viaABC — Variational Inference Assisted ABC

Jun Won Park[1] and Sanket Rane[1]

[1]Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA

March 19, 2025

Contents

Abstract

-Likelihood free parameter estimation. -Reliance on summary statistics – issues with that. -Multidimensional data – ill posed inverse problems. -Use of generative models to convert inverse problem into a (machine) learning problem

We introduce viaABC (Variational Inference Assisted Approximate Bayesian Computation), a novel framework that integrates Approximate Bayesian Computation (ABC) with Variational Autoencoders (VAEs). By leveraging the representational learning of VAEs, viaABC learns a manifold representation of data beyond Euclidean space, which is particularly advantageous for complex datasets, such as hierarchical time-series, spatial, spatio-temporal, and stochastic dynamical system data.

1 Introduction

Need for a new approach Blurb about inverse problems, Bayesian approaches to solve inverse problems, their shortcom- ings.

ABC for likelihood free inference, summary statistics are NOT always the answer, discuss con- temporary alternatives for summary statistics.

Multidimensional data is a norm in biology now, how to handle large scale data or generate summary statistics from them?

Our solution is to perform inference in the latent space – follows manifold hypothesis, which posits that high-dimensional data often lie in low dimensional manifolds within the high-dimensional data. Better representation to perform causal inference on dynamical data.

In this publication, we introduce viaABC (Variational Inference Assisted Approximate Bayesian Computation), a novel framework that integrates Approximate Bayesian Computation (ABC) with Variational Autoencoders (VAEs). By leveraging the representational learning of VAEs, viaABC learns a manifold representation of data beyond Euclidean space, which is particularly advan- tageous for complex datasets, such as hierarchical time-series, spatial, spatio-temporal, and stochastic dynamical system data.

Mechanistic inference is essential for understanding phenomena across various scientific disciplines, as it enables the mathematical modeling of dynamical systems and the extraction of their underlying parameters. Therefore, accurately inferring model parameters, denoted as θ, from experimentally observed data, y^{obs}, is crucial. In the Bayesian framework, this involves determining the posterior distribution

$$\pi(\theta|y^{obs}) \propto f(y^{obs}|\theta) \cdot \pi(\theta)$$

of the parameters from the data by calculating the likelihood $f(y^{obs}|\theta)$ of such observations under the model with given parameter values. However, deriving the likelihood function theoretically is often intractable, computationally expensive, or too complex for direct optimization. Approximate Bayesian Computation (ABC) was introduced as an alternative framework for approximating posterior distributions when traditional likelihood-based methods are impractical[1]. This approach either employs summary statistics to represent both simulated and observed data using a distance metric to gauge their similarity in the case high-dimensional data or it directly compares the raw simulated and observed datasets. In both cases, ABC aims to identify the parameter values that minimize the discrepancy between the two. Numerous variants of the ABC algorithm have since been developed, including ABC Monte Carlo Markov Chain[2], ABC Sequential Monte Carlo[3], and ABC Population Monte Carlo[4], each of which has demonstrated effectiveness across a wide range of scientific disciplines.

Despite their success, the accuracy and efficiency of existing Approximate Bayesian Computation (ABC) methods heavily depend on the selection of hyperparameters, including summary statistics and distance

metrics, which are often data-dependent and require extensive tuning. In particular, reducing data to a few informative summary statistics is a critical step that directly influences inference accuracy. The choice of these statistics must be made carefully, as it determines the quality of the representation and, consequently, the reliability of the results[5]. Additionally, the choice of distance metric, including how to weight each statistics, also influences the accuracy of the parameter estimation. **FIX THIS FIX THIS FIX THIS**

To address these limitations, we introduce viaABC, a novel statistical inference framework that leverages techniques from computer vision and natural language processing to enhance approximate posterior inference in mechanistic modeling. By learning a structured representation of data through variational inference, viaABC reduces the reliance on manually selected summary statistics and distance metrics, thereby improving both the accuracy and robustness of ABC-based inference.

1. We show that VAE can learn a dense, vector representation of the data, eliminating the need for using summary statistic to represent data
2. We show that our framework is metric invariant, eliminating the need for choosing an appropriate distance metric function and how to weight each statistics

Our work will primarily focus on multivariate time-series data. Our method employs self-supervised pre-training on simulated multivariate time-series data. During pre-training, the model captures temporal and cross-channel dependencies, constructing a latent distribution for each time step. This process yields an entangled representation of the time-series data. The pre-trained model then generates latent representations for simulated data, which are subsequently compared to the representation of y^{obs} for parameter inference using ABC-SMC.

2 Related Work

2.1 ABC

This section provides a comprehensive review and development of the theoretical foundations of Approximate Bayesian Computation (ABC), with a particular focus on its applications to dynamical systems and statistical parameter inference before introducing the variational inference approach in the context of ABC.

ABC Rejection

Approximate Bayesian Computation (ABC) rejection is a method for parameter inference when likelihoods are intractable or computationally expensive to evaluate[1]. The process begins by defining prior distributions for the parameters, a distance metric, and summary statistics. Common distance metrics include L1 and L2 distances, while statistics often include the mean, variance, or combinations of moments. A parameter set is then sampled from the prior distributions and passed through a predefined simulator, a mathematical model describing the system's dynamics. The resulting statistics are computed and compared to those from the observed data. If the distance between the simulated and observed statistics falls within an acceptance threshold, the parameter set is retained for the approximate posterior; otherwise, it is rejected.

ABC Sequential Monte Carlo

Approximate Bayesian Computation (ABC) rejection sampling relies on several hyperparameters, including the rejection threshold and the choice of distance metric for comparing simulated and observed data. A small rejection threshold typically results in a low acceptance rate, leading to high computational costs,

especially when the prior distribution differs significantly from the posterior[3]. To address this, ABC Sequential Monte Carlo (ABC-SMC) was introduced, which begins with a large threshold and progressively reduces it. In each iteration, samples, referred to as particles, are drawn from the approximate posterior of the previous iteration, improving efficiency.

2.2 Self-supervised Learning (SSL)

Self-supervised learning (SSL) has achieved remarkable success in natural language processing and computer vision, with applications extending beyond text and images to modalities such as video, audio, and time series[6]. Broadly, SSL can be categorized into contrastive learning and masked learning.

Contrastive Loss

The core idea of contrastive loss is to generate two augmented versions of a given input: a positive sample that is similar to the input and a negative sample that is dissimilar. The objective is to encourage similar representations to cluster together while pushing dissimilar ones apart.
PLACEHOLDER FOR CONTRASTIVE LOSS APPROACHES IN TIME-SERIES

Masked Modeling

The masked modeling approach learns useful representations by masking parts of the input and reconstructing them using the unmasked portions. This technique has been successful in natural language processing, as demonstrated by masked language modeling in BERT, where some tokens are randomly masked, and the model predicts the missing tokens based on the surrounding context[7]. In computer vision, an image is divided into patches, with some randomly masked, and the model reconstructs the original image using the remaining patches[8].
PLACEHOLDER FOR MASKED MODELING APPROACHES IN TIME-SERIES

2.3 Varitional Inference

Variational inference has been proposed as a machine learning approach to approximate intractable probability densities[9][10]. Given a dataset $\mathbf{X} = {x^{(i)}}_{i=1}^{N}$ consisting of N i.i.d. samples of a continuous or discrete variable x, we assume that the data are generated by an underlying random process involving an unobserved continuous latent variable z[11]. The generative process follows two steps:

1. A latent variable $z^{(i)}$ is drawn from a prior distribution $p_{\theta^*}(z)$.
2. A corresponding observation $x^{(i)}$ is sampled from a conditional distribution $p_{\theta^*}(x \mid z)$.

We assume that both the prior $p_{\theta^*}(z)$ and the likelihood $p_{\theta^*}(x \mid z)$ belong to parametric families of distributions $p_\theta(z)$ and $p_\theta(x \mid z)$, respectively, and that their probability density functions are differentiable with respect to both θ and z. However, much of this process remains hidden: the true parameters θ^*, as well as the values of the latent variables $z^{(i)}$, are unknown. **PARAPHRASE THIS AND CITE KINGMA**

3 Methodology

3.1 Model Architecture

We extend the transformer-based Masked Autoencoder (MAE), originally developed for computer vision, where it employs the Vision Transformer (ViT) as the backbone model for both the encoder and decoder,

to the domain of multivariate time-series data[8,12]. While MAE has been widely employed for image representation learning, its direct application to time-series data presents challenges due to the temporal dependencies and multivariate structure inherent in such data. To address these challenges, we introduce the Time-Series Masked Variational Autoencoder (TSMVAE), a novel adaptation that integrates the principles of MAE with a variational autoencoder framework tailored for time-series modeling.

Similar to all other autoencoders, our model consists of an encoder that maps the observed signal into a latent representation and a decoder that reconstructs the original signal from this latent space. However, TSMVAE diverges from the original MAE in two fundamental ways to better capture the structure of multivariate time-series data:

Temporal Patching: Given a D-dimensional time-series, $\{X_t\}_{t=1}^{T}$, where $X_t \in \mathbb{R}^D$, TSMVAE partitions the sequence along the time axis, forming patches that correspond to different time intervals. Each patch is then projected into a higher-dimensional space via a linear transformation, facilitating the encoding of temporal dependencies.

Variational Latent Space Representation: The encoded output undergoes two additional linear transformations. The first maps the encoded representations into a predefined latent dimension, structuring the latent space effectively. The second transformation introduces variational layers, which impose a probabilistic structure on the latent variables, thereby enabling the learning of meaningful latent distributions.

These modifications enable TSMVAE to effectively learn representations from multivariate time-series data time-stamp wise while leveraging the power of masked self-supervised learning. Figure 1 provides a high-level illustration of our proposed model. For further details on the model architecture, we refer the reader to prior work.

3.2 Pretraining

Suppose we have a mathematical model or data-generating process, denoted as f, which maps a k-dimensional parameter vector to a d-dimensional multivariate time series:

$$f(\theta) = \{X_t\}_{t=1}^{T}, \quad \theta \in \mathbb{R}^k.$$

To construct a training dataset, we employ Latin hypercube sampling (LHS) to generate N parameter sets, each of which is used to simulate an associated multivariate time series. Specifically, we denote the i-th simulated time series as

$$\{X_t^{(i)}\}_{t=1}^{T}, \quad X_t^{(i)} \in \mathbb{R}^d,$$

where each sample represents a realization of the underlying process. This procedure yields a dataset consisting of N multivariate time series samples. To facilitate the study of the model's dynamics across a diverse range of parameter configurations, we employ Latin hypercube sampling to ensure a well-distributed exploration of the parameter space, thereby mitigating the risk of overfitting to any particular mode.

During training, Gaussian noise ϵ is dynamically injected at each time step such that

$$\tilde{X}_t = X_t + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

The variational autoencoder (VAE) is trained using noise-injected time series data as input, learning to reconstruct the original, denoised data. In addition to minimizing the reconstruction loss, the VAE also minimizes the Kullback–Leibler (KL) divergence to regularize the latent space and encourage a structured representation. In this study, we enforce a Gaussian prior on the latent space.

Loss function: We define the loss function as a combination of reconstruction loss and KL divergence, applied at each time step. Since we employ the VAE in a time-step-wise manner, the loss function for each time step t consists of the reconstruction loss and the Gaussian KL divergence loss:

$$L_t = \frac{1}{2} MSE(x_t, \hat{x}_t) + \beta D_{KL}(q_\phi(z_t|x_t)||p_\theta(z_t))$$

Aggregating the losses over the entire time sequence, the total loss is given by:

$$\begin{aligned} L &= \frac{1}{T}\sum_{t=1}^{T} L_t \\ &= \frac{1}{T}(\frac{1}{2N \cdot D}\sum_{i=1}^{N}\sum_{j=1}^{d}(x_{t,j}^{(i)} - \hat{x}_{t,j}^{(i)})^2 + \beta D_{KL}(q_\phi(z_t|x_t)||p_\theta(z_t))) \end{aligned}$$

where β is a hyper-parameter that balances latent channel capacity and independence constraints with reconstruction accuracy[13] **[rewrite this]**.

The TSMVAE model has a special CLS token, which will be used to output a latent vector representation of a multi-variate time-series data. This is often known as the CLS pooling.

3.3 ABC-SMC

INSERT ALGORITHM HERE

3.4 Dataset

Deterministic Lotka-Volterra model

The Lotka-Volterra (LV) model describes the interaction between predators and prey[14 15]. This deterministic model, which represents the populations of prey x and predators y, is expressed as a system of nonlinear differential equations:

$$\begin{aligned} \frac{dx}{dt} &= ax - cxy \\ \frac{dy}{dt} &= bxy - dy \end{aligned} \tag{1}$$

where a denotes the intrinsic growth rate of the prey, d is the mortality rate of the predators, c represents the predation rate coefficient, and b characterizes the efficiency with which consumed prey are converted into predator offspring. In this study, we set $c = d = 1$ and aim to infer the parameters $\theta = (a, b)$ from prior distributions where $a, b \sim \mathrm{Uniform}(-10, 10)$. Following the methodology of Toni etal.[3], we solve the system of ordinary differential equations with the initial condition $(x(0), y(0)) = (1, 0.5)$ using $\theta = (1, 1)$. The state variables (x, y) are sampled at eight distinct time points between $t = 0$ and $t = 15$.

Subsequently, normally distributed noise $\epsilon_t \sim \mathcal{N}(0, 0.5^2)$ is added to each data point, thereby forming the observed dataset $s^{\text{obs}} = \{x_1, y_1, \ldots, x_8, y_8\}$

To evaluate viaABC's performance, we compare the final posterior distributions with ABC-DRF, ABC-SMC, and ABC-SMC-DRF. Algorithm drf is blah blah. Talk about table sizes, different kernel in each algorithm. We impose the series of tolerance thresholds $\epsilon_1 = 0.3, \epsilon_2 = 0.2, \epsilon_1 = 0.1, \epsilon_4 = 0.04$. $\epsilon_4 = 0.04$ is the cosine dissimilarity between the ground truth and the observed data. To produce 1,000 accepted particles, viaABC requires $N = 34,474$ simulations.

Stochastic Susceptible-Infected-Recovered model

The stochastic Susceptible-Infected-Recovered (SIR) model is a fundamental framework in epidemiology for capturing the spread of infectious diseases within a population[?] . Unlike its deterministic counterpart, which employs ordinary differential equations, the stochastic SIR model incorporates intrinsic randomness in disease transmission and recovery, making it particularly suitable for modeling outbreaks in finite populations. The system tracks the evolution of three state variables: S, I, and R, representing the number of susceptible, infected, and recovered individuals, respectively. The model dynamics are governed by the following reaction scheme:

$$S + I \xrightarrow{\beta} 2I$$
$$I \xrightarrow{\gamma} R$$

where β denotes the transmission rate per susceptible-infected pair, and γ represents the recovery rate of infected individuals. To capture the stochastic nature of infection and recovery events, we employ the Gillespie algorithm[16], which simulates the discrete event-driven evolution of the system.
In this study, we set $\beta = 2$ and $\gamma = 0.5$ and simulate the epidemic dynamics over a time horizon of $T = 15$ days, starting from an initial condition of $(S(0), I(0), R(0)) = (290, 10, 0)$ in a closed population of size $N = 300$. The stochastic trajectories are recorded at thirty-five distinct time points, and to account for observational uncertainty, normally distributed noise $\epsilon_t \sim \mathcal{N}(0, 1^2)$ is added to each recorded data point. The resulting dataset, denoted as $s^{\text{obs}} = \{S_1, I_1, R_1, \ldots, S_{35}, I_{35}, R_{35}\}$ serves as the empirical basis for inference in this study. Since events in the Gillespie algorithm occur at irregular time intervals, we interpolate the simulated trajectories to align their time indices with those of the observed data, ensuring comparability in the inference process.

Deterministic Marginal Zone B Cell Dynamics

PLACEHOLDER[17]

4 Results

4.1 Deterministic Lotka-Volterra model

4.2 Stochastic Susceptible-Infected-Recovered model

Figure Something: a) Predicted trajectories of stochastic sir and the observed data points. (b) Final approximate posterior distribution of parameters inferred by the ABC-SMC sampler.

Statistics	viaABC iterations				ABC-DRF	ABC-SMC	ABC-SMC-DRF
	1	2	3	4 (Final posterior)			
$\mathbb{E}$(a)	0.9259	1.1958	1.0152	**1.0579**	0.7562	1.2912	1.1215
Var(a)	0.6339	0.3658	0.0435	**0.0078**	0.5953	0.1104	0.0333
$\mathbb{E}$(b)	1.6541	1.0758	1.2000	1.0806	1.3092	**1.0269**	0.9704
Var(b)	1.4106	0.3181	0.1369	**0.0274**	0.3593	0.1049	0.0313

Table 1: Means and variances of marginal posterior distributions for the deterministic Lotka-Volterra model from viaABC, ABC-DRF, ABC-SMC and ABC-SMC-DRF. The best result for each statistic across different algorithms is in bold (E(a), E(b) closest to true values (a, b) = (1, 1), and lowest variance in each statistic). Results are taken from Dinh et al.[18].

4.3 Deterministic Marginal Zone B Cell Dynamics

Figure Something: a) Predicted trajectories of marginal bzone. (b) Final approximate posterior distribution of parameters inferred by the ABC-SMC sampler.

Figure Something: Comparison between

Cited literature

1. Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.
2. Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.
3. Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.
4. Christian P Robert, Mark A Beaumont, Jean-Michel Marin, and Jean-Marie Cornuet. Adaptivity for abc algorithms: the abc-pmc scheme. *arXiv preprint arXiv:0805.2256*, 2008.
5. Mattias Åkesson, Prashant Singh, Fredrik Wrede, and Andreas Hellander. Convolutional neural networks as summary statistics for approximate bayesian computation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6):3353–3365, 2021.
6. Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.
7. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
8. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
9. Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

10. David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017.

11. Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

12. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

13. Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

14. Alfred James Lotka. *Elements of physical biology*. Williams & Wilkins, 1925.

15. Vito Volterra. Variations and fluctuations of the number of individuals in animal species living together. *ICES Journal of Marine Science*, 3(1):3–51, 1928.

16. Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

17. Melissa Verheijen, Sanket Rane, Claire Pearson, Andrew J Yates, and Benedict Seddon. Fate mapping quantifies the dynamics of b cell development and activation throughout life. *Cell reports*, 33(7), 2020.

18. Khanh N Dinh, Zijin Xiang, Zhihan Liu, and Simon Tavaré. Approximate bayesian computation sequential monte carlo via random forests. *arXiv preprint arXiv:2406.15865*, 2024.