}(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim\delta_{\mathbf{y}}}[\text{Sig}(\mathbf{y})] \right\|^2 = \left\| \text{Sig}(\mathbf{x}) - \text{Sig}(\mathbf{y}) \right\|^2, \tag{8}$$

can be generalised to either the scoring rule

$$\mathcal{D}_{\text{S}}(\delta_{\mathbf{x}}, Q) := \left\| \mathbb{E}_{\mathbf{x}\sim\delta_{\mathbf{x}}}[\text{Sig}(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim Q}[\text{Sig}(\mathbf{y})] \right\|^2 \tag{9}$$

if the simulation budget permits only one simulation per $\boldsymbol{\theta}$, or the distance on distributions on sequences

$$\mathcal{D}_{\text{M}}(P_{\boldsymbol{\theta}}, Q) := \left\| \mathbb{E}_{\mathbf{x}\sim P_{\boldsymbol{\theta}}}[\text{Sig}(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim Q}[\text{Sig}(\mathbf{y})] \right\|^2$$

if the simulation budget allows for multiple draws from the simulator at parameters $\boldsymbol{\theta}$. Both distances may be estimated as in Equation (7). The parameter posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ may then be approximated through, for example, rejection ABC, such that the ABC posterior follows

$$\pi_{\text{REJ}}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int \mathbb{1}\left[\mathcal{D}_{\text{A}} \leq \varepsilon\right] p(\mathbf{x} \mid \boldsymbol{\theta}) \mathrm{d}\mathbf{x}, \tag{10}$$

where $\mathbf{y}$ is the collection of observed, real-world sequences and $\mathcal{D}_{\text{A}}$ is Equation (9) or (6) as desired.

4 Experiments

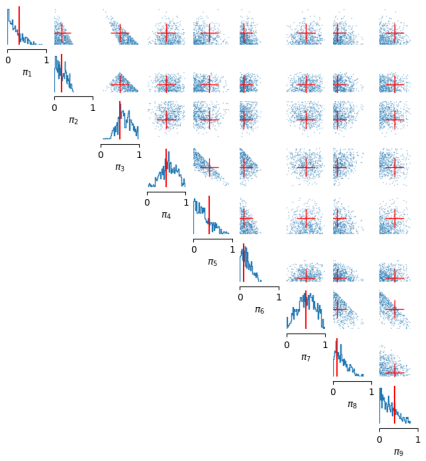
To illustrate our proposed method, we perform experiments with a dynamic graph simulator based on a larger model which has been used to simulate evolving functional brain network connectivities in patients (18). The model may be summarised as follows:

$$s_t \mid s_{t-1} = \ell \sim \text{Multinomial}(\pi_{\ell 1}, \ldots, \pi_{\ell S}) \qquad \text{(state transition)} \tag{11}$$

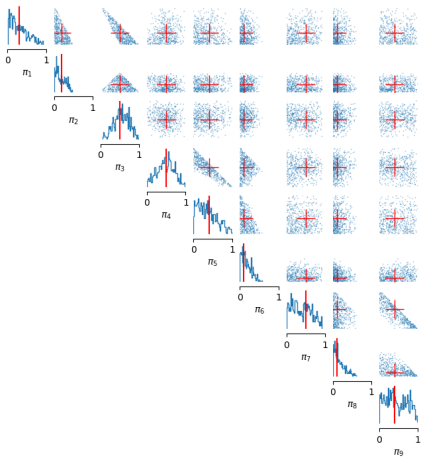
$$\beta_t \mid s_t \sim \mathcal{N}(\mu_s, \Sigma_s) \qquad \text{(logit of connection probabilities)} \tag{12}$$

$$W_t \mid \beta_t \sim \text{Bernoulli}(f(\beta_t, \Omega)). \qquad (t\text{-th adjacency matrix}) \tag{13}$$

In the equations above, the $s_t \in \{1, \ldots, S\}$ denote the state of the subject at time t; the π_{rs} denote the transition probabilities between states $r, s = 1, \ldots, S$; $\beta_t \in \mathbb{R}^{Q^2}$ consist of the logits of the connection probabilities between each of the Q communities of nodes in the brain; μ_s and Σ_s are parameters depending on the current state s; W_t is the adjacency matrix at time t; and $\Omega \in \mathbb{R}^{N\times Q}$ is a matrix denoting community membership, where $\Omega_{nq} = 1$ if node n is in community q and 0 otherwise. In our experiments, we observe only the brain connectivity matrices W_t, such that the simulated data $\mathbf{x}$ is a sequence of graphs, while the pseudo-observed (simulated) data $\mathbf{y}$ consists of $R > 1$ *iid* sequences of graphs and represent a collection of dynamic functional connectivity matrices observed in the "real world" from "real" patients.



(a) Signature MMD posterior.



(b) Posterior with hand-crafted summary statistics.

We consider the task of inferring the transition matrix elements π_{ij} using our proposed ABC pipeline based on the signature MMD. We use a Weisfeiler–Lehman graph kernel (15) to construct the static kernel κ employed in the signature kernel computations, in order for the signature MMD to be applied immediately to dynamic graph data. We place uniform Dirichlet priors on each row of the transition matrix π (such that the hyperparameters are $(1, 1, 1)$); further experimental details are provided in Appendix A.1. We use a rejection ABC scheme in which 10^4 parameters are sampled from the prior, from which we retain the top 5% of samples generating the smallest distances. As a baseline comparison method, we perform rejection ABC using the squared Euclidean distance between hand-crafted summary statistics, which we take to be the counts of edges between each pair of communities at each timestep. These are well-known sufficient statistics for the static SBM, see *e.g.* (8).

In Figures 1a and 1b, we show the ABC posteriors obtained using the distance in Equation 9 and the hand-crafted summary statistics, respectively. The marginal densities are shown on the diagonal, while the off-diagonals show the joint bivariate densities. Both posteriors are fairly diffuse and assign relatively high density to the true parameter values, shown as the red lines and crosses. We further show in Table 1 statistics describing the distribution of the L_2 distances between summary statistics of the generated data, and those of samples from the posterior predictive distributions associated with the ABC posteriors based on (i) Equation (9), and (ii) directly using these hand-crafted summary statistics. From this, we see that our signature-based approach tends to generate smaller distances than the approach based on hand-crafted summary statistics, even compared on this same basis. Overall, this suggests that our method is a useful automatic approach to constructing approximate parameter posteriors in settings involving multiple *iid* sequences from the real-world.

5 Conclusion

In this paper, we investigate the use of the signature maximum mean discrepancy in approximate Bayesian computation when panel data is available from the real-world, as may be encountered in medical settings or in the natural or behavioural sciences. Through experiments with a dynamic graph simulator used to model evolving functional brain connectivity, we demonstrate that such an approach may be successfully applied to temporal graph data to relate mechanistic models to data.

Method	Min	25%	50%	75%	Max
Signature (ours)	2420	2780	2947	3130	4800
Hand-crafted	2623	3012	3171	3369	5824

Table 1: Posterior predictive checks for the dynamic graph simulator. Columns show percentiles of the distribution of L_2 distances between summary statistics. Lower values indicate better predictions.

References

[1] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[2] 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.

[3] Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *arXiv preprint arXiv:1810.10971*, 2018.

[4] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[5] Joel Dyer, Patrick Cannon, and Sebastian M Schmon. Approximate Bayesian Computation with Path Signatures. *arXiv preprint arXiv:2106.12555*, 2021.

[6] Joel Dyer, Patrick Cannon, J Doyne Farmer, and Sebastian M Schmon. Calibrating agent-based models to microdata with graph neural networks. In *ICML 2022 Workshop AI for Agent-Based Modelling*, 2022.

[7] Joel Dyer, Patrick W Cannon, and Sebastian M Schmon. Amortised likelihood-free inference for expensive time-series simulators with signatured ratio estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 11131–11144. PMLR, 2022.

[8] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[9] Franz J. Kiraly and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31):1–45, 2019. URL `http://jmlr.org/papers/v20/16-314.html`.

[10] 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.

[11] James Morrill, Andrey Kormilitzin, Alejo Nevado-Holgado, Sumanth Swaminathan, Sam Howison, and Terry Lyons. The Signature-Based Model for Early Detection of Sepsis from Electronic Health Records in the Intensive Care Unit. *Computing in Cardiology*, 2019-Septe: 2–5, 2019. ISSN 2325887X. doi: 10.23919/CinC49843.2019.9005805.

[12] 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.

[13] Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[14] Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The signature kernel is the solution of a goursat pde. *SIAM Journal on Mathematics of Data Science*, 3(3): 873–899, 2021.

[15] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(77): 2539–2561, 2011.

[16] Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis, and Michalis Vazirgiannis. Grakel: A graph kernel library in python. *Journal of Machine Learning Research*, 21(54):1–5, 2020.

[17] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

[18] Chee-Ming Ting, S Balqis Samdin, Meini Tang, and Hernando Ombao. Detecting dynamic community structure in functional brain networks across individuals: a multilayer approach. *IEEE Transactions on Medical Imaging*, 40(2):468–480, 2020.

A Appendix

A.1 Further details on experimental setup

In the reported numerical experiments, we take: the number of states $S = 3$; number of brain regions of interest (nodes) $N = 20$; number of time steps $T = 30$; number of "observed" subjects in $\mathbf{y}$ to be $R = 10$; and four communities of nodes (such that $\Omega \in \{0, 1\}^{N \times 4}$) for which the network partition is assumed to be known.

A.2 Further details on signature kernel computations

For all signature kernel computations, we use the `sigkernel`[1] package (14), and use a dyadic order 1 in the finite element scheme to solve the corresponding Goursat partial differential equation. As the static kernel κ, we use a Weisfeiler-Lehman (WL) kernel in which all nodes are assigned the same initial label of 1 and we use 2 iterations of the WL procedure and a normalised Vertex Histogram as the base kernel within the WL kernel. For the graph kernel computations, we use the `grakel`[2] package (16).

[1] `https://github.com/crispitagorico/sigkernel`

[2] `https://github.com/ysig/GraKeL`