Statistics and Computing (2023) 33:142
https://doi.org/10.1007/s11222-023-10319-y

ORIGINAL PAPER

Manifold-constrained Gaussian process inference for time-varying parameters in dynamic systems

Yan Sun[1] · **Shihao Yang**[1]

Received: 26 September 2022 / Accepted: 27 September 2023 / Published online: 16 October 2023

Abstract
Identification of parameters in ordinary differential equations (ODEs) is an important and challenging task when modeling dynamic systems in biomedical research and other scientific areas, especially with the presence of time-varying parameters. This article proposes a fast and accurate method, TVMAGI (Time-Varying MAnifold-constrained Gaussian process Inference), to estimate both time-constant and time-varying parameters in the ODE using noisy and sparse observation data. TVMAGI imposes a Gaussian process model over the time series of system components as well as time-varying parameters, and restricts the derivative process to satisfy ODE conditions. Consequently, TVMAGI does not require any conventional numerical integration such as Runge–Kutta and thus achieves substantial savings in computation time. By incorporating the ODE structures through manifold constraints, TVMAGI enjoys a principled statistical construct under the Bayesian paradigm, which further enables it to handle systems with missing data or unobserved components. The Gaussian process prior also alleviates the identifiability issue often associated with the time-varying parameters in ODE. Unlike existing approaches, TVMAGI can be applied to general nonlinear systems without specific structural assumptions. Three simulation examples, including an infectious disease compartmental model, are provided to illustrate the robustness and efficiency of our method compared with numerical integration and Bayesian filtering methods.

Keywords Ordinary differential equations · Inverse problem · Time-varying parameter estimation · Gaussian process · Bayesian inference

1 Introduction

Ordinary Differential Equations (ODEs) are often used to analyze the behavior of dynamic systems, such as the spread of infectious diseases (Li and Muldowney 1995), interactions between species (Takeuchi et al. 2006), and viral dynamics (Perelson et al. 1996). This paper studies a general formulation of ODE equations, where some of the parameters are allowed to be time-varying:

$$\dot{\boldsymbol{x}}(t) \equiv \frac{d\boldsymbol{x}(t)}{dt} = \mathbf{f}(\boldsymbol{x}(t), \boldsymbol{\theta}(t), \boldsymbol{\psi}, t), t \in [0, T] \quad (1)$$

Here, $\boldsymbol{x}(t)$ is the series of system outputs from time 0 to T, $\boldsymbol{\psi}$ denotes time-constant parameters, $\boldsymbol{\theta}(t)$ denotes time-varying parameters, and $\mathbf{f}$ is a set of general functions that characterize the derivative process. When $\mathbf{f}$ is non-linear, the system outputs $\boldsymbol{x}(t)$ typically do not have analytic solutions. To solve $\boldsymbol{x}(t)$ given initial conditions $\boldsymbol{x}(0)$ and parameters $\boldsymbol{\theta}(t)$ and $\boldsymbol{\psi}$, numerical integration methods are often required, such as Euler's Method or Runge-Kutta Method (Lapidus and Seinfeld 1971).

This paper focuses on the inverse problem that, given the observations, how to efficiently draw inference on the ODE parameters. Our goal is to estimate time-constant parameters $\boldsymbol{\psi}$ and time-varying parameters $\boldsymbol{\theta}(t)$ inside the ODE from data. In real world, the observation data of system components $\boldsymbol{x}$ are often obtained at discrete time points and are subject to measurement errors. We thus assume that we

This research was partially supported by NSF award DMS-2318883.

✉ Shihao Yang
shihao.yang@isye.gatech.edu

Yan Sun
ysun614@gatech.edu

[1] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Dr NW, Atlanta, GA 30332, USA

observe $\boldsymbol{y}(\boldsymbol{\tau}) = \boldsymbol{x}(\boldsymbol{\tau}) + \boldsymbol{\epsilon}(\boldsymbol{\tau})$, where $\boldsymbol{\tau}$ denotes the observation time points while error $\boldsymbol{\epsilon}(\boldsymbol{\tau})$ denotes Gaussian noise. We focus on the inference of $\boldsymbol{\theta}(t)$ and $\boldsymbol{\psi}$ given $\boldsymbol{y}(\boldsymbol{\tau})$, with emphasis on nonlinear structure $\mathbf{f}$.

The time-varying parameter $\boldsymbol{\theta}(t)$ in the ODE is often important yet challenging to recover from real-world data. For example, during a pandemic, the time-varying disease reproduction number is critical for public health policy decisions. However, its estimation can still be crude despite the best effort (Abbott et al. 2020). The time-varying $\boldsymbol{\theta}(t)$ provides too much degree of freedom to the ODE system, and two different $\boldsymbol{\theta}(t)$ can both give $\boldsymbol{x}(t)$ that fits the observation data, resulting in identifiability issues (Miao et al. 2011). Such high degree of freedom in the time-varying parameters also gives overfitting issues for the usual gold standard numerical integration method, which tends to produce $\boldsymbol{\theta}(t)$ that fits the observation data exactly.

2 Review of related literature

Surrogate models (including Gaussian Processes) have long been proposed to approximate the dynamic systems and facilitate computation (Calderhead et al. 2008; Dondelinger et al. 2013; Barber and Wang 2014; Ghosh et al. 2017; Lazarus et al. 2018; Wenk et al. 2019), although most of them only accommodate time-constant parameters (Dondelinger et al. 2013; Yang et al. 2021; Wenk et al. 2019; Calderhead et al. 2008; Girolami and Calderhead 2011). Early Bayesian works include (Skilling 1992), which brought out the idea of modeling the derivative process as Gaussian convolution of some hidden function (i.e. radial basis function interpolation in discrete setting). More recent works include (Tronarp et al. 2019; Krämer et al. 2022), which combined Gaussian process with Bayesian filtering for non-linear ODEs/PDEs. However, all these works (Skilling 1992; Tronarp et al. 2019; Krämer et al. 2022) focused mainly on solving the initial value problems of the ODE (i.e. forward problems) rather than the inverse problems. For the inverse problems, Calderhead et al. (2008) used the product-of-expert heuristic to construct GP posterior to estimate time-constant parameter. With notable lack of theoretical rigor, the model artificially injected noise to the posterior so as to balance the overconfidence issue from product-of-expert approach. Other ideas, such as Bayesian sampling with manifold constraints, has also been studied (Diaconis et al. 2013; Girolami and Calderhead 2011), but the usual approaches still suffer from the computational burden of conventional numerical integration methods when applying to ODEs.

For time-varying parameters inference in the ODE system, existing methods all have their deficiencies (Wu 2005). For example, Li et al. (2002) relied on time-consuming numerical integration; Huang et al. (2006) proposed a Bayesian parametric approach to model time-varying coefficients in the HIV-1 dynamic model, sacrificing some flexibility; Cao et al. (2012) developed an efficient two-stage local polynomial estimation method that circumvents conventional numerical integration for a non-parametric time-varying parameters, but required ODE system to have linear dependency on the time-varying parameters (see Eq.(A2) in Supplementary Material). Bayesian filtering methods are also explored in the time-varying ODE parameter inferences, although lacking some statistical rigor. For example, Pei and Shaman (2020) and Shaman and Karspeck (2012) applied Ensemble Adjustment Kalman Filter (EAKF) algorithm to estimate parameters in a metapopulation SEIR model. Schmidt et al. (2021) proposed an extended Kalman Filter approach based on Gauss-Markov process that can infer time-varying parameter but cannot accommodate time-constant parameters any more. The use of GP to model time-varying parameters have been previously explored on the Stochastic Differential Equation (SDE) (Pokern et al. 2013; Papaspiliopoulos et al. 2012; Hairer et al. 2011), but its applicability with respect to the ODE remains more open, although its potential for mitigating ill-posedness has been recognized (Cotter et al. 2010).

One concurrent work with ours that worth special note is the article of Chen et al. (2021). In this article, Chen et al. introduced the PDE-constraint GP optimization to solve inverse problems. The paper also proposed to model the location-dependent parameters as another GP, which resulted in a seemingly similar objective function to our proposed posterior function. However upon close examination, we can see that the approaches are different, with each of its own merits. We highlight a few key differences here. (1) The motivations are different. The paper Chen et al.(2021) incorporates the PDE structure using an engineering approach through constraint optimization where the PDE constraint is directly added to the objective function. We took a more statistical approach through conditioning in the probability space where the ODE information in the posterior function is naturally derived from Bayesian principles. (2) The end objective function and posterior function are seemingly similar but with major differences. In the objective function of Chen et al., the PDE constraint is exact at the discretization points, and it would need equality constraint optimization such as Lagrange multiplier. In our approach, the ODE information is incorporated through posterior conditioning and the posterior function only needs unconstrained optimization to get the Maximum A Posteriori (MAP). Looking only at the posterior function, our approach is more like "penalizing" the deviation from ODE information, which is naturally derived through Bayesian principles. (3) There is a small but solid difference in the GP kernel selection: Chen et al.(2021) uses the radial basis function (RBF) kernel, while we propose to use Matern kernel in the least degree of freedom possible.

The Matern kernel, thanks to the finite degree of smoothness, tends to give much better numerical stability than the RBF kernel, which eliminated the need for artificial perturbation or nugget in the kernel and removed assumptions on the smoothness of the time-varying/location-dependent parameters. We will further highlight our distinct contribution in the next section.

3 Our contribution

We propose a fast and statistically principled method to infer time-varying $\boldsymbol{\theta}(t)$ and time-constant $\boldsymbol{\psi}$ from noisy observations of ODE. The key idea is to use Bayesian approach and place Gaussian process (GP) prior on $\boldsymbol{x}(t)$ and time-varying parameters $\boldsymbol{\theta}(t)$, thus the identifiability issue is mitigated using the informative prior that favors smoother parameter curves. Our method is built upon the prior work of MAnifold-constrained Gaussian process Inference (Yang et al. 2021) where the Gaussian process $\boldsymbol{x}(t)$ is restricted on a manifold that satisfies the ODE system. Placing a Gaussian process on $\boldsymbol{x}(t)$ facilitates a fast inference on $\boldsymbol{\theta}(t)$, as it no longer requires conventional numerical integration such as Runge–Kutta. Our approach also adheres to the classical Bayesian paradigm with principled posterior derivation. Through a Gaussian process model on $\boldsymbol{\theta}(t)$, we are able to generalize the MAnifold-constrained Gaussian process Inference to the situation where time-varying and time-constant parameters co-exist. We name our method TVMAGI (Time-Varying MAnifold-constrained Gaussian process Inference), emphasizing its capability in handling time-varying parameters. We demonstrate the effectiveness of TVMAGI through three realistic simulation examples, where TVMAGI works well even when some of the system components $\boldsymbol{x}(t)$ are partially observed. Through these simulation examples, we also show that TVMAGI can outperform benchmark methods including a numerical integration approach, a Bayesian filtering approach, and a two-stage approach. Thanks to the computational savings of skipping conventional numerical integration step such as Runge-Kutta, TVMAGI has great potential to be generalized in high-dimensional and large-scale systems. TVMAGI has a distinct contribution from the previously-proposed time-constant parameter inference methods (Wenk et al. 2019; Yang et al. 2021) by investigating a much more complicated problem with functional estimate of time-varying parameters. The change from time-constant parameter to time-varying parameter also creates a notable difference in the scientific context, as parameters to be inferred in most real-world phenomena are non-stationary or changing over time.

4 Method of TVMAGI

4.1 The prior

Following standard Bayesian notation, the D-dimensional dynamic system $\boldsymbol{x}(t)$ is a realization of stochastic process $\boldsymbol{X}(t) = (X_1(t), ..., X_D(t))$, and the P-dimensional time-varying parameters $\boldsymbol{\theta}(t)$ is a realization of stochastic process $\boldsymbol{\Theta}(t) = (\Theta_1(t), ..., \Theta_P(t))$. We assume that $\boldsymbol{\Theta}(t)$ is continuous and differentiable in t during time period $[0, T]$, which helps to prevent overfit and to alleviate identifiability issue, but can be relaxed later. The prior distribution of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$ in each dimension is independent Gaussian process in each dimension d. That is,

$$\Theta_p(t) \sim \mathcal{GP}(\mu_p^{\Theta}, \mathcal{K}_p^{\Theta}), t \in [0, T], p \in \{1, \ldots, P\} \tag{2}$$

$$X_d(t) \sim \mathcal{GP}(\mu_d^{X}, \mathcal{K}_d^{X}), t \in [0, T], d \in \{1, \ldots, D\} \tag{3}$$

where $\mathcal{K}_d^X$ and $\mathcal{K}_p^{\Theta}: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are positive definite covariance kernels for GP, while μ_d^X and $\mu_d^{\Theta}: \mathbb{R} \to \mathbb{R}$ denote mean functions.

4.2 The likelihood

The observations are denoted as $\boldsymbol{y}(\boldsymbol{\tau}) = (y_1(\boldsymbol{\tau}_1), ..., y_D(\boldsymbol{\tau}_D))$, where $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ..., \boldsymbol{\tau}_D)$ is the collection of observation time points across all components. Each component $X_d(t)$ can have its own set of observation time points $\boldsymbol{\tau}_d = (\tau_{d,1}, \ldots, \tau_{d,N_d})$, where N_d is the number of observations of the d-th component. If the d-th component is not observed, then $N_d = 0$, and $\boldsymbol{\tau}_d = \emptyset$. The observation is thus assumed to be

$$Y_d(\boldsymbol{\tau}_d) = X_d(\boldsymbol{\tau}_d) + \epsilon(\boldsymbol{\tau}_d), \quad \epsilon(\boldsymbol{\tau}_d) \overset{iid}{\sim} \mathcal{N}(0, \sigma_d^2) \tag{4}$$

In this paper, notation t shall refer to time generically, and $\boldsymbol{\tau}$ shall denote specifically the observation time points.

4.3 The manifold constraint

We introduce a variable W to quantify the difference in the derivative process $\dot{\boldsymbol{X}}(t)$ between Gaussian process and ODE:

$$W = \sup_{t \in [0,T], d \in \{1,\ldots,D\}} |\dot{X}_d(t) - \mathbf{f}(\boldsymbol{X}(t), \boldsymbol{\Theta}(t), \boldsymbol{\Psi}, t)_d| \tag{5}$$

Intuitively, W is the L_∞ norm of derivative difference, and $W = 0$ if and only if $\dot{\boldsymbol{X}}(t)$ strictly satisfies the ODE structure, which is equivalent to constraining $\boldsymbol{X}(t)$ on the manifold of the ODE solutions. The advantage of L_∞ norm is further discussed in Supplementary Material Section A.1. In the ideal situation where $W \equiv 0$, the posterior distribution of $\boldsymbol{\Theta}(t)$, $\boldsymbol{\Psi}$, and $\boldsymbol{X}(t)$ shall be formulated as

$P_{\boldsymbol{\Theta}(t),\boldsymbol{\Psi},\boldsymbol{X}(t)|W,\boldsymbol{Y}(\boldsymbol{\tau})}(\boldsymbol{\theta}(t),\boldsymbol{\psi},\boldsymbol{x}(t)|W=0,\boldsymbol{Y}(\boldsymbol{\tau})=\boldsymbol{y}(\boldsymbol{\tau}))$. However, such ideal posterior is not computable in practice. Therefore, we approximate W by finite discretization on time points $\boldsymbol{I}=\{t_1,t_2,\ldots,t_n\}$, such that $\boldsymbol{\tau}\subset\boldsymbol{I}\subset[0,T]$. We similarly define $W_{\boldsymbol{I}}$ on the discretization set $\boldsymbol{I}$ as the L_∞ distance of the derivative from GP and that from ODE:

$$W_{\boldsymbol{I}}=\sup_{t\in\boldsymbol{I},d\in\{1,\ldots,D\}}|\dot{X}_d(t)-\mathbf{f}(\boldsymbol{X}(t),\boldsymbol{\Theta}(t),\boldsymbol{\Psi},t)_d| \tag{6}$$

Here $W_{\boldsymbol{I}}$ is the maximum on a finite set, and $W_{\boldsymbol{I}}\to W$ monotonically as $\boldsymbol{I}$ becomes dense. The associated computable Bayesian probability of the discretized manifold constraint $W_{\boldsymbol{I}}=0$ is

$$\begin{aligned}P(W_{\boldsymbol{I}}=0|\boldsymbol{X}(\boldsymbol{I})=\boldsymbol{x}(\boldsymbol{I}),\boldsymbol{\Theta}(\boldsymbol{I})=\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\Psi}=\boldsymbol{\psi})&\\=P(\dot{\boldsymbol{X}}(\boldsymbol{I})-\mathbf{f}(\boldsymbol{X}(\boldsymbol{I}),\boldsymbol{\Theta}(\boldsymbol{I}),\boldsymbol{\Psi},t_{\boldsymbol{I}})&\\=\mathbf{0}|\boldsymbol{X}(\boldsymbol{I})=\boldsymbol{x}(\boldsymbol{I}),\boldsymbol{\Theta}(\boldsymbol{I})=\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\Psi}=\boldsymbol{\psi})&\\=P(\dot{\boldsymbol{X}}(\boldsymbol{I})=\mathbf{f}(\boldsymbol{x}(\boldsymbol{I}),\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\psi},t_{\boldsymbol{I}})|\boldsymbol{X}(\boldsymbol{I})=\boldsymbol{x}(\boldsymbol{I}))&\end{aligned} \tag{7}$$

which is a multivariate Gaussian distribution since the time derivative $\dot{X}_d(t)$ of GP is also a GP with specific mean and covariance kernel.

about $\boldsymbol{\theta}$. Supplementary Material Section A.2 presents additional intuition regarding the manifold constraint $W_{\boldsymbol{I}}=0$.

4.4 The posterior

Therefore, a computable discretized posterior for TVMAGI inference of $\boldsymbol{X}(t)$, $\boldsymbol{\Theta}(t)$, and $\boldsymbol{\Psi}$ is:

$$\begin{aligned}&P_{\boldsymbol{\Theta}(\boldsymbol{I}),\boldsymbol{\Psi},\boldsymbol{X}(\boldsymbol{I})|W_{\boldsymbol{I}},\boldsymbol{Y}(\boldsymbol{\tau})}\\&(\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\psi},\boldsymbol{x}(\boldsymbol{I})|W_{\boldsymbol{I}}=0,\boldsymbol{Y}(\boldsymbol{\tau})=\boldsymbol{y}(\boldsymbol{\tau}))\end{aligned} \tag{8}$$

Equation (8) is the computable discretized posterior of TVMAGI inference. In this paper, we consider the Maximum A Posteriori (MAP) as the fast point estimate from TVMAGI, while the Posterior Mean and the Posterior Interval are the formal Bayesian inference results that further quantify the uncertainty.

4.5 Closed-form derivation

The posterior distribution of $\boldsymbol{X}(t)$, $\boldsymbol{\Theta}(t)$, and $\boldsymbol{\Psi}$ in Eq.(8) can be further derived as

$$\begin{aligned}&p_{\boldsymbol{\Theta}(\boldsymbol{I}),\boldsymbol{\Psi},\boldsymbol{X}(\boldsymbol{I})|W_{\boldsymbol{I}},\boldsymbol{Y}(\boldsymbol{\tau})}(\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\psi},\boldsymbol{x}(\boldsymbol{I})|W_{\boldsymbol{I}}=0,\boldsymbol{Y}(\boldsymbol{\tau})=\boldsymbol{y}(\boldsymbol{\tau}))\propto P(\boldsymbol{\Theta}(\boldsymbol{I})=\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\Psi}=\boldsymbol{\psi},\boldsymbol{X}(\boldsymbol{I})\\&=\boldsymbol{x}(\boldsymbol{I}),W_{\boldsymbol{I}}=0,\boldsymbol{Y}(\boldsymbol{\tau})=\boldsymbol{y}(\boldsymbol{\tau}))\end{aligned} \tag{9}$$

$$\begin{aligned}&\propto\pi_{\boldsymbol{\Psi}}(\boldsymbol{\psi})\times\underbrace{P(\boldsymbol{\Theta}(\boldsymbol{I})=\boldsymbol{\theta}(\boldsymbol{I})|\boldsymbol{\Psi}=\boldsymbol{\psi})}_{\text{1st Part, which is Eq.(2)}}\underbrace{P(\boldsymbol{X}(\boldsymbol{I})=\boldsymbol{x}(\boldsymbol{I})|\boldsymbol{\Theta}(\boldsymbol{I})=\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\Psi}=\boldsymbol{\psi})}_{\text{2nd Part, which is Eq.(3)}}\\&\times\underbrace{P(\boldsymbol{Y}(\boldsymbol{\tau})=\boldsymbol{y}(\boldsymbol{\tau})|\boldsymbol{X}(\boldsymbol{I})=\boldsymbol{x}(\boldsymbol{I}),\boldsymbol{\Theta}(\boldsymbol{I})=\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\Psi}=\boldsymbol{\psi})}_{\text{3rd Part, which is Eq.(4)}}\underbrace{P(W_{\boldsymbol{I}}=0|\boldsymbol{Y}(\boldsymbol{\tau})=\boldsymbol{y}(\boldsymbol{\tau}),\boldsymbol{X}(\boldsymbol{I})=\boldsymbol{x}(\boldsymbol{I}),\boldsymbol{\Theta}(\boldsymbol{I})=\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\Psi}=\boldsymbol{\psi})}_{\text{4th Part, which is Eq.(7)}}\end{aligned} \tag{10}$$

$$\begin{aligned}&=\pi_{\boldsymbol{\Psi}}(\boldsymbol{\psi})\times\exp\Big\{-\frac{1}{2}\Big(\underbrace{\sum_{p=1}^{P}\Big[|\boldsymbol{I}|\log(2\pi)+\log|\mathcal{K}_p^{\Theta}(\boldsymbol{I})|+\|\theta_p(\boldsymbol{I})-\mu_p^{\Theta}(\boldsymbol{I})\|^2_{\mathcal{K}_p^{\Theta}(\boldsymbol{I})^{-1}}\Big]}_{\text{1st Part, which is Eq.(2)}}\\&+\sum_{d=1}^{D}\Big[\underbrace{|\boldsymbol{I}|\log(2\pi)+\log|\mathcal{K}_d^X(\boldsymbol{I})|+\|x_d(\boldsymbol{I})-\mu_d^X(\boldsymbol{I})\|^2_{\mathcal{K}_d^X(\boldsymbol{I})^{-1}}}_{\text{2nd Part, which is Eq.(3)}}+\underbrace{N_d\log(2\pi\sigma_d^2)+\|x_d(\boldsymbol{\tau}_d)-y_d(\boldsymbol{\tau}_d)\|^2_{\sigma_d^{-2}}}_{\text{3rd Part, which is Eq.(4)}}\\&+\underbrace{|\boldsymbol{I}|\log(2\pi)+\log|C_d|+\|\mathbf{f}_{d,\boldsymbol{I}}^{\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\psi}}-\dot{\mu}_d^X(\boldsymbol{I})-{}'\mathcal{K}_d^X(\boldsymbol{I})\mathcal{K}_d^X(\boldsymbol{I})^{-1}\{x_d(\boldsymbol{I})-\mu_d^X(\boldsymbol{I})\}\|^2_{C_d^{-1}}\Big]}_{\text{4th Part, which is Eq.(7)}}\Big)\Big\}\end{aligned} \tag{11}$$

Note that $\boldsymbol{\theta}$ enters into the posterior only through the manifold constraint in Eq.(5)(6)(7). For the equations in the previous sections, Eq.(3) is the prior for the system component (without ODE information), and Eq.(4) is the observation noise, both of which would have no information

where $\|v\|_A^2=v^TAv$, $|\boldsymbol{I}|$ is the cardinality of $\boldsymbol{I}$, and $\mathbf{f}_{d,\boldsymbol{I}}^{\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\psi}}$ is short for the d-th component of $\mathbf{f}(\boldsymbol{x}(\boldsymbol{I}),\boldsymbol{\theta}(\boldsymbol{I}),\boldsymbol{\psi},t_{\boldsymbol{I}})$, and $C_d=\mathcal{K}''^X_d(\boldsymbol{I})-{}'\mathcal{K}_d^X(\boldsymbol{I})\mathcal{K}_d^X(\boldsymbol{I})^{-1}\mathcal{K}'^X_d(\boldsymbol{I})$ is the conditional covariance matrix of $\dot{X}_d(\boldsymbol{I})$ given $X_d(\boldsymbol{I})$.

A deeper look into the above equation reveals that Eq.(9) is the joint probability in Bayesian statistics, and Eq.(10) further decomposes it into parts. The 1st Part (which is Eq.(2)) corresponds to independent GP prior distribution of $\boldsymbol{\Theta}(\boldsymbol{I})$, as the prior of $\boldsymbol{\Theta}(t)$ and $\boldsymbol{\Psi}$ are independent. The 2nd Part (which is Eq.(3)) is the prior of GP on $\boldsymbol{X}(\boldsymbol{I})$, because the prior of $\boldsymbol{X}(\boldsymbol{I})$ is independent from $\boldsymbol{\Theta}(t)$ and $\boldsymbol{\Psi}$. The 3rd Part (which is Eq.(4)) is the level of observation noise, and given the value of underlying true components $\boldsymbol{X}(\boldsymbol{\tau})$, the additive Gaussian observation noise $\boldsymbol{\epsilon}(\boldsymbol{\tau})$ is independent from everything else. The 4th Part (which is Eq.(7)) can be simplified to be the conditional probability of $\dot{\boldsymbol{X}}(\boldsymbol{I})$ given $\boldsymbol{X}(\boldsymbol{I})$ evaluated at $\mathbf{f}(\boldsymbol{x}(\boldsymbol{I}), \boldsymbol{\theta}(\boldsymbol{I}), \boldsymbol{\psi}, t_{\boldsymbol{I}})$. All four parts are multivariate Gaussian distributed. Especially, The 4th Part (which is Eq.(7)) is Gaussian because conditional $\dot{\boldsymbol{X}}(\boldsymbol{I})$ given $\boldsymbol{X}(\boldsymbol{I})$ has a multivariate Gaussian distribution, provided that the GP kernel $\mathcal{K}^X$ is twice differentiable.

We choose Matern kernel with degree of freedom $\nu = 2.01$ for both $\boldsymbol{\Theta}(t)$ and $\boldsymbol{X}(t)$ to guarantee a differentiable GP that allows more flexible patterns:

$$\mathcal{K}_\nu(l) = \phi_1^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu}\frac{l}{\phi_2})^\nu K_\nu(\sqrt{2\nu}\frac{l}{\phi_2}), \quad l = |s-t| \quad (12)$$

where K_ν denotes the modified Bessel function of the second kind. In this case, $'\mathcal{K} = \frac{\partial}{\partial s}\mathcal{K}(s,t)$, $\mathcal{K}' = \frac{\partial}{\partial t}\mathcal{K}(s,t)$, and $\mathcal{K}'' = \frac{\partial^2}{\partial s \partial t}\mathcal{K}(s,t)$ are all well-defined.

5 Algorithm

This section provides a detailed computational scheme of TVMAGI, including the hyper-parameter settings. The implementation is available on GitHub. Overall, the Maximum A Posteriori (MAP) of $\boldsymbol{X}(\boldsymbol{I})$, $\boldsymbol{\Theta}(\boldsymbol{I})$, and $\boldsymbol{\Psi}$ is obtained by optimization, while the posterior mean/interval is obtained by Hamiltonian Monte Carlo. To set the hyper-parameters and initiate the optimizer, we introduce a multi-stage approach in the algorithm. The advantages of the multi-stage mechanism are discussed in Supplementary Material Section A.16.

5.1 Initialization and inference of the mean

At the first stage, we impose a GP only on $\boldsymbol{X}(t)$ and substitute the time-varying $\boldsymbol{\theta}(t)$ with its unknown mean $\boldsymbol{\mu}^\Theta$ in the entire model. This formulation ignores the time-varying property of $\boldsymbol{\theta}(t)$ and treats it as time-constant, which fits in the time-constant parameters inference framework of Yang et al. (2021). As such, we can use MAGI package ((Yang et al. 2021)) to obtain point estimates for the parameters and system components, denoted as $\hat{\boldsymbol{\mu}}^\Theta$, $\boldsymbol{\psi}^{(0)}$, and $\boldsymbol{x}(\boldsymbol{I})^{(0)}$. The $\hat{\boldsymbol{\mu}}^\Theta$ is subsequently used as the prior mean value for the time-varying $\boldsymbol{\theta}(t)$ in an empirical Bayes fashion, and will be plugged in Eq.(11). The $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{x}(\boldsymbol{I})^{(0)}$ will be used as the initial values for $\boldsymbol{\psi}$ and $\boldsymbol{x}(\boldsymbol{I})$ in the later MAP optimization.

The hyper-parameters $(\phi^X_{1,d}, \phi^X_{2,d})$ for kernel $\mathcal{K}^X_d$ in Eq.(12) and the noise level σ for each system component $X_d, d = 1, ..., D$, are also estimated in MAGI package, using the Gaussian process smoothing marginal likelihood ((Yang et al. 2021)). The MAGI estimated noise level $\boldsymbol{\sigma}^{(0)}$ will serve as initial value for later joint MAP optimization.

5.2 Point-wise inference of the time-varying parameters

At the second stage of TVMAGI, we obtain an initial estimate for time-varying $\boldsymbol{\theta}(\boldsymbol{I})$ by removing the smoothing GP prior. That is, we maximize the partial posterior Eq.(13) conditioning on $\boldsymbol{\Theta}(\boldsymbol{I})$, without considering Part 1 Eq.(2):

$$\begin{aligned}
&\tilde{\boldsymbol{x}}(\boldsymbol{I}), \tilde{\boldsymbol{\theta}}(\boldsymbol{I}), \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\sigma}} \\
&= \underset{\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\psi},\sigma}{\arg\max}\, p_{\boldsymbol{\Psi},\boldsymbol{X}(\boldsymbol{I})|W_{\boldsymbol{I}},\boldsymbol{Y}(\boldsymbol{\tau}),\boldsymbol{\Theta}(\boldsymbol{I})} \\
&(\boldsymbol{\psi}, \boldsymbol{x}(\boldsymbol{I})|W_{\boldsymbol{I}} = 0, \boldsymbol{Y}(\boldsymbol{\tau}) = \boldsymbol{y}(\boldsymbol{\tau}), \boldsymbol{\Theta}(\boldsymbol{I}) = \boldsymbol{\theta}(\boldsymbol{I})) \\
&\propto \pi_{\boldsymbol{\Psi}}(\boldsymbol{\psi}) \times \underbrace{P(\boldsymbol{X}(\boldsymbol{I}) = \boldsymbol{x}(\boldsymbol{I}))}_{\text{2nd Part Eq.(3)}} \qquad (13) \\
&\times \underbrace{P(\boldsymbol{Y}(\boldsymbol{\tau}) = \boldsymbol{y}(\boldsymbol{\tau})|\boldsymbol{X}(\boldsymbol{I}) = \boldsymbol{x}(\boldsymbol{I}))}_{\text{3rd Part Eq.(4)}} \times \\
&\underbrace{P(\dot{\boldsymbol{X}}(\boldsymbol{I}) = \mathbf{f}(\boldsymbol{x}(\boldsymbol{I}), \boldsymbol{\theta}(\boldsymbol{I}), \boldsymbol{\psi}, t_{\boldsymbol{I}})|\boldsymbol{X}(\boldsymbol{I}) = \boldsymbol{x}(\boldsymbol{I}))}_{\text{4th Part Eq.(7)}} \qquad (14)
\end{aligned}$$

The optimization is initialized at $\boldsymbol{x}(\boldsymbol{I})^{(0)}$, $\boldsymbol{\psi}^{(0)}$, $\boldsymbol{\sigma}^{(0)}$, and the $\boldsymbol{\theta}(\boldsymbol{I})$ is initialized at $\hat{\boldsymbol{\mu}}^{\boldsymbol{\Theta}}$. We denote the optimized $\boldsymbol{\theta}(\boldsymbol{I})$ as $\tilde{\boldsymbol{\theta}}(\boldsymbol{I})$, and $\boldsymbol{x}(\boldsymbol{I})^{(0)}$, $\boldsymbol{\psi}^{(0)}$, $\boldsymbol{\sigma}^{(0)}$ are updated to a new optimum $\tilde{\boldsymbol{x}}(\boldsymbol{I})$, $\tilde{\boldsymbol{\psi}}$ and $\tilde{\boldsymbol{\sigma}}$. We call $\tilde{\boldsymbol{\theta}}(\boldsymbol{I})$ the point-wise estimate since there is no requirement on the smoothness or continuity of $\tilde{\boldsymbol{\theta}}(t)$ on $\boldsymbol{I}$. Although wiggling and possibly overfitting the data, the point-wise estimate $\tilde{\boldsymbol{\theta}}(\boldsymbol{I})$ captures the trend of parameter changes, which provides information to set the hyper-parameters of GP kernels $\mathcal{K}^\Theta_p$ for $\boldsymbol{\Theta}(t)$.

5.3 GP hyper-parameters for time-varying ODE parameters

Length scale parameter ϕ^Θ controls how fast $\boldsymbol{\theta}(t)$ could change. Provided the point-wise estimate $\tilde{\boldsymbol{\theta}}(\boldsymbol{I})$, we use Gaussian Process smoothing method to set the hyper-parameters $\phi^\Theta_{1,p}, \phi^\Theta_{2,p}$ of GP kernels $\mathcal{K}^\Theta_p$ in Eq.(12). We shall treat $\tilde{\boldsymbol{\theta}}(\boldsymbol{I})$ as observations of $\boldsymbol{\Theta}(\boldsymbol{I})$, and operate on each dimension of time-varying ODE parameters separately.

Recall the prior $\Theta_p(\boldsymbol{I}) \sim \mathcal{GP}(\hat{\mu}^\Theta_p, \mathcal{K}^\Theta_p(\boldsymbol{I}, \boldsymbol{I}))$, where the mean $\hat{\mu}^\Theta_p$ is obtained in Sect. 5.1. We use the empirical Bayes approach again to set $\phi^\Theta_{1,p}, \phi^\Theta_{2,p}$ by maximizing its posterior

density at $\tilde{\theta}_p(\boldsymbol{I})$:

$$\hat{\phi}^{\Theta}_{1,p}, \hat{\phi}^{\Theta}_{2,p} = \underset{\phi_1,\phi_2}{\arg\max}\, \pi_{\boldsymbol{\Phi}_p}(\boldsymbol{\phi}) P(\tilde{\theta}_p(\boldsymbol{I})|\boldsymbol{\phi}) \tag{15}$$

where $\tilde{\theta}_p(\boldsymbol{I})|\boldsymbol{\phi} \sim \mathcal{N}(\hat{\mu}^{\Theta}_p, \mathcal{K}(\boldsymbol{\phi}) + \text{diag}(\delta^2))$, the δ is the nuisance parameter governing the induced noise in point-wise estimate $\tilde{\theta}_p(\boldsymbol{I})$, and the $\pi_{\boldsymbol{\Phi}_p}(\cdot)$ is the hyper-prior. In practice, the hyper-prior $\pi_{\boldsymbol{\Phi}_p}(\cdot)$ is often set to be uniform on a reasonable interval depending on the context to ensure desired level of smoothness for the time-varying ODE parameter $\theta_p(t)$.

Once the hyper-parameters $\hat{\phi}^{\Theta}_{1,p}, \hat{\phi}^{\Theta}_{2,p}$ are estimated through maximum marginal likelihood, we fix the hyper-parameters at their optimized values in all subsequent posterior inference. This is also called modularization and is justified in Bayarri et al. (2009).

5.4 Maximum a posteriori (MAP) optimization

All hyper-parameters are now set and will be held as constant when optimizing Eq.(11) to get the MAP, with initial values $\boldsymbol{\theta}(\boldsymbol{I})^{(0)} = \tilde{\boldsymbol{\theta}}(\boldsymbol{I})$, $\boldsymbol{x}(\boldsymbol{I})^{(0)} = \tilde{\boldsymbol{x}}(\boldsymbol{I})$, $\boldsymbol{\psi}^{(0)} = \tilde{\boldsymbol{\psi}}$ and $\boldsymbol{\sigma}^{(0)} = \tilde{\boldsymbol{\sigma}}$, all from Sect. 5.2. The joint posterior function Eq.(11) of $\boldsymbol{x}(\boldsymbol{I})$, $\boldsymbol{\theta}(\boldsymbol{I})$, $\boldsymbol{\psi}$ and $\boldsymbol{\sigma}$ is optimized with Adam optimizer (Kingma and Ba 2014) in PyTorch to get the MAP estimation of TVMAGI. Finally, to mitigate the potential issue of Adam optimizer converging to local optimum, we suggest trying multiple initial values, including starting $\boldsymbol{x}(\boldsymbol{I})$ at linear interpolations from the observations $\boldsymbol{y}(\boldsymbol{\tau})$.

5.5 Interval estimation of parameters

In addition to the MAP point estimate, we also quantify the parameter uncertainty in TVMAGI using posterior samples. In particular, we sample the posterior function Eq(11) using Hamilton Monte Carlo (HMC), while holding all the hyper-parameters at the same constant value as in Sect. 5.4. Details about the HMC algorithm can be found in Supplementary Material Section A.3. The interested reader may refer to Neal (2011) for more thorough introduction to HMC. Specifically in all illustration examples of this paper, we set step size $\epsilon = 10^{-5}$, number of leap-frog steps $L = 100$, sample size 8000, burn-in ratio 0.5, and the HMC is initialized at the MAP estimate.

6 Benchmark methods and evaluation metrics

6.1 Benchmark methods

We compare our method with two common approaches for time-varying parameter inference in ODE: numerical integration methods, represented by Runge–Kutta method (Lapidus and Seinfeld 1971), and Bayesian filtering methods, represented by Ensemble Adjustment Kalman Filter (EAKF) (Shaman and Karspeck 2012), which has been used in estimating the influenza disease spread SIRS model parameter (Shaman and Karspeck 2012) and studying time-varying fatality rate In COVID-19 disease spread modeling (Yang et al. 2020). Supplementary Material Section A.4 provides the review of two approaches and some additional theoretical discussion about the limitations and the statistical rigor of the benchmark methods for ODE inference when time-varying parameters and time-constant parameters co-exist.

6.2 Evaluation metrics

To assess the quality of the parameter estimates and the system recovery, we consider two metrics based on root mean squared error (RMSE). First, we examine the accuracy of the parameter estimates, using *parameter RMSE*. For the time-constant parameters, we directly calculate the RMSE of the parameter estimates to the true parameter value across simulations. For the time-varying parameters, we additionally average over discretization set $\boldsymbol{I}$ for the RMSE. Second, we examine the system recovery, using *trajectory RMSE*. Due to the potential identifiability issue that different parameters can give similar system observations, we measure how well the system components are recovered as another independent evaluation. To calculate the trajectory RMSE, we use numerical integration to reconstruct the trajectory based on the TVMAGI inferred parameters and initial conditions. The RMSE of the reconstructed trajectory to the true system is then calculated at observation time points.

We emphasize that the numerical integration is only used for evaluation purpose, and throughout our TVMAGI approach, no numerical integration is ever needed. For better distinction, we refer to the MAP of $\boldsymbol{x}(\boldsymbol{I})$ directly from TVMAGI as the *inferred trajectory*, and refer to the numerically integrated $\boldsymbol{x}(t)$ based on the TVMAGI inferred parameters and initial conditions as the *reconstructed trajectory*.

To assess the quality of the interval estimates, we consider the Frequentist coverage of our posterior intervals. For the time-constant parameters, we directly calculate the proportion of repeated simulations where our posterior interval covers the truth. For the time-varying parameters, we additionally average over discretization set $\boldsymbol{I}$ for the coverage. The coverage of the inferred trajectory can be similarly calculated, averaging over discretization set $\boldsymbol{I}$. We do not compute the coverage of the reconstructed trajectory as it will require numerical solver for each posterior sample of the parameters and initial conditions.

7 Results

We illustrate the accuracy and efficiency of TVMAGI through three realistic simulation studies of ODE models in epidemiology, ecology, and system biology. We begin with a disease compartmental model that demonstrates the effectiveness of TVMAGI for problems with partially observed system component(s). We then use an ecology example to show how TVMAGI can mitigate the identifiability issue through the informative GP prior that favors smoother time-varying parameters. Lastly, we apply TVMAGI on a system biology example with non-stationary rapid-changing time-varying parameters, and presents TVMAGI's competitive performance with one additional tailor-made benchmark method for such ODE.

7.1 SEIRD model

Consider a COVID-19 cases/deaths modeling using an infectious disease Susceptible-Exposed-Infectious-Recovered-Deceased (SEIRD) compartmental ODE model (Hethcote 2000; Hao et al. 2020), where the entire population is classified into S, E, I, R, D components, and any transitions from one state to another state (i.e., the disease spreading dynamics) are modeled as ODE:

$$\begin{aligned} &\frac{dS}{dt} = -\frac{\beta I S}{N}, \quad \frac{dE}{dt} = \frac{\beta I S}{N} - v^e E, \\ &\frac{dI}{dt} = v^e E - v^i I, \\ &\frac{dD}{dt} = v^i I \cdot p^d \end{aligned} \tag{16}$$

N is the total population, and the cumulative recovered population is $R = N - S - E - I - D$. The S, E, I and D denote the susceptible, exposed, infected population and cumulative death respectively. The 4 parameters of interest are investigated: rate of contact by an infectious individual (β), rate of transferring from state of exposed to infectious (v^e), rate of leaving infectious period (v^i) and fatality rate (p^d). During a pandemic, parameters in the SEIRD model can evolve over time due to pharmaceutical and non-pharmaceutical interventions. We assume that β is time-varying due to the mutation of disease and policy interventions during a specific time; p^d is time-varying depending on the sufficiency of medical treatments; v^e is time-varying due to the different levels of public awareness or complacency, and v^i is assumed to be unknown time-constant parameter to avoid identifiability issues.

In the experiment we set $v^i = 0.1$, $\beta_t = 1.8 - \cos(\pi t/8)$, $v^e_t = 0.1 - 0.02\cos(\pi t/8)$, $p^d_t = 0.05 + 0.025\cos(\pi t/8)$, and focus on a time horizon of 32 days. The initial values of four components are set as (100000, 100, 50, 50) for (S, E, I, D). We assume S, I, D are observed on daily frequency with log-normal multiplicative observation noise at 3% level. The exposed population E is assumed to be only sparsely observable at 3% noise level, with one observation per two days, due to the high cost of data acquisition from sampling test. Such pandemic settings (Dong et al. 2020; Mwalili et al. 2020) capture the periodic fluctuation of parameters often observed in the real world.

We apply TVMAGI on a log-transformed system (by taking the log of populations in each of the S, E, I, D state) over 100 simulation datasets, with 2 discretizations per day. Figure 1 shows the results of parameter inference and the TVMAGI reconstructed trajectory $\boldsymbol{X}(\boldsymbol{I})$ of the ODE system. The parameter RMSE and trajectory RMSE introduced in Sect. 6.2 are presented in Table 1, where TVMAGI is shown to be more accurate than Runge–Kutta or EAKF.

Table 1 Accuracy comparison for the SEIRD model based on 100 simulation datasets. The mean of RMSE is reported first with the standard deviation across 100 replications followed after ± for the parameters and the reconstructed trajectories. The last column is the coverage of interval estimates for the parameters and the inferred trajectories. The last row shows the computing time (in seconds) needed to obtain point estimates from all methods

		Point Estimate						TVMAGI Posterior Samples	
	RMSE	TVMAGI-MAP		Runge–Kutta		EAKF		Posterior Mean RMSE	Interval Coverage
Parameter	β	**0.114**	**0.039**	0.178	0.094	0.706	0.010	0.110 ± 0.043	98.2%
	v^e	**0.009**	**0.010**	0.051	0.030	0.057	0.003	0.007 ± 0.005	97.4%
	v^i	0.005	0.003	**0.004**	**0.003**	0.151	0.008	0.007 ± 0.004	91.0%
	p^d	**0.019**	**0.029**	0.083	0.073	0.039	0.003	0.011 ± 0.008	98.0%
Trajectory	S	**581.7**	**272.1**	1084.8	195.3	3868.3	132.2	615.3 ± 294.5	98.8%
	E	**704.7**	**218.3**	951.7	142.3	5376.1	496.6	660.7 ± 202.3	96.6%
	I	**439.0**	**140.4**	556.2	90.6	3167.0	404.9	415.0 ± 158.7	96.4%
	D	38.3	4.9	**33.3**	**5.0**	907.3	48.6	14.0 ± 5.2	94.2%
Computing Time (s)		1006.7	115.54	2904.4	195.5	7.3	0.4	-	

The bold indicates that the performance is the best among all methods

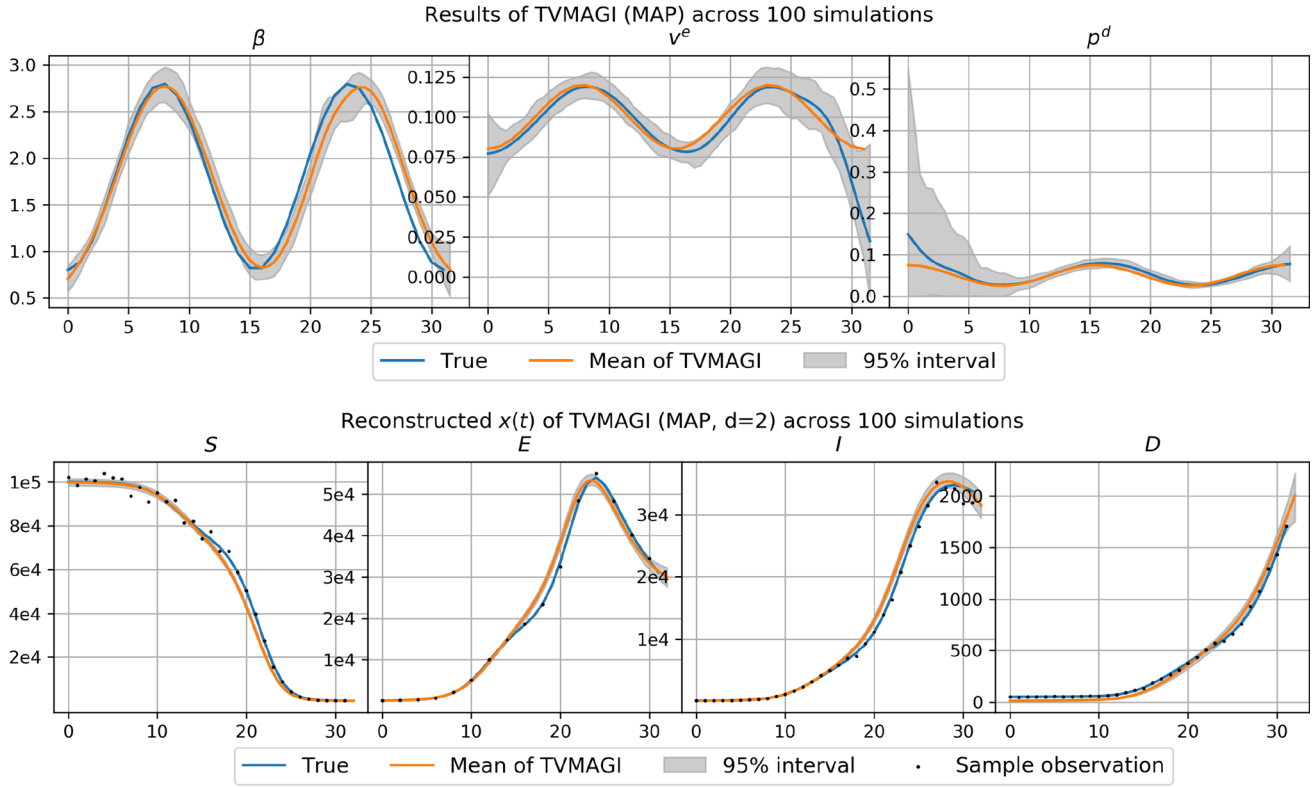


Fig. 1 Results of parameter inference (upper) and reconstructed trajectory (lower) of TVMAGI in 100 simulated datasets for SEIRD model. The mean and the 95% interval here refer to the point estimates across 100 simulated datasets. One sample simulation dataset is also presented to visualize the noise level and observation schedule.

For point estimates, Fig. 1 and Table 1 show that, even when the exposed population is sparsely observed, TVMAGI is still capable of providing good results of inference. As the most important parameter when assessing the spread of disease, β_t can be accurately and robustly inferred. v^i can also be accurately inferred as constant. p_t^d has larger variability at the start, as initial deaths are too few to provide enough information. In comparison, the variability of v_t^e inference increases at the end of the period, because susceptible population has decreased to nearly zero while infectious population reaches plateau. Despite variations in the inferred parameters, the inferred system trajectories are all very close to the truth, confirming the intuition that the system is possibly less sensitive to p^d in earlier state and v^e in later stage. Supplementary Material Section A.5 has the visual illustration for Runge-Kutta or EAKF, and their accuracy is far from satisfactory: Runge-Kutta method will overfit the observation noise, and EAKF reconstructed trajectory completely misses the truth. We also tried to increase the computation budget for EAKF, by increasing the discretization level and the ensemble size. The result is shown in Supplementary Material Section A.17. It can be seen that the performance of EAKF would not improve given the increased computation budget, suggesting that the limitation of EAKF is inherent, as discussed in Supplementary Material Section A.8.

For Interval estimates, Figure A10 in Supplementary Material Section A.9 gives a visual illustration for 10 sample datasets. The coverage of Posterior Interval across 100 simulated datasets is included in Table 1. The emperical coverage of the interval is reasonable around the 95% nominal value. The intervals are wider for p^d at the starting time, and wider for v^e at the ending time, which are consistent with the intuition about their sensitivity discussed above. More interval estimation results are available in Supplementary Material Section A.9.

On the computational cost, Table 1 also shows that TVMAGI is much faster than the Runge–Kutta numerical integration methods. EAKF is fast, but gives unreliable results (see Supplementary Material Section A.8 for more discussion on the reliability of EAKF).

7.2 Lotka–Volterra model

Lotka-Volterra (LV) model (a.k.a. predator–prey model) is widely used to describe population fluctuation of predators and preys and their interactions in the ecosystem (Goel et al.

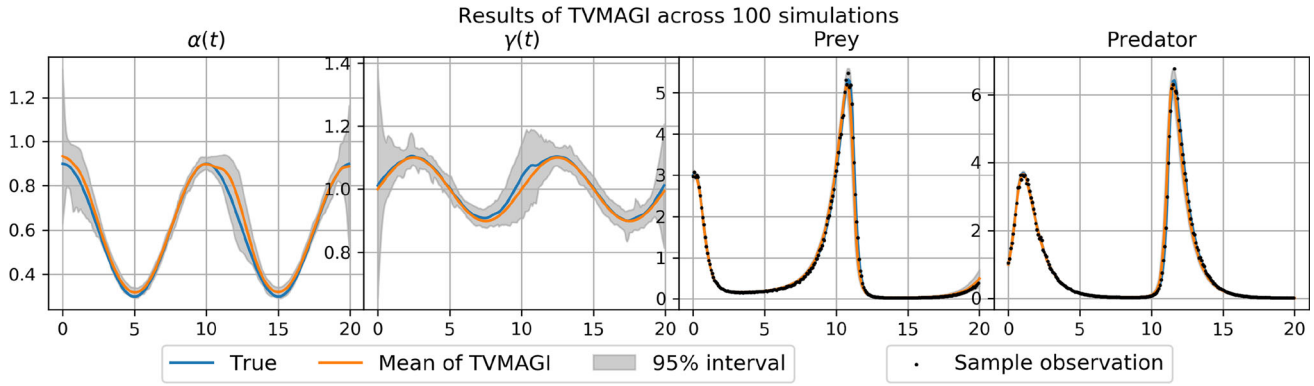


Fig. 2 Comparison of inferred $\boldsymbol{\theta}(t)$ and reconstructed $\boldsymbol{X}(t)$ of LV model. The mean and the 95% interval here refer to the point estimates across 100 simulated datasets. One sample simulation dataset is also plotted to visualize the noise level.

Table 2 Accuracy comparison of estimated parameters and reconstructed trajectory in the LV model based on 100 simulation datasets. See legend of Table 1 for detailed description

		Point Estimate						TVMAGI Posterior Samples	
	RMSE	TVMAGI		Runge–Kutta		EAKF		Posterior mean RMSE	Interval Coverage
Parameter	$\alpha(t)$	**0.0395**	**0.0302**	0.1155	0.0920	0.2330	0.1589	0.0450 ± 0.0346	40.5%
	β	**0.0154**	**0.0098**	0.0480	0.0069	0.1041	0.0621	0.0187 ± 0.0122	62.0%
	δ	**0.0156**	**0.0111**	0.0166	0.0058	0.1605	0.0831	0.0160 ± 0.0119	58.0%
	$\gamma(t)$	**0.0304**	**0.0226**	0.0863	0.0756	0.0974	0.1847	0.0341 ± 0.0201	46.0%
Traj.	x (prey)	0.0606	0.0550	**0.0314**	**0.0085**	0.4701	0.0903	0.0838 ± 0.0568	69.1%
	y (predator)	0.0813	0.0622	**0.0384**	**0.0131**	0.2773	0.0646	0.0989 ± 0.0714	62.2%
Computing Time (s)		910.8	113.7	2042.1	85.2	22.7	1.1	-	

The bold indicates that the performance is the best among all methods

1971). With the introduction of time-varying parameters, the system becomes weakly identifiable during certain time range, which creates a challenge in the inference. Specifically, the ODE system is characterized as:

$$\frac{dx}{dt} = \alpha_t x - \beta xy, \quad \frac{dy}{dt} = \delta xy - \gamma_t y \tag{17}$$

where x and y denote the population of preys and predators. α_t indicates the birth rate of the prey and γ_t denotes the death rate of the predator, both of which are assumed to fluctuate according to seasonality. β and δ describe the interaction relationships between predators and preys, and are assumed constant. We set the parameters $\beta = 0.75$, $\delta = 1$, $\alpha_t = 0.6 + 0.3\cos(\pi t/5)$, and $\gamma_t = 1 + 0.1\sin(\pi t/5)$. The time is measured on a yearly basis, and data for 20 years are generated with monthly observations contaminated by 3% multiplicative log-normal noise. The initial values of predators and preys are 1 and 3, as an ideal ratio in real ecology systems (Donald and Stewart Anderson 2003).

Figure 2 shows the estimated time-varying parameters and the reconstructed trajectory $\boldsymbol{X}(\boldsymbol{I})$, with parameter RMSE and trajectory RMSE presented in Table 2. Our recovered system components x and y are very close to the truth, despite the weak identifiability of the parameter α_t and γ_t when the x and y are at peak (year=12). Most notably, α_t could deviate from the truth in the attempt to best fit the observed noisy data at the peak of x_t, resulting in a biased inference of the time-varying parameters at the weakly identifiable time points, although all deviations are still within the range of smoothness constraints on α_t. Nevertheless, both TVMAGI inferred system components x and y are still accurate.

Comparing with benchmark models in Table 2, TVMAGI gives the most accurate parameter inference thanks to the GP smoothing prior that mitigates the identifiability issue. The numerical method of Runge–Kutta gives better trajectory inference, but it cannot handle the identifiability issue in the parameters. The coverage from TVMAGI is not ideal, possibly due to the bias in $\alpha(t)$ estimate and the variance in $\gamma(t)$ estimate – if the GP smoothing prior is too strong, the point estimates will be biased, and if the GP smoothing prior is too weak, the point estimates will have large variance (see Supplementary Material Figure A12. The comparison on computational cost again demonstrates the expected advantage of TVMAGI over Runge–Kutta, while EAKF is the fastest method with the worst accuracy.

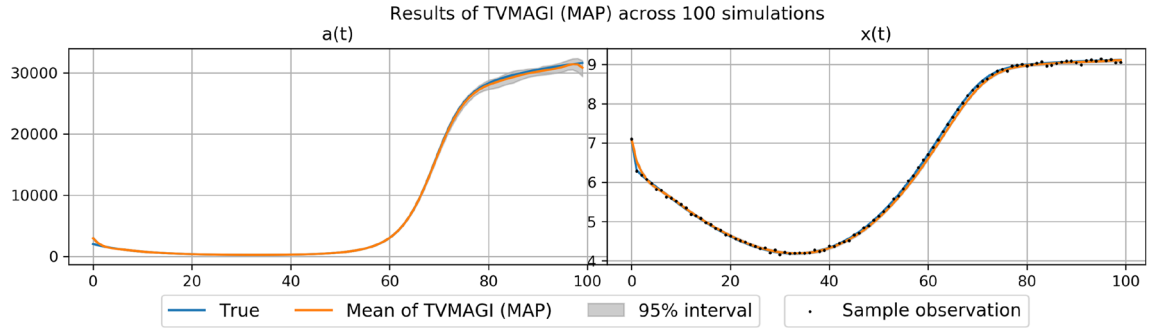


Fig. 3 TVMAGI inferred $a(t)$ and reconstructed $\boldsymbol{X}(t)$ of HIV model.

Table 3 Accuracy comparison of estimated parameters and trajectories in the HIV model based on 100 simulation datasets. See legend of Table 1 for detailed description

RMSE	Point Estimate TVMAGI		Runge–Kutta		EAKF		ELE		TVMAGI Posterior Samples Posterior mean RMSE	Interval Coverage
$a(t)$	**281.9**	**71.8**	695.7	50.9	818.6	54.9	291.5	48.4	359.0 ± 127.8	74.1%
$x(t)$	0.057	0.002	**0.038**	**0.003**	0.181	0.004	0.075	0.003	0.069 ± 0.004	65.9%
Computing Time (s)	897.6	72.1	1940.2	79.7	5.4	0.3	10.7	0.1	-	-

The bold indicates that the performance is the best among all methods

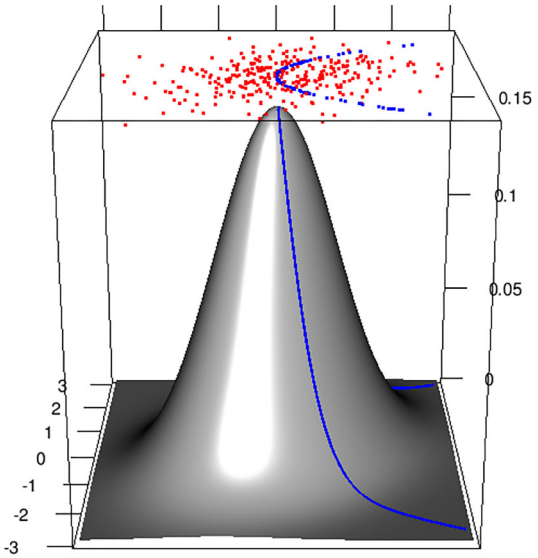


Fig. 4 Illustration of the manifold constraint on bi-variate Gaussian. Left panel: blue dots are the samples from joint density of (Z_1, Z_2) without the manifold constraint $W = 0$. The red curve is the manifold constraint $W = 0$ to be imposed. Right panel: blue dots are the samples from $(Z_1, Z_2)|W = 0$ where all points lie on the parabola. The density is proportional to the original bi-variate Gaussian but only on the parabola curve.

This example illustrates the performance of TVMAGI in the presence of weak identifiability – the inferred time-varying parameters at the weakly identified time points could subject to deviation from the truth, although the parameters are smooth and still fit the observed data well.

7.3 HIV model

In this example, we compare TVMAGI with a state-of-the-art two-stage Efficient Local Estimation (ELE) method proposed by Chen and Wu (2008) in an HIV dynamic model that they studied. This is a challenging case for GP modeling as the true time-varying parameter has non-periodic non-stationary trends with rapid changes (Perelson et al. 1996; Huang et al. 2003). To use the ELE method of Chen and Wu (2008), the ODE system must fit in the linear form of Eq.(18):

$$X'(t) = \sum_{i=1}^{d} a_i(t) Z_i(t) - cX(t) \tag{18}$$

where $Z_i(t)$ is the known covariate, and $a_i(t)$ is the unknown time-varying coefficient. For benchmark comparison, we treat $a_1(t)$ and $a_2(t)$ in Eq.(18) as unknown time-varying parameters for TVMAGI. Detailed illustration of HIV model formulation is provided in Supplementary Material A.15.

Figure 3 shows the TVMAGI inferred parameter $a(t) = \sum_{i=1}^{d} a_i(t) Z_i(t)$ and the reconstructed trajectory $X(t)$. The parameter/trajectory RMSEs of TVMAGI and the benchmark methods are reported in Table 3. TVMAGI has a small advantage over the state-of-the-art method on HIV model inference of $X(t)$. Further visual comparison to benchmark methods (Supplementary Material Figure A9) shows that TVMAGI is slightly more accurate at the beginning phase of the system, which is in fact the most challenging phase for HIV inference as viral load drops sharply due to the drug

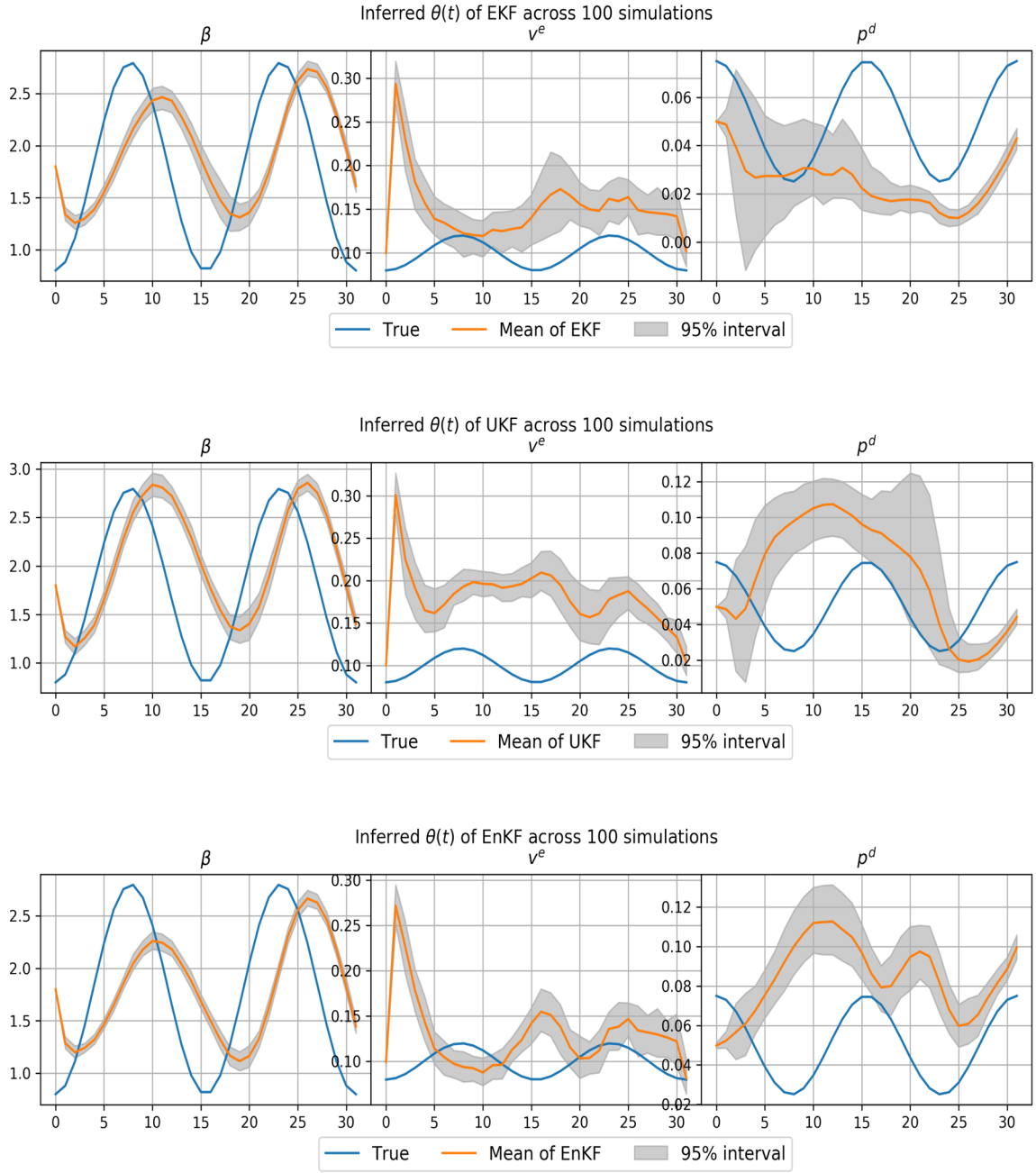


Fig. 5 Inferred $\theta(t)$ of EKF, UKF and EnKF approach for SEIRD model. The inferred parameters cannot capture the ODE structure, thus yielding inaccurate reconstructed trajectory.

effect. TVMAGI also achieves competitive inference result on $a(t)$, which is of clinical importance for the generation rate of HIV virus (Cao et al. 2012). The TVMAGI posterior interval coverage is less ideal because of the decreased accuracy in $a(t)$ towards the ending period (Supplementary Material Figure A13). Most importantly, while the benchmark method requires a highly restricted form of ODE formulation, TVMAGI assumes no specific form of ODE equations, and is thus applicable for general ODE systems, albeit with longer computing time.

Overall in this example, we compare TVMAGI with an additional benchmark method that can only be applied to the ODEs with a specific form, where TVMAGI is shown to provide competitive inference accuracy while having much more general applicability. The application in HIV model also illustrates that TVMAGI could work well with non-stationary trends in the time-varying parameters, where the time-varying parameters is not periodic and have rapid changes in part of the time horizon.

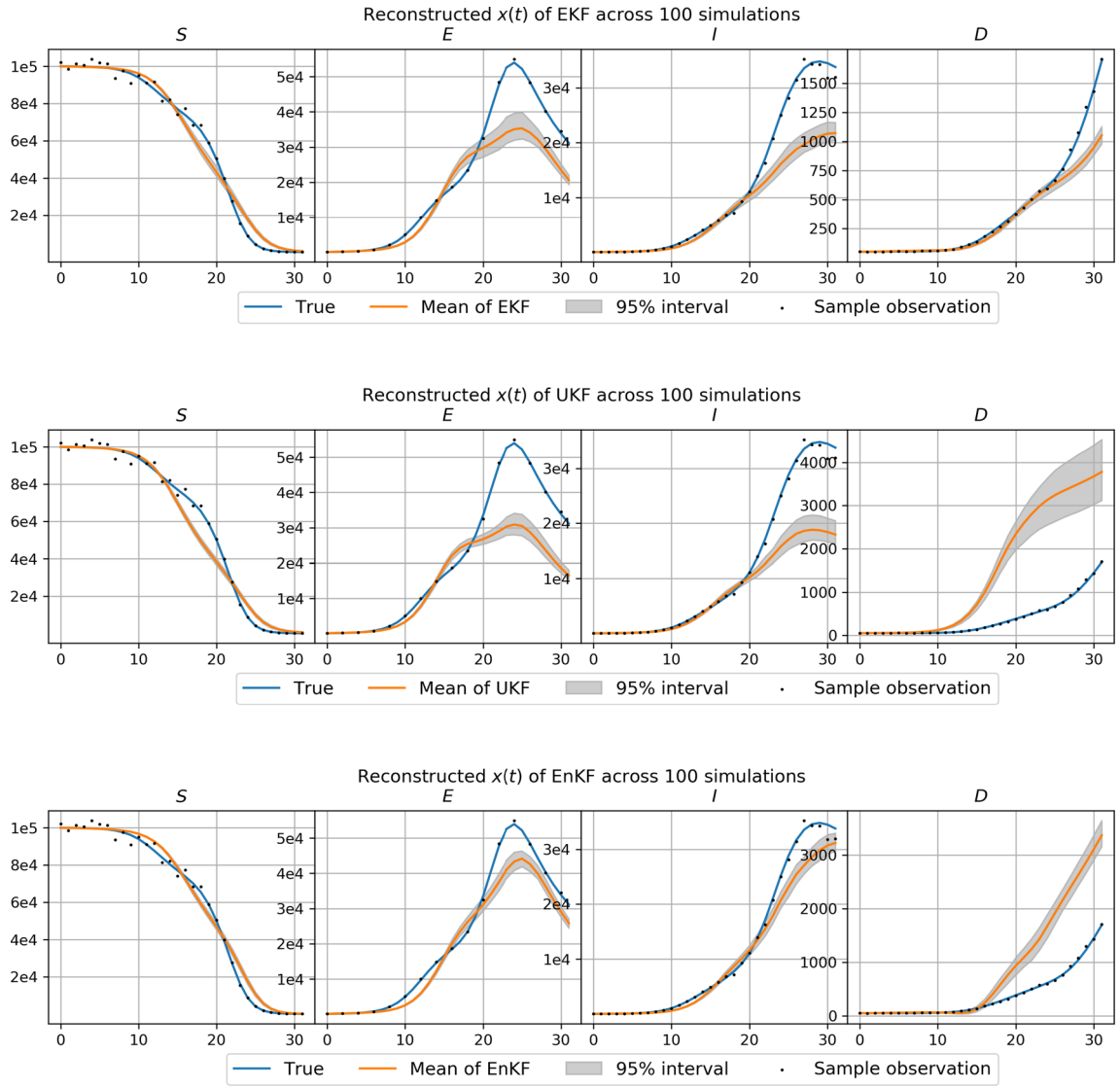


Fig. 6 Reconstructed trajectory of EKF, UKF and EnKF approaches for SEIRD model. The accuracy is far from satisfactory.

8 Sensitivity analysis

We conduct three sensitivity analysis to show the robustness of our approach: the number of discretization, the selection of GP kernel, and the mis-specified time-varying parameters. Detailed discussion is provided in Supplementary Material Section A.10-A.14, along with tables and visualizations of SEIRD model results.

9 Discussion

In this paper, we introduce a Bayesian approach, TVMAGI, for time-varying parameters inference in ODE dynamic systems. TVMAGI models time-varying parameters and system components as Gaussian process, and is constrained to have the derivative processes satisfy the ODE dynamics. We show that TVMAGI is statistically principled and illustrate its general applicability through three simulation examples. Results have shown that TVMAGI yields accurate and robust parameter inference from noisy observations, with reasonable interval estimates as well. Moreover, TVMAGI can mitigate the identifiability issue and the over-fitting issue in the time-varying parameters using the informative GP smoothing prior. TVMAGI is also generally applicable in the presence of missing observations.

TVMAGI is more accurate than the benchmark methods because TVMAGI addresses the challenges of the numerical integration method and the Bayesian filtering method for ODE time-constant and time-varying parameter inference. Numerical integration methods are the gold standard for the ODE parameter inference when all parameters are

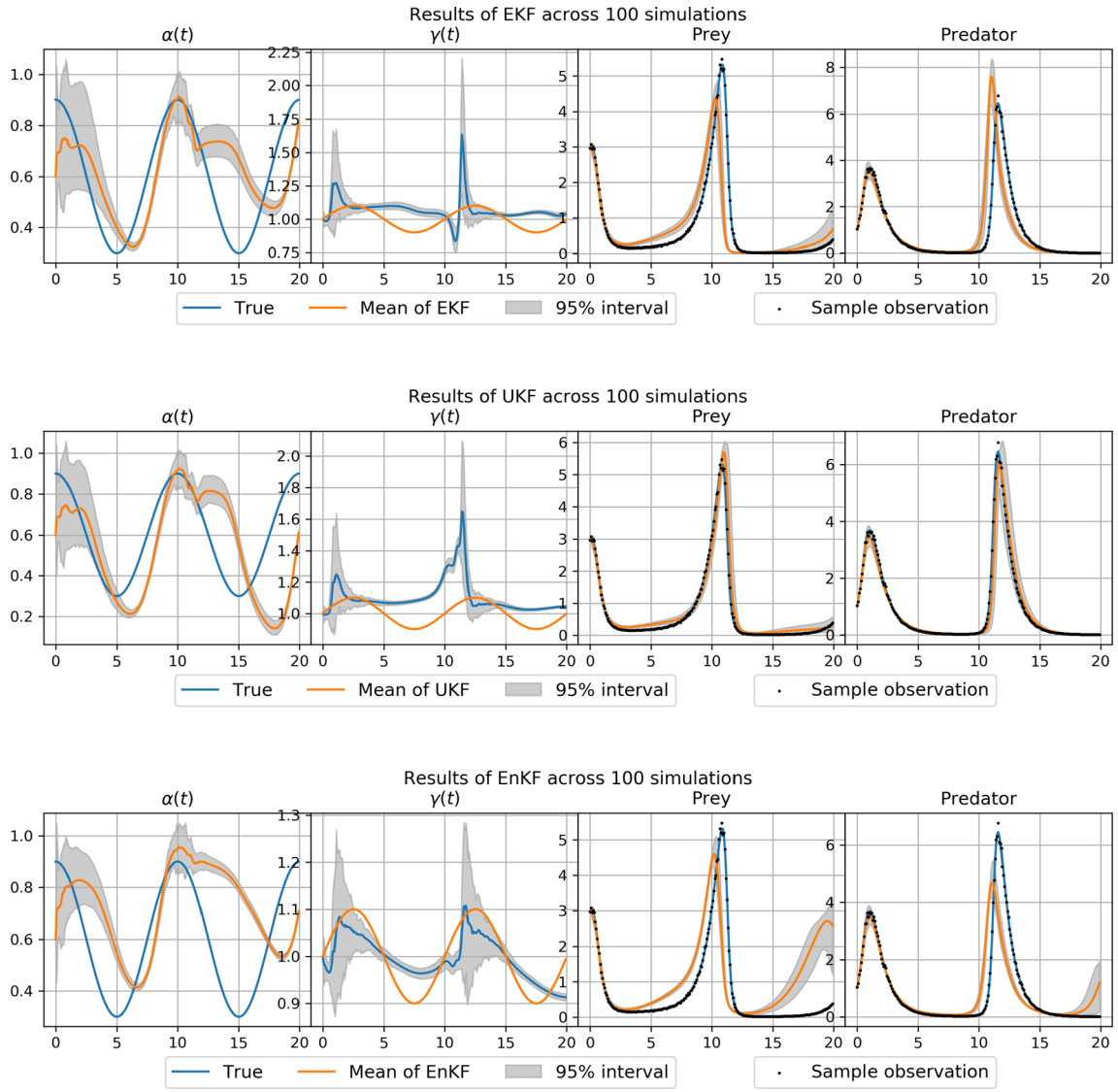


Fig. 7 Inferred parameter and reconstructed trajectory of EKF, UKF and EnKF approaches for LV model. The accuracy is far from satisfactory.

time-constant. However, with the presence of time-varying parameters, due to the lack of smoothness structure on $\boldsymbol{\theta}(t)$, the inferred time-varying parameters from Runge-Kutta will overfit the noisy observation data, resulting in volatile $\boldsymbol{\theta}(t)$ with little information about the true trends. The Bayesian filtering approach, on the other hand, cannot infer time-constant parameter $\boldsymbol{\psi}$ because the update in $\boldsymbol{\psi}$ is not permissible in a state-space model fashion. We can nevertheless enforce an update on the time-constant parameter, but there will be no guarantee on the accuracy of the reconstructed trajectory. The ELE two-stage approach relies on a regression technique that can only be used if the ODE has linear dependency on the time-varying parameters. Therefore, TVMAGI is the only approach that is theoretically sound, practically accurate, and generally applicable for the ODE inference problem when time-constant and time-varying parameters co-exist.

On the computational time comparison, TVMAGI has notable advantage of reduced computation cost compared to numerical integration method, while the inference is more accurate compared to the fast-yet-unreliable Bayesian filtering methods. Even for the three small-sized problems in this paper, TVMAGI is more than twice as fast than the numerical integration method of Runge–Kutta with better accuracy. When dealing with large-scale system, the gain in computational time is likely to be even larger, as TVMAGI computational time would scale linearly as the dimension of system components grow, while inference with numerical integration method typically scales super linearly. Therefore, TVMAGI has strong potential in large-scale systems, where numerical integration is expensive.

There are two settings that may require tuning in TVMAGI. First, the number of discretization can affect the inference

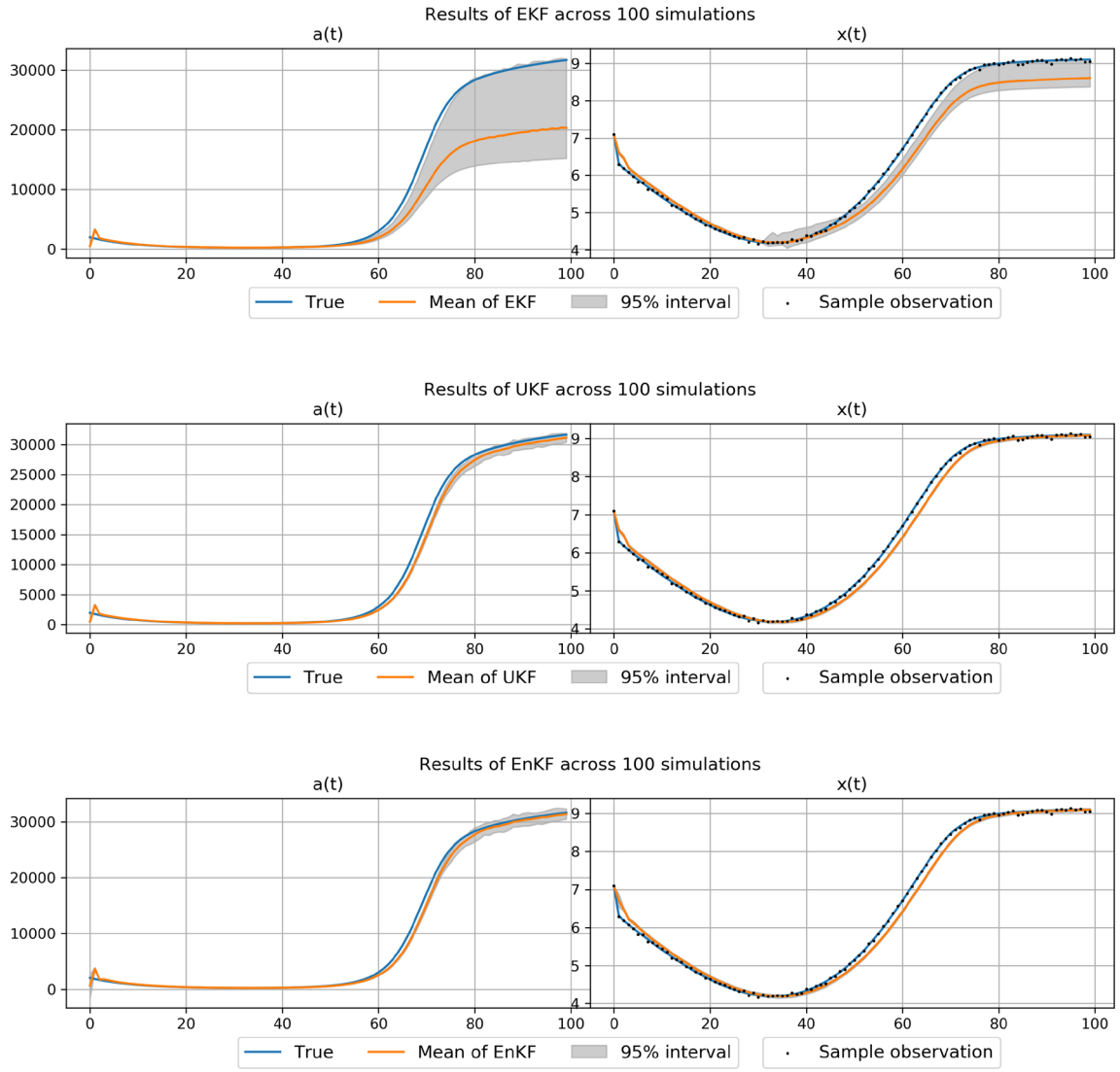


Fig. 8 Inferred parameter and reconstructed trajectory trajectory of EKF, UKF and EnKF approaches for HIV model. The accuracy is worse than TVMAGI, Runge–Kutta, or ELE.

results. When observed components are sparse, the number of discretization should increase until the results are stabilized. However, over-densed discretization will lead to higher computation cost. For example, in SEIRD model, we set discretization as 2 data points per day for optimized performance, as further increasing the discretization will not improve the result accuracy. Second, the inference results on TVMAGI can be affected by hyper-parameter settings of the GP kernel for $\boldsymbol{\theta}(t)$. To achieve the desired variability level of time-varying parameters, we find it helpful to use informative hyper-prior that specifies the range of length-scale (a.k.a. bandwidth) parameter of the GP kernel for $\boldsymbol{\theta}(t)$ to prevent obvious over-smoothing or over-fitting.

The Gaussian process modeling of $\boldsymbol{\theta}(t)$ with Matern kernel $\nu = 2.01$ ensures continuously differentiable time-varying parameters, which prevents overfitting the parameter to the observation noise. The variability in time-varying parameter $\boldsymbol{\theta}(t)$ can be further controlled by the length-scale GP hyper-parameter ϕ_2 through its hyper-prior. The Matern kernel together with the hyper-prior on the length-scale hyperparameter ensures the smoothness and the degree of variability in $\boldsymbol{\theta}(t)$, which in turn prevents over-fitting and mitigates identifiability issues. If a more flexible $\boldsymbol{\theta}(t)$ is desired, Matern kernel $\nu = 1.5$ with hyper-prior favoring smaller GP hyper-parameter ϕ_2 can be used to allow rapid non-differentiable changes in $\boldsymbol{\theta}(t)$.

One limitation of TVMAGI is its inherent bias. Just like any other Bayesian approaches, TVMAGI could be biased towards smoother curves due to the GP prior. The results of inference would be less accurate when the true time-varying parameters have rapid changes, and the posterior interval coverage could suffer. But as shown in the

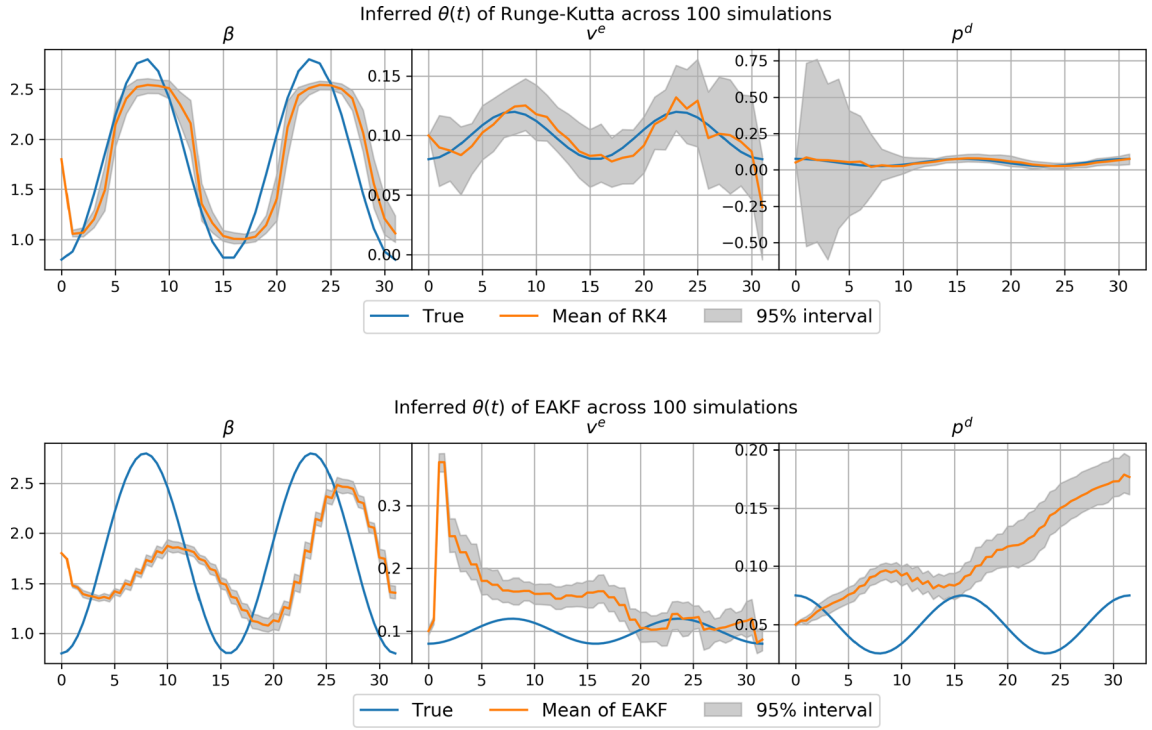


Fig. 9 Results of parameter inference in 100 simulated datasets using Runge–Kutta and EAKF for SEIRD model.

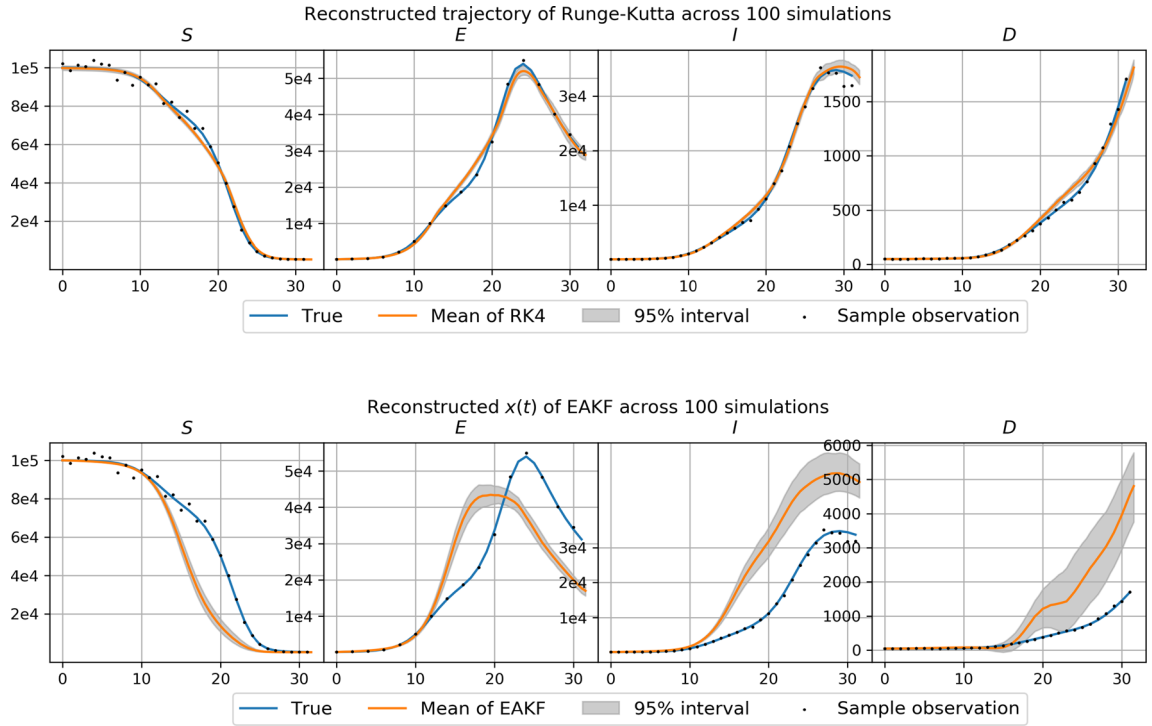


Fig. 10 Reconstructed trajectory using inferred parameters of Runge-Kutta and EAKF methods for SEIRD model. One sample simulation dataset is also presented to visualize the noise level and observation schedule.

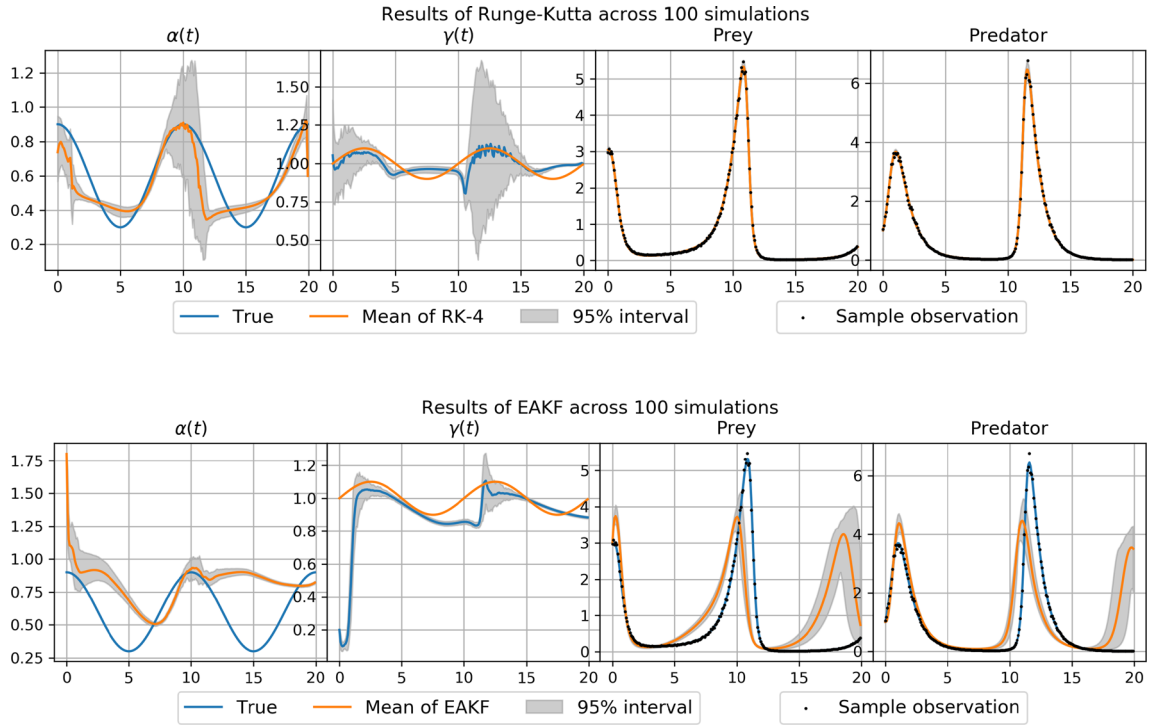


Fig. 11 Comparison of inferred $\boldsymbol{\theta}(t)$ and reconstructed $\boldsymbol{X}(t)$ of LV model. We also plot one sample simulation dataset to visualize the noise level.

examples, the magnitude of such bias is small in practice, and our accuracy is still comparable with state-of-the-art approaches while TVMAGI having much better universal applicability. TVMAGI is also not suitable if the underlying time-varying parameter is a jump process. In this case, methods in change point detection literature might be more applicable (Cuenod et al. 2011). Alternatively, we can place the prior of continuous-time Markov chain or Poisson process on $\theta(t)$, instead of Gaussian process, to model the jump process.

There are also many other interesting future directions for TVMAGI. We currently focus on empirical performance of TVMAGI through simulation examples. More theoretical study on the convergence property, identifiability issue, and asymptotic behavior of the time-varying parameter estimate are all natural directions of future research. It would be of future interest to extend TVMAGI for partial differential equation of spatial-temporal dynamics (Xun et al. 2013), or stochastic differential equation of inherent noise modeling (Kou and Xie 2004).

Declarations

financial interests The authors have no financial or non-financial interests that are directly or indirectly related to this paper.

Appendix A: Supplementary materials

A.1 Advantage of the infinity norm in Eq. (4.5)

In this section we illustrate how L_∞ norm in Eq.(4.5) of main text facilitates theoretical construction, compared to L_2 norm. First, with L_∞ norm in W, it is clear that on the discretization subset I, the corresponding W_I will simply be the maximum over I. However, with the L_2-norm of $\int_0^T (X(t) - f(X(t), \Theta(t), \Psi, t))^2 dt$, the formulation of the corresponding W_I is not as clear. Second, using L_∞ makes the theoretical justification easier. To mathematically study the properties of TVMAGI while avoiding Borel paradox, one can use the fact that $\{W_I < \epsilon\} \equiv \cap_{i\in I}\{W_i < \epsilon\}$, thanks to W_I being the L_∞ norm over the set I. Third, the L_∞ norm in Eq.(4.5) and Eq.(4.6) automatically transforms into L_2 loss for likelihood calculation in Eq.(4.7) and Eq.(4.11) through a simple mathematical derivation, which facilitates computation while maintaining the theoretical rigor. This is because when a Gaussian distributed vector is constrained to have zero deviation with some fixed value (i.e., vector L_∞ distance to the fixed value is zero), the fixed value will be plugged into the Gaussian probability density function, inducing an L_2 loss in the target function Eq.(4.11).

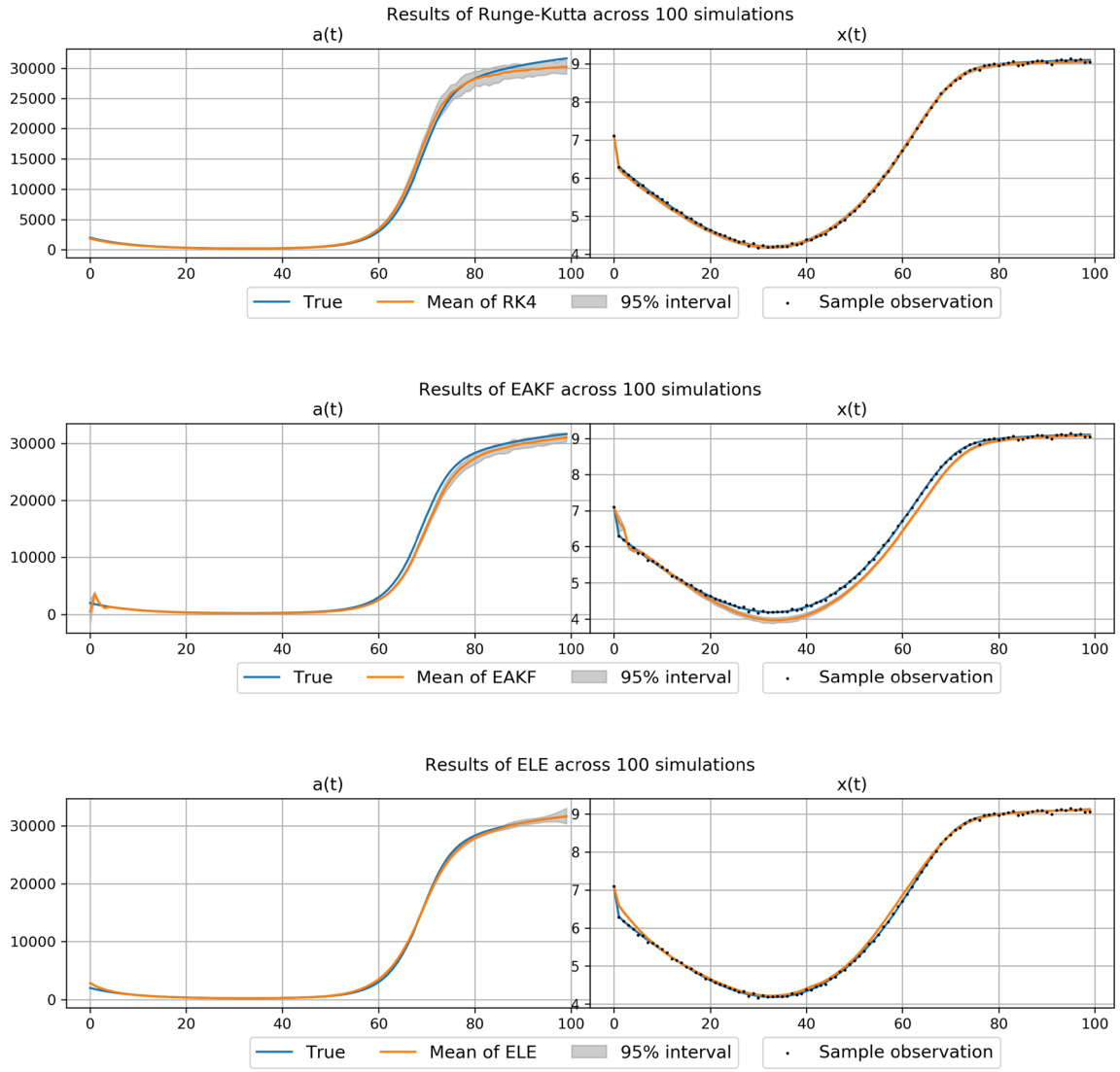


Fig. 12 Comparison of inferred $\boldsymbol{\theta}(t)$ and reconstructed $\boldsymbol{X}(t)$ of HIV model. We also plot one sample simulation dataset to visualize the noise level.

A.2 Intuition of the manifold constraint on Gaussian process

In this section we illustrate the intuition behind the manifold constraint $W = 0$ (or $W_I = 0$) for the Gaussian process.

We consider the following simple bi-variate Gaussian example:

$$(Z_1, Z_2) \sim \mathcal{N}_2\left(0, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

Let $W = Z_1 - 0.5 \times (Z_2^2 + 1)$. Then conditioning on $W = 0$, the (Z_1, Z_2) is distributed on the parabola as in the left panel of Fig. 4. The sampling is possible as the blue dots in the right panel of Fig. 4, where the density is proportional to the original bi-variate Gaussian but only on the parabola curve.

The Gaussian process X_I in the main text and the constraint $W_I = 0$ apply the same intuition on $|I|$-dimensional Gaussian vector, thus having a manifold constraint induced by $W_I = 0$.

A.3 HMC algorithm

We outline the HMC procedure for sampling from a target probability distribution. Algorithm 1 provides the details of our HMC implementation.

A.4 Additional benchmark methods of Bayesian filtering approaches

Compared with Ensemble Adjustment Kalman Filter (EAKF), other Bayesian filtering methods, such as Extended Kalman

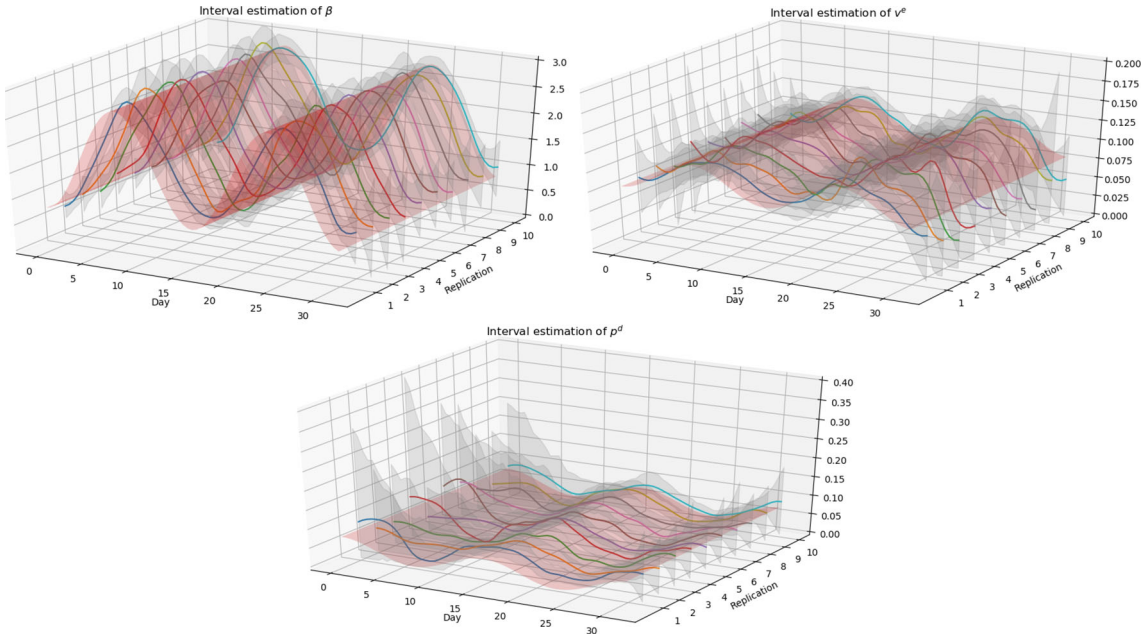


Fig. 13 Illustration of interval estimation of SEIRD model of 10 sample datasets for β (upper), v^e (middle), p^d (lower). The shadow indicates the 95% posterior interval from HMC samples, the solid lines indicate the posterior mean, and the red surface indicates the true value.

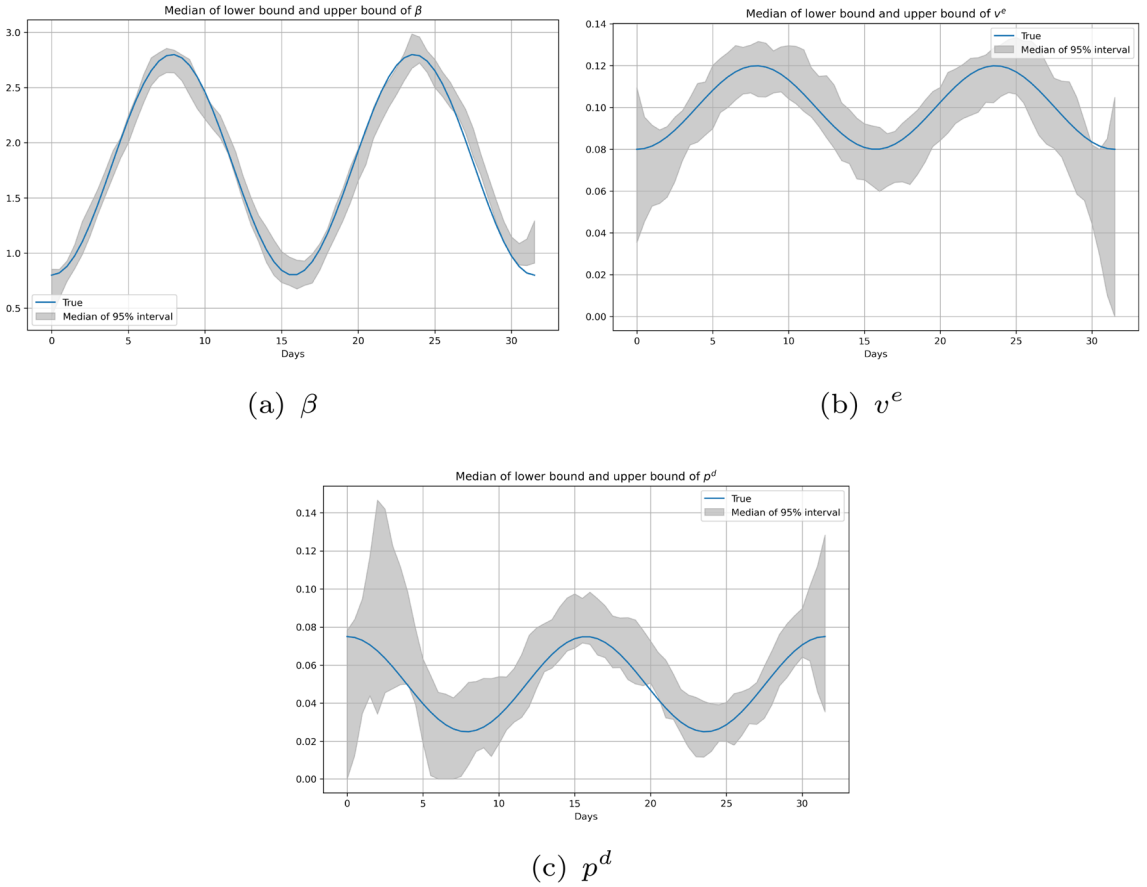
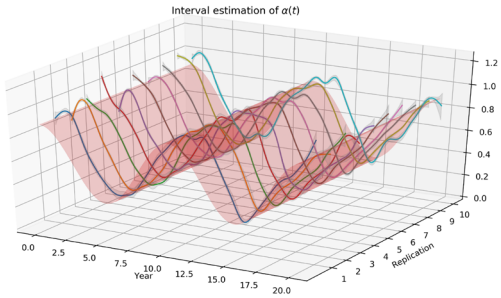
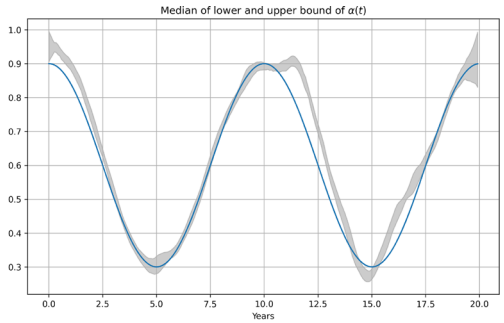


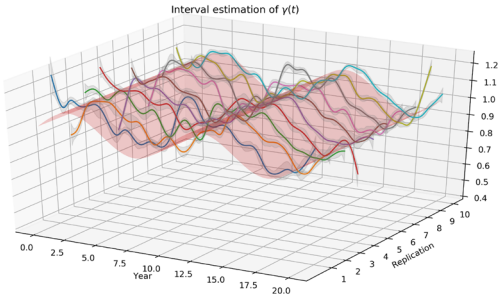
Fig. 14 Median of upper & lower bound of β, v^e and p^d in SEIRD model. We randomly plot 10 replications.



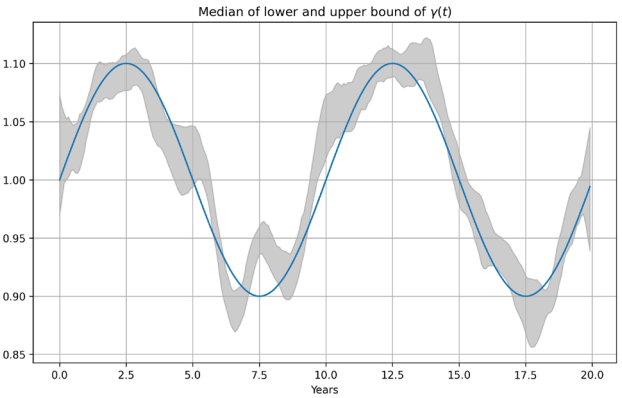
(a) Interval estimation of $\alpha(t)$



(b) Median of upper & lower bound of $\alpha(t)$ interval

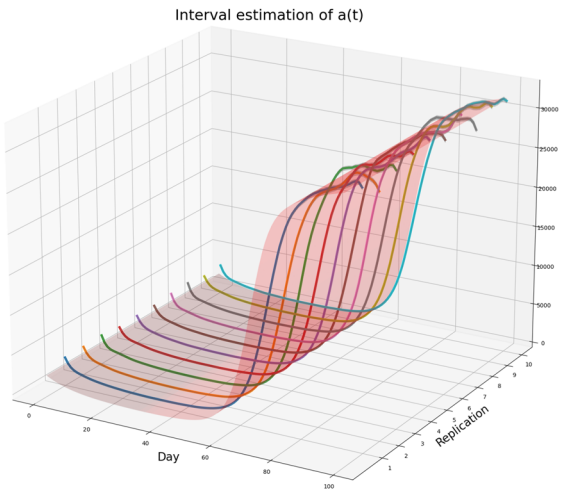


(c) Interval estimation of $\gamma(t)$

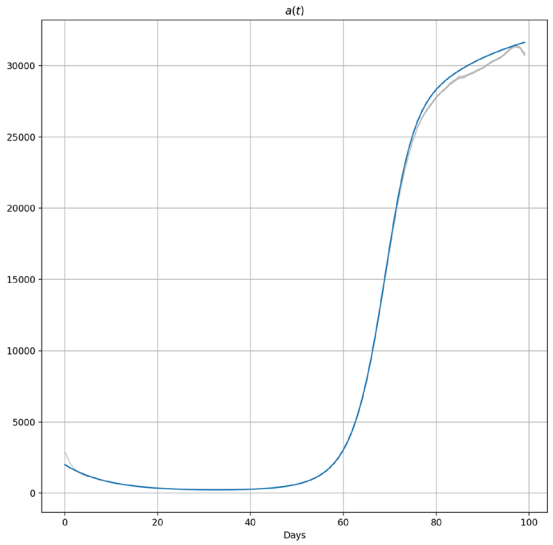


(d) Median of upper & lower bound of $\gamma(t)$ interval

Fig. 15 Estimated interval of $\alpha(t)$ and $\gamma(t)$ in LV model. We randomly plot 10 replications.



(a) Interval estimation of $a(t)$



(b) Median upper and lower bound of $a(t)$

Fig. 16 Estimated interval of $a(t)$ in HIV model. We randomly plot 10 replications.

Algorithm 1 HMC sampling in TVMAGI

Input:
U: Log likelihood function in Eq.(4.11)
ϵ: step size of HMC
L: number of leaf frog steps
N: number of samples
Initialize: $\boldsymbol{x}(\boldsymbol{I}), \boldsymbol{\theta}(\boldsymbol{I}), \boldsymbol{\psi}, \boldsymbol{\sigma}$

for i in 1:N **do**
  $\boldsymbol{q}_{\text{current}} = \textbf{vector}(\boldsymbol{x}(\boldsymbol{I}), \boldsymbol{\theta}(\boldsymbol{I}), \boldsymbol{\psi}, \boldsymbol{\sigma})$
  $\boldsymbol{q} = \boldsymbol{q}_{\text{current}}$
  $\boldsymbol{p} = \textbf{rnorm}(\textbf{length}(\boldsymbol{q}), 0, 1)$
  $\boldsymbol{p}_{\text{current}} = \boldsymbol{p}$
  $\boldsymbol{p} = \boldsymbol{p} - \epsilon \nabla U(\boldsymbol{q})/2$
  for j in 1:L **do**
    $\boldsymbol{q} = \boldsymbol{q} + \epsilon \boldsymbol{p}$
    $\boldsymbol{p} = \boldsymbol{p} - \epsilon \nabla U(\boldsymbol{q})$
  end for
  $\boldsymbol{p} = \boldsymbol{p} - \epsilon \nabla U(\boldsymbol{q})/2$
  if runif$(1) < \textbf{exp}(U(\boldsymbol{q}_{\text{current}}) - U(\boldsymbol{q}) + \textbf{sum}(\boldsymbol{p}_{\text{current}}^2 - \boldsymbol{p}^2)/2)$ **then**
    return $\boldsymbol{q}$ ▷ (Accept)
  else
    return $\boldsymbol{q}_{\text{current}}$ ▷ (Reject)
  end if
end for

Filter (EKF), Unscented Kalman Filter (UKF) and Ensemble Kalman Filter (EnKF) are less discussed in ODE parameter inference applications. In this section we also include a few more Bayesian filtering benchmark methods of Extended Kalman Filter (EKF), Unscented Kalman Filter (UKF), and EnKF. All the Bayesian filtering methods have the inherent limitation that all parameters must be assumed time-varying, and thus cannot accommodate time-constant parameters. To further illustrate the Bayesian filtering approaches, we also provide inference results using the other three methods. Figure 5 and Fig. 6 illustrate the results of SEIRD model. Figure 7 shows the results of LV model and Fig. 8 shows the results of HIV model.

A.5 Additional results of the main benchmark methods

Due to the limited space of the main text, we include visualizations of the main benchmark methods of Runge–Kutta method and EAKF method for the three examples in the main text here. Figure 9 is for SEIRD model parameter inference, and Fig. 10 is for SEIRD model reconstructed trajectory. Figure 11 is for LV model. Figure 12 is for HIV model.

A.6 Review on benchmark methods

A.6.1 Runge–Kutta method

Runge–Kutta methodis a brute-force way for parameter inference in ODE systems. As a non-linear least square method, Runge–Kutta method minimizes the MSE of observations and reconstructed trajectory using numerical integration from the proposed initial conditions and parameters. The objective function is given by:

$$\min_{\boldsymbol{x_0}, \boldsymbol{\theta}(\boldsymbol{I}), \boldsymbol{\psi}} \sum_{\tau \in \boldsymbol{\tau}_d} \sum_{d=1}^{D} (y_d(\tau) - \mathbb{X}_{\tau}^{RK4}(\boldsymbol{x_0}, \boldsymbol{\theta}(t), \boldsymbol{\psi})_d)^2 \tag{19}$$

where $\mathbb{X}^{RK4}$ denotes the reconstructed trajectory using the 4th Order Runge–Kutta method.

A.6.2 Ensemble adjustment Kalman filter

Ensemble Adjustment Kalman Filter (EAKF) is a variation of Kalman Filter that is popular for parameter calibration of ODE systems in practice. It is a specially designed fast Bayesian filtering method. As a data assimilation technique, EAKF represents filtered distribution using Monte Carlo samples, and replaces the covariance matrix with sample covariance. The Kalman update assumes all probability distributions involved are Gaussian. As a major difference with Ensemble Kalman Filter (EnKF), EAKF uses a deterministic update instead of stochastic update.

A.7 Limitations of Runge–Kutta method

Numerical integration methods are the gold standard for the ODE parameter inference when all parameters are time-constant. However, with the presence of time-varying parameters, there are several inherent disadvantages of numerical integration methods. First and foremost, without any structure on the time-varying parameters, the numerical integration method will give time-varying parameter estimate that perfectly fits the observation data, resulting in overfitting issues in the time-varying parameter. Second, with the increase of time points and size of the system, the objective function becomes expensive to evaluate, resulting in high computation cost. Third, the numerical methods are sensitive to the initial value of the optimization, while searching for a good initial point can be challenging in the high-dimensional scenarios, as optimization of objective functions using algorithms such as Adam can be easily stuck at local minimum. We point out that using random initial values in our examples can lead to high level of error, making numerical integration methods completely fail. In this case, all the optimization for Runge–Kutta method are initialized at the TVMAGI initial points from Section 5.1 in the main text examples.

Table 4 Parameter and trajectory RMSE of TVMAGI in SEIRD model under different discretization level. The computing time (in seconds) is reported in the last row

	RMSE	Discretization Level 1	2	4
Parameter	β	0.156 ± 0.083	0.114 ± 0.039	0.102 ± 0.036
	v^e	0.007 ± 0.007	0.009 ± 0.010	0.010 ± 0.010
	v^i	0.006 ± 0.004	0.005 ± 0.003	0.005 ± 0.003
	p^d	0.018 ± 0.027	0.019 ± 0.029	0.021 ± 0.033
Trajectory	S	1253.4 ± 332.0	581.7 ± 272.1	603.9 ± 355.2
	E	1010.6 ± 231.9	704.7 ± 218.3	679.5 ± 209.1
	I	584.6 ± 195.6	439.0 ± 140.4	401.0 ± 132.8
	D	44.7 ± 8.8	38.3 ± 4.9	39.0 ± 3.4
computing time (s)		622.1 ± 49.6	1013.3 ± 109.1	1773.9 ± 161.3

Table 5 Parameter and trajectory RMSE of TVMAGI in SEIRD using different kernels

	RMSE	ν for $\boldsymbol{\theta}(t)$ 1.5	2.5	2.01
Parameter	β	0.085 ± 0.005	0.139 ± 0.021	0.114 ± 0.039
	v^e	0.012 ± 0.004	0.015 ± 0.005	0.009 ± 0.010
	v^i	0.004 ± 0.005	0.003 ± 0.004	0.005 ± 0.003
	p^d	0.058 ± 0.004	0.044 ± 0.012	0.019 ± 0.029
Trajectory	S	423.0 ± 130.2	853.3 ± 288.5	581.7 ± 272.1
	E	412.3 ± 111.7	836.1 ± 186.5	704.7 ± 218.3
	I	356.3 ± 94.1	569.4 ± 171.8	439.0 ± 140.4
	D	69.1 ± 21.5	20.4 ± 8.6	38.3 ± 4.9

Table 6 Parameter and reconstructed RMSE table for mis-specified time-varying parameter v^e

RMSE Parameter			Reconstructed		
β	0.120	0.046	S	513.13	232.08
v^e	0.008	0.009	E	647.16	232.81
v^i	0.007	0.004	I	364.94	197.92
p^d	0.022	0.036	D	27.61	6.52

A.8 Limitations of Bayesian filtering methods

Although Bayesian filtering method is the fastest, examples have shown that applying Bayesian filtering methods to ODE parameter inference problems has failed to provide satisfactory results. Even though we included the Bayesian filtering methods as baseline comparison methods, we emphasize that state-space model (Bayesian filtering) and ODE parameter inference (TVMAGI) are fundamentally different problems. To illustrate the difference in a simplified framework from a theoretical perspective, consider the following example of time-constant parameter inference where all parameters are denoted as $\boldsymbol{\theta}$.

The fundamental difference is the lack of randomness in the state transition given the model parameters, and thus the Bayesian update given the observation will have zero effect. With the ODE structure, there is no randomness in $\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}$. The state transition distribution $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta})$ essentially has shifted Dirac delta distribution. Therefore, regardless of emission probability $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$, the hidden state $\boldsymbol{x}_t$ will not depend on $\boldsymbol{y}_t$. As such, all Bayesian updates will have zero effect to shift distribution of $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta})$, which is still shifted Dirac delta distribution. The exact Bayesian filtering results will simply be the solution of ODE dynamics given the initial sample of $\boldsymbol{x}_0$ and the parameter $\boldsymbol{\theta}$. In this case, the parameter estimation in the exact Bayesian filtering reduces to using numerical solver to generate the entire ODE curve given $\boldsymbol{x}_0, \boldsymbol{\theta}$, and then using a least square approach to compare the solved curve and the observations to find the best $\boldsymbol{x}_0, \boldsymbol{\theta}$. The exact Bayesian filtering in this case degenerates to a numerical integration method. From another particle filter perspective, each particle of sampled $\boldsymbol{x}_0, \boldsymbol{\theta}$ will evolve in time according to ODE without any randomness, and the parameter estimation becomes a brute-force search of the particle of sampled $\boldsymbol{x_0}, \boldsymbol{\theta}$ that provides the smallest mean squared error to the observation. In light of this, the Bayesian filtering/smoothing is more suitable for inference of Stochastic Differential Equations (SDEs) parameters.

However, Bayesian filter methods still can be applied in ODE inference problem if we can forego some statistical rigor. We can treat **all** parameters as time-varying, and

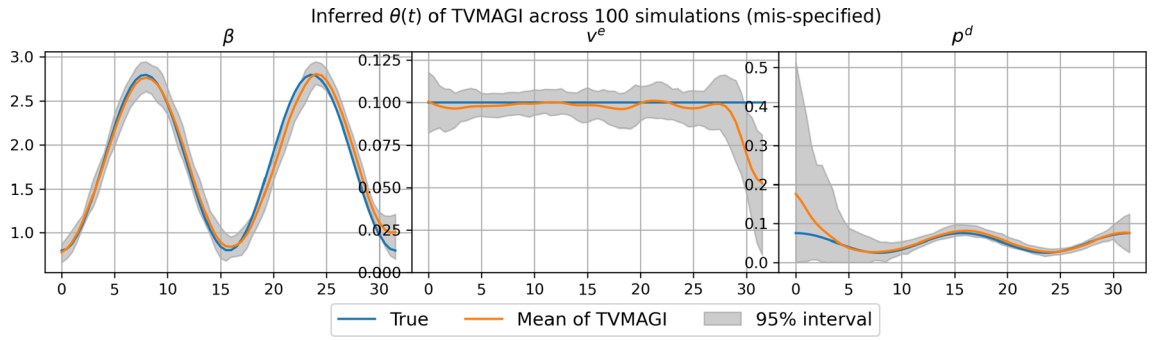
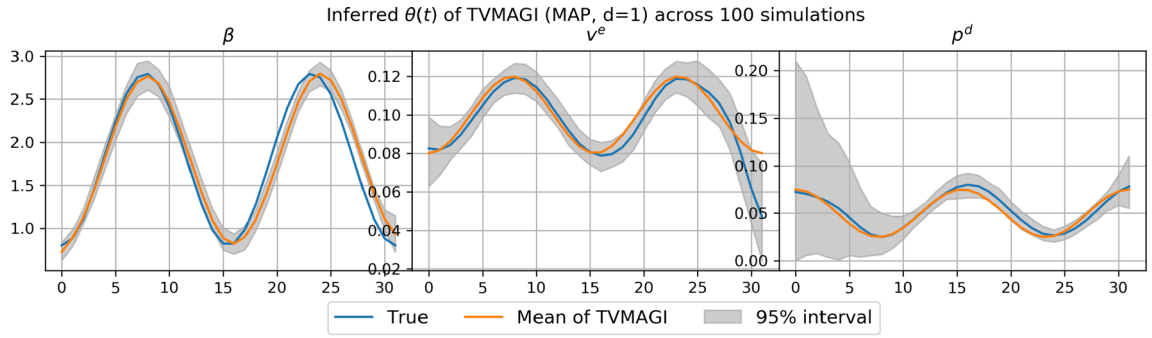
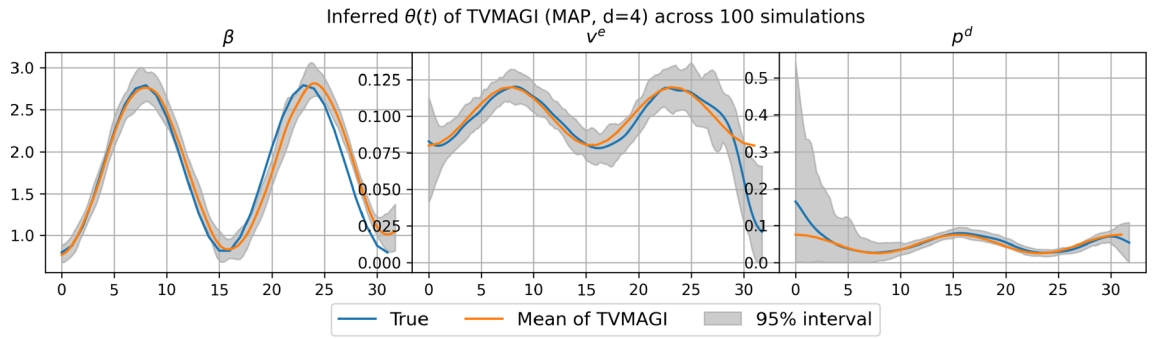


Fig. 17 Mis-specified parameter v^e in SEIRD model. Figure y-scale is the same as Figure 1 for better visualization.



(a) Discretization level = 1



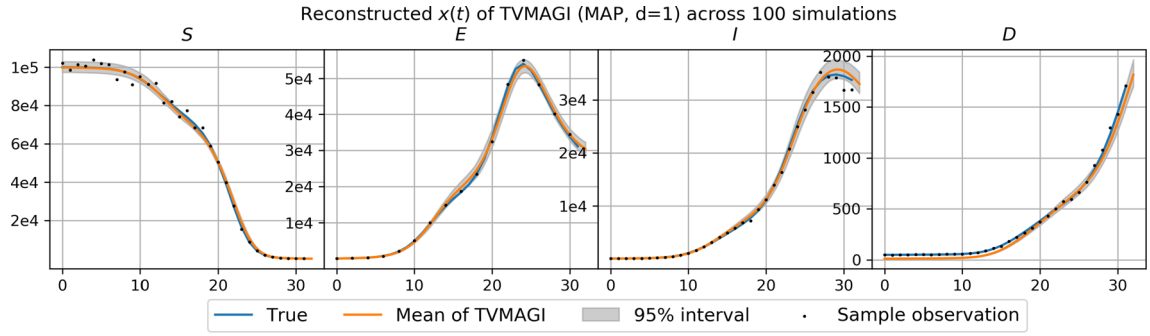
(b) Discretization level = 4

Fig. 18 TVMAGI inferred parameters in SEIRD model with different discretization level.

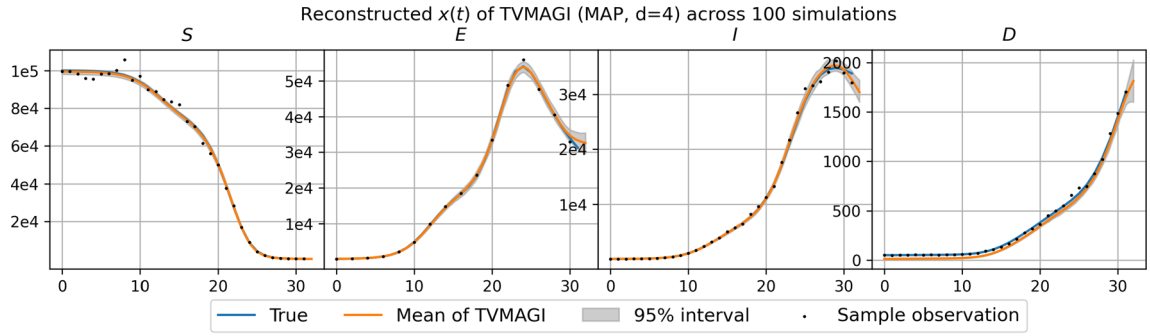
artificially introduce additional randomness in $\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}$ and $\boldsymbol{\psi}_t|\boldsymbol{\psi}_{t-1}$. In this case, the simultaneous estimation of system components $\boldsymbol{x}(t)$, time-constant parameter $\boldsymbol{\psi}$ and time-varying parameter $\boldsymbol{\theta}(t)$ becomes an estimation of joint hidden state $(\boldsymbol{x}_t, \boldsymbol{\psi}_t, \boldsymbol{\theta}_t)$. Then Bayesian filtering methods such as EKF, UKF, EnKF and EAKF become applicable. However, this is not a statistically principled approach, because (1) time-constant parameter $\boldsymbol{\psi}$ is now changing with time, and (2) the inference on the distribution is not exact as Gaussian distributional approximation is used on system components $\boldsymbol{x}(t)$. Nevertheless, this method could work empirically, with the notable success of SIRS-EAKF model.

In our example, we found the Bayesian filter methods are highly sensitive to the initialization of the parameters. Randomized initialization often lead to insensible results. Therefore, all the parameters for Bayesian filter methods are initialized at the TVMAGI initial points from Section 5.1 in the main text examples, which is the same as Runge–Kutta method.

Even then, our numerical experiments suggest that Bayesian filtering methods are the fastest, but often yield unreliable inference results. Among EKF, UKF, EnKF and EAKF, we see that UKF, EnKF and EAKF yield similar results, all outperforming EKF in LV and HIV examples, while EKF has a



(a) Discretization level = 1



(b) Discretization level = 4

Fig. 19 TVMAGI reconstructed trajectories in SEIRD model with different discretization level.

slight advantage in SEIRD model inference compared with other Bayesian filtering methods, although not robust. All of them yield orders of magnitude worse trajectory RMSE compared to TVMAGI or Runge–Kutta. Two factors contribute to the unsatisfactory trajectory RMSE of Bayesian filtering methods. First, all filtering approaches fail to yield an accurate and robust parameter estimation, especially in SEIRD model when the observation points are limited. Second, the variance of time-varying parameters is large at weakly identifiable time points (Fig. 7), and the estimates for the later time points are no longer accurate due to the cascading effect.

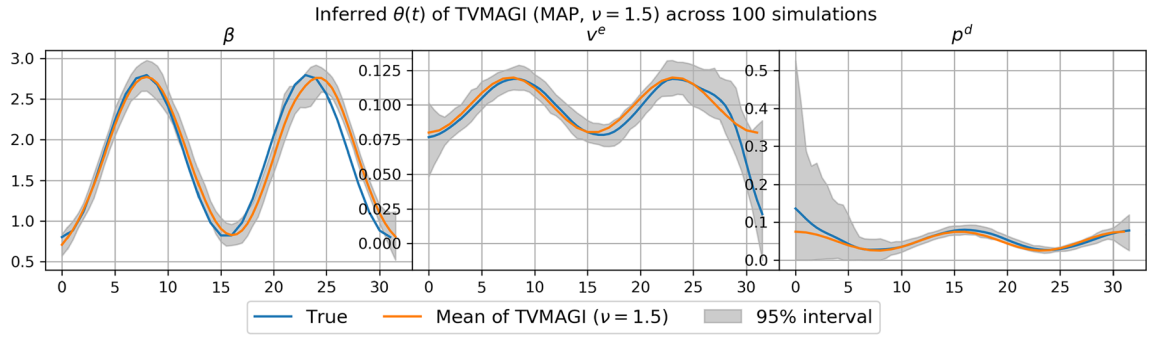
The failure of Bayesian filtering in our setting is not surprising. First, it violates the assumption of the model, as it is not principled to allow time-constant parameters $\boldsymbol{\psi}$ to change over time. Although we used the average $\bar{\boldsymbol{\psi}}$ of the filtered parameter $\boldsymbol{\psi}_t$ to be the final estimate for the $\boldsymbol{\psi}$, the approximation error can still be large. The idea of changing a time-constant parameter to be time-varying during inference and later plugging in the average of the inferred values is not theoretically sound. The trajectory RMSE precisely evaluates how accurate the estimated parameters can be used to reconstruct the entire system given the ODE structure, of which $\bar{\boldsymbol{\psi}}$ would fail. Second, contrary to the typical setting of Bayesian filtering in machine learning where there is a long sequence of data, our experiments are designed to see how the method performs with short time series and sparse observations, as in most scientific experiment settings. The lack of long series of data poses a challenge to Bayesian filtering methods. Third, the ODE structure is no longer exactly followed in Bayesian filtering, which loosens the structure constraints and creates additional loss of information from the observation data that is already sparse.

A.9 Additional results of TVMAGI interval estimates

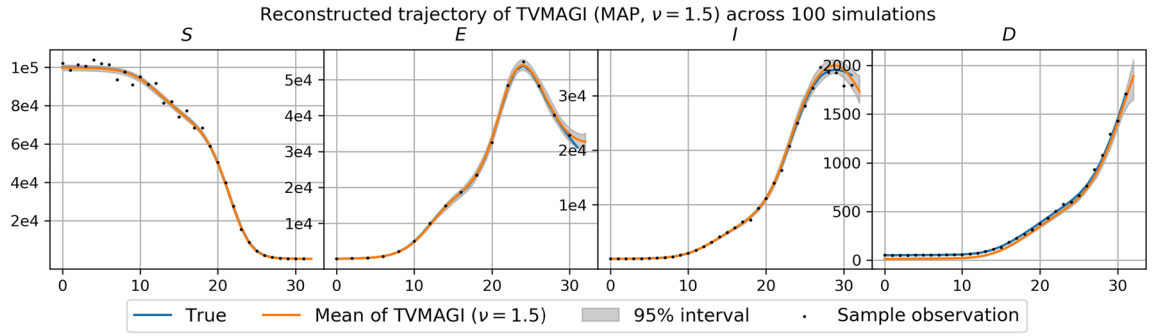
In this section we present the visualization for the interval estimation results. We see that for long time series of observations, the estimated intervals tend to be narrow and may not contain the true values, which is a limitation of TVMAGI.

A.10 Additional results on TVMAGI sensitivity study

In this section we conduct three sensitivity analysis: the number of discretization, the selection of GP kernel, and the mis-specified time-varying parameter. For the number of discretization, our theoretical derivation ensures that the inference result will converge as the discretization increase, and we recommend gradually increasing the number of



(a) Inferred parameters of $\nu = 1.5$



(b) Reconstructed trajectory of $\nu = 1.5$

Fig. 20 TVMAGI inferred parameters and reconstructed trajectories of SEIRD model with Matern kernel $\nu = 1.5$.

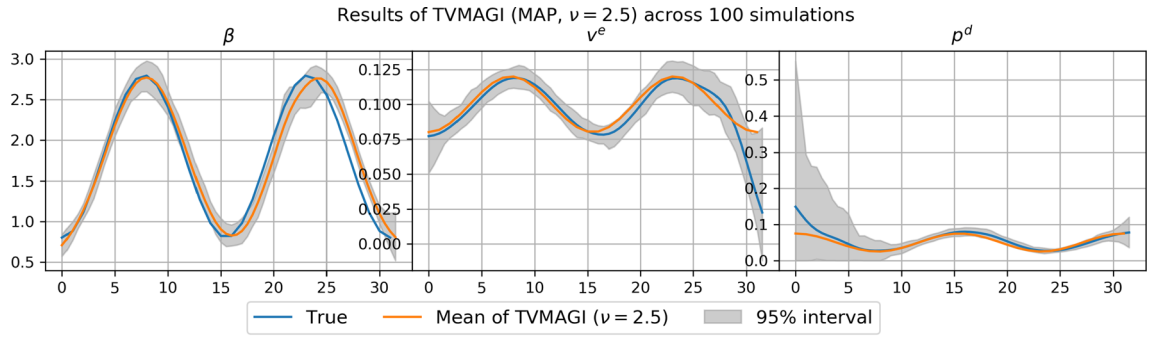
Table 7 Accuracy comparison of estimated parameters and reconstructed trajectory in SEIRD model based on 100 simulation datasets. d is the discretization size. For example, $d = 5$ means 5 discretizations per day, that is, 5 discreization points per single observation points. n indicates the ensemble size in the Ensemble Adjustment Kalman Filter (EAKF). See legend of Table 1 for additional description

	RMSE	$d = 2, n = 1e5$		$d = 5, n = 1e5$		$d = 5, n = 2e5$		$d = 10, n = 2e5$		$d = 20, n = 2e5$	
Parameter	β	0.706	0.010	0.688	0.0186	0.683	0.0179	0.702	0.0145	0.737	0.0206
	v^e	0.057	0.003	0.044	0.005	0.040	0.006	0.061	0.006	0.056	0.0008
	v^i	0.151	0.008	0.146	0.006	0.146	0.005	0.153	0.002	0.155	0.0004
	p^d	0.039	0.003	0.043	0.007	0.046	0.009	0.034	0.005	0.0.032	0.006
Traj.	S (prey)	3868.3	132.2	3786.7	152.9	4379.3	158.8	3879.3	158.8	4083.6	177.5
	E (predator)	5376.1	496.6	5618.4	522.7	6769.0	525.9	5469.0	540.4	5718.9	623.7
	I (predator)	3167.0	404.9	1908.6	358.4	3081.3	392.1	2008.9	449.4	2168.2	620.9
	D (predator)	907.3	48.6	924.6	52.6	72.4	30.8	317.5	35.7	130.4	10.6
Computing Time (s)		7.3	0.4	20.4	3.5	53.4	11.9	121.8	25.7	267.7	40.8

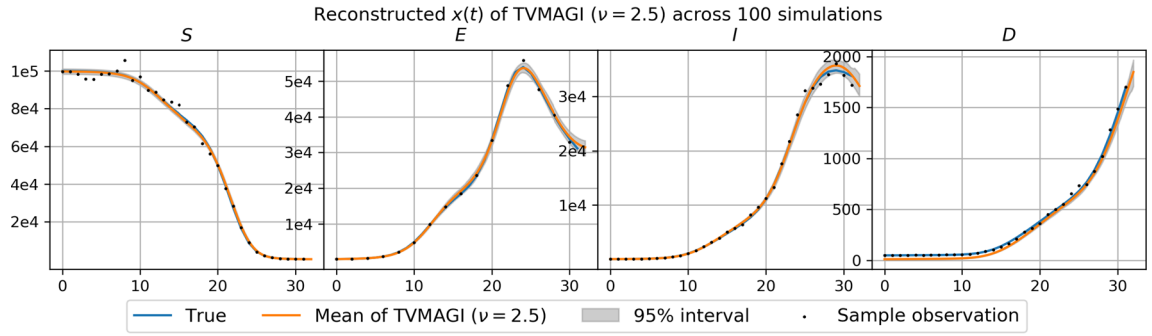
discretization points until the result is stabilized. Here we empirically demonstrate such convergence by presenting results with various discretization level. For the GP kernel selection, we relax the GP kernel of time-varying parameter to be Matern kernel with different degrees of freedom $\nu = 2.5$ and $\nu = 1.5$. When $\nu = 1.5$, $\boldsymbol{\theta}(t)$ is only required to be continuous, but not necessarily differentiable. For the misspecified time-varying parameter, we examine the TVMAGI time-varying estimate when one parameter is in fact time-constant.

A.11 Number of discretization

In this section we explore the sensitivity of TVMAGI to the number of discretization. In the paper we used discretization level of 2 in the SEIRD model, that is, with 32 observation points, we have a total of 64 discretization points (2



(a) Inferred parameters of $\nu = 2.5$



(b) Reconstructed trajectory of $\nu = 2.5$

Fig. 21 TVMAGI inferred parameters and reconstructed trajectories of SEIRD model with Matern kernel $\nu = 2.5$.

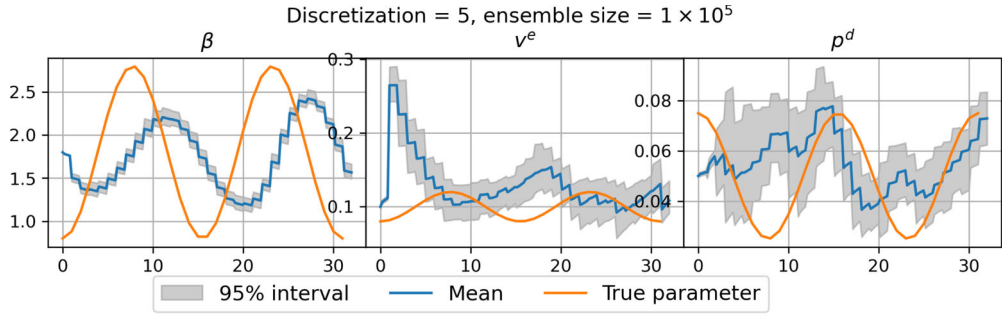
discretization per observation). For comparison, we use the same observation data with discretization level of 1 and level of 4, corresponding to 32 discretization points (1 discretization per observation) and 128 discretization points (4 discretization per observation), respectively. Table 4 shows that the inference accuracy indeed converges. However, the computational time scales up linearly with the number of discretization points. In practice, we recommend gradually increasing the number of discretization points until the result is stabilized, trying to balance the inference accuracy with the computing cost.

A.12 Selection of kernel

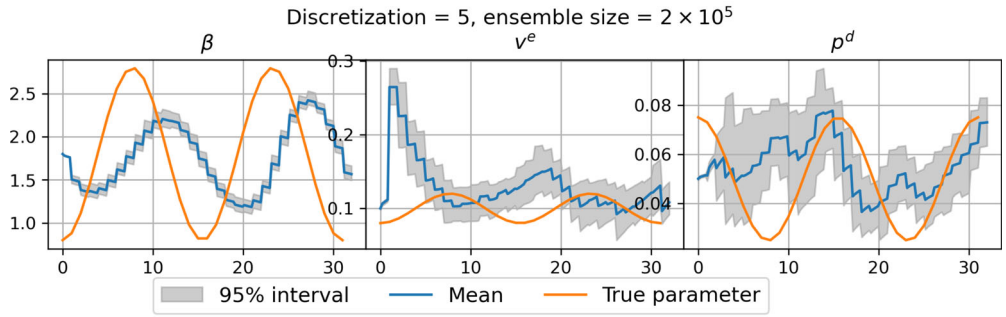
In this section we discuss how the kernel selection will affect the performance of TVMAGI. In the paper we recommend modeling $\boldsymbol{\theta}(t)$ as Gaussian process with Matern kernel $\nu = 2.01$ to guarantee a continuous and differentiable time-varying parameters while maintaining high flexibility. We can also use other GP kernels or hyperparameters to control the smoothness. For example, Matern kernel with $\nu = 2.5$ can be used for even smoother GP with a simple closed-form kernel. The condition of differentiability can also be further relaxed if we substitute the kernel with $\nu = 1.5$, and then the parameters are only assumed with continuity without differentiability, allowing more flexible patterns for the time-varying parameters $\boldsymbol{\theta}(t)$. Table 5 shows the result under both kernels, where the the performance is similar to the recommended $\nu = 2.01$ in SEIRD model, indicating that TVMAGI is not sensitive to the choice of kernels.

A.13 Mis-specified time-varying parameters

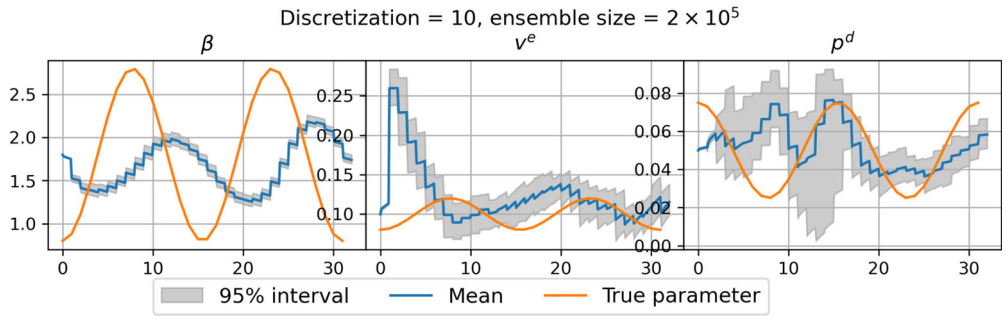
In this section, we explore the TVMAGI estimation when a time-constant parameter is mis-specified to be time-varying, i.e., a time-constant ODE parameter is falsely recognized as a time-varying parameter. Ideally, the inferred parameter curve from TVMAGI should be a horizontal line which is close to the true constant value, while still maintaining smoothness. In this example, we alter the settings of SEIRD model by setting parameter v^e as constant $v^e = 0.1$, and treat it as time-varying parameter in the TVMAGI estimation. As shown in Fig. 17, TVMAGI is capable of dealing with the mis-specified parameters, where the inferred v^e is approximately a horizontal line close to true parameter value (except the ending time when the v^e is difficult to estimate), while the inference of other parameters are not affected. The parameter and reconstructed RMSE results are presented in



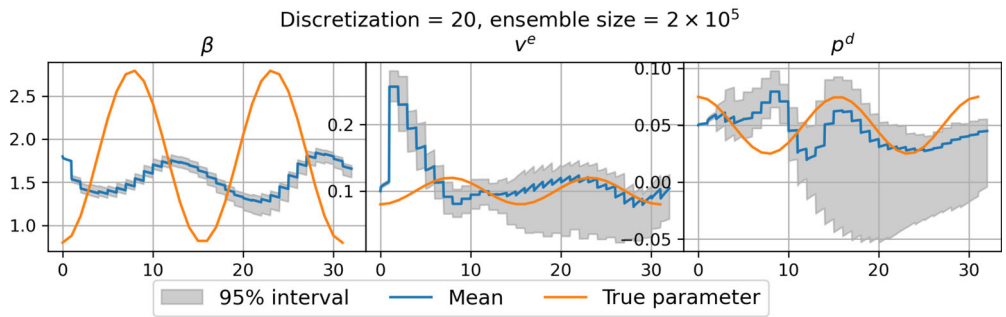
(a) $d = 5$, ensemble size $= 1 \times 10^5$



(b) $d = 5$, ensemble size $= 2 \times 10^5$

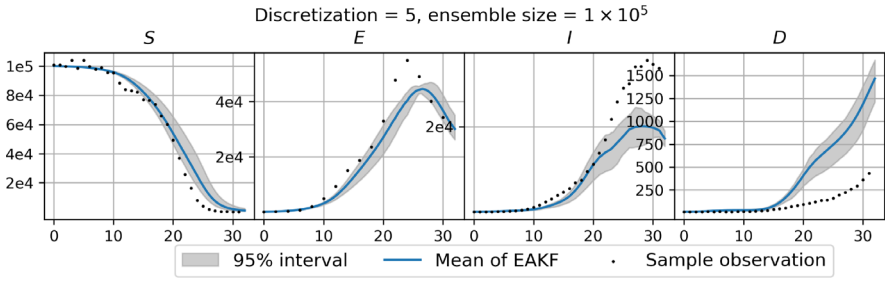


(c) $d = 10$, ensemble size $= 2 \times 10^5$

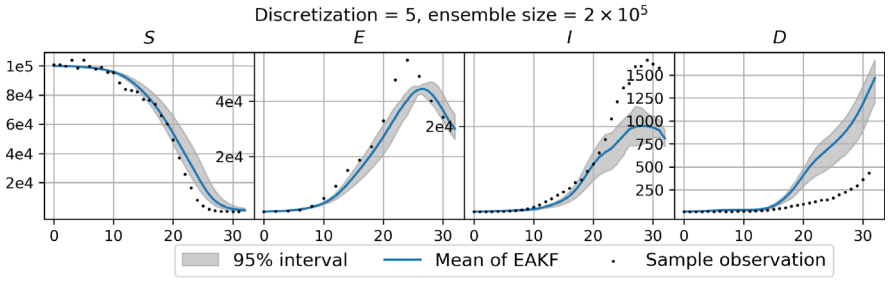


(d) $d = 20$, ensemble size $= 2 \times 10^5$

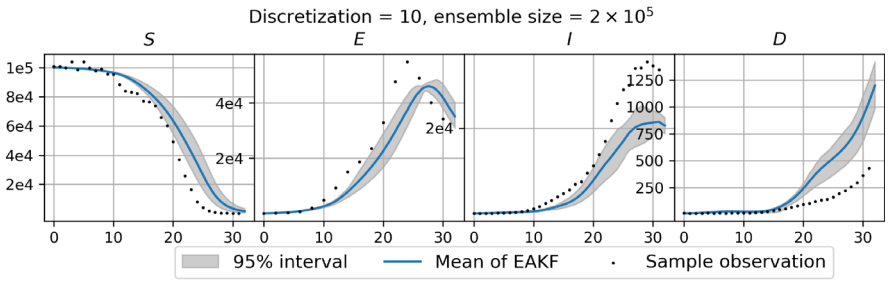
Fig. 22 Comparison of inferred parameters under different EAKF settings. d is the discretization size. For example, $d = 5$ means 5 discretizations per day, that is, 5 discreization points per single observation points. n indicates the ensemble size in the Ensemble Adjustment Kalman Filter (EAKF).



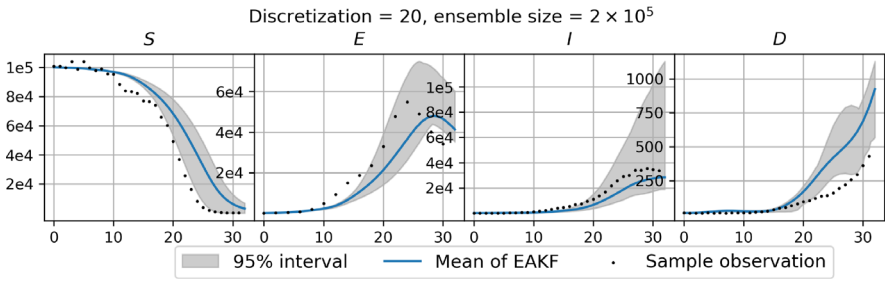
(a) $d = 5$, ensemble size $= 1 \times 10^5$



(b) $d = 5$, ensemble size $= 2 \times 10^5$



(c) $d = 10$, ensemble size $= 2 \times 10^5$



(d) $d = 20$, ensemble size $= 2 \times 10^5$

Fig. 23 Comparison of reconstructed trajectory under different EAKF settings. d is the discretization size. For example, $d = 5$ means 5 discretizations per day, that is, 5 discreization points per single observation points. n indicates the ensemble size in the Ensemble Adjustment Kalman Filter (EAKF).

Supplementary Material Table A1, and the accuracy is comparable to the Table 1 in main text where all time-varying parameters are correctly specified.

A.14 Discretization and choice of kernels

In this section we give additional visualizations for TVMAGI sensitivity study when varying the discretization level, and Matern kernel degree of freedom ν.

A.15 Structure of HIV model

The ODE model that characterizes the response of anti-viral regimens during HIV infection is given by:

$$\begin{cases} T'(t) \equiv \dfrac{d}{dt}T(t) = \lambda - \rho T(t) - k[1-r(t)]T(t)X(t) \\ T^{*\prime}(t) \equiv \dfrac{d}{dt}T^*(t) = k[1-r(t)]T(t)X(t) - \delta T^*(t) \\ X'(t) \equiv \dfrac{d}{dt}X(t) = N\delta T^*(t) - cX(t) \end{cases} \tag{20}$$

$T(t)$ denotes the concentration of uninfected CD4+ T cells, which can be accurately measured clinically; $T^*(t)$ denotes the unknown unobservable concentration of infected T cells; $X(t)$ is the HIV-1 viral load in plasma, and can be observed with noise. λ is the rate of new T cell generation; ρ is the death rate of T cells; k is the infection rate of T cells by HIV virus; δ is the death rate of infected cells; N is the total production of new virions by an infected T cell; c denotes the known constant rate of free virion clearance; $r(t)$ is the time-varying antiviral drug efficacy coefficient, which may decay through time due to drug resistance. Our simulation settings are based on as $\lambda = 36$, $\rho = 0.108$, $k = 5 \times 10^{-4}$, $\delta = 0.1$, $N = 1000$, $c = 3.5$, $X(0) = 1000$, $T(0) = 350$, $T^*(0) = 20$, and $r(t) = \cos(\pi t/500)$. Time horizon is set as 100 days, with observation noise level at 5%. Hulin Wu (2008) transformed the system Eq.(20) into Eq.(7.17) by taking $d = 2$, $a_1(t) = -NT^{*\prime}(t)$, $a_2(t) = Nk[1-r(t)]X(t)$, $Z_1(t) = 1$ and $Z_2(t) = T(t)$, and then used their ELE method estimate time-varying coefficients $a_i(t)$.

A.16 Advantages of multi-stage algorithm

Compared with joint optimization of hyperparameters and parameters together, the multi-stage optimization method enjoys several advantages. First, the GP hyperparameters $\mathbf{\Phi}^X$ for the system components are set at the first stage and held as constant in the rest of the optimization so that the inverse of kernel matrix only needs to be computed once. Second, GP hyperparameters $\mathbf{\Phi}^\Theta$ for the time-varying parameters could not be set without any information about $\mathbf{\Theta}(I)$. Therefore, a multi-stage procedure is necessary, where a point-wise $\tilde{\boldsymbol{\theta}}(I)$ is obtained in one stage without GP, and then GP hyperparameters $\mathbf{\Phi}^\Theta$ is estimated in the following stage based on $\tilde{\boldsymbol{\theta}}(I)$. Lastly, The multi-stage optimization ensures that each step of the optimization starts with sensible initial value obtained from previous modularized optimization, thus drastically decreasing the chance of Adam optimizer stuck in local mode. Experiments have shown that our carefully designed multi-stage optimization is faster and achieves better results than joint optimization with randomized starting values.

However, although we carefully designed the multi-stage optimization and sampling schedule, occasionally the Adam optimizer or the HMC sampler could still get stuck. Among the total of 300 simulated datasets across three examples, the algorithm got stuck in one particular dataset of the SEIRD model. In the stuck case, some manual tuning of the hyperparameters or jittering of the sampled parameters might be needed. We will continue to improve the robustness of our proposed algorithm and our software implementation.

A.17 Results of EAKF under different discretization and sample size

The computation time of Ensemble Adjustment Kalman Filter (EAKF) increases with denser discretization and more sample sizes. In this section, the results of EAKF under different settings are exhibited and compared. It can be seen that the accuracy of EAKF will not improve dispite the increased computation cost.

References

Li, M.Y., Muldowney, J.S.: Global stability for the SEIR model in epidemiology. Math. Biosci. **125**(2), 155–164 (1995). https://doi.org/10.1016/0025-5564(95)92756-5

Takeuchi, Y., Du, N.H., Hieu, N.T., Sato, K.: Evolution of predator–prey systems described by a Lotka–Volterra equation under random environment. J. Math. Anal. Appl. **323**(2), 938–957 (2006). https://doi.org/10.1016/j.jmaa.2005.11.009

Perelson, A.S., Neumann, A.U., Markowitz, M., Leonard, J.M., Ho, D.D.: Hiv-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **271**(5255), 1582–1586 (1996)

Lapidus, L., Seinfeld, J.H.: Numerical Solution of Ordinary Differential Equations. Academic Press, Cambridge (1971)

Abbott, S., Hellewell, J., Thompson, R.N., Sherratt, K., Gibbs, H.P., Bosse, N.I., Munday, J.D., Meakin, S., Doughty, E.L., Chun, J.Y.: Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Res. **5**(112), 112 (2020)

Miao, H., Xia, X., Perelson, A.S., Wu, H.: On identifiability of nonlinear ode models and applications in viral dynamics. SIAM Rev. **53**(1), 3–39 (2011)

Calderhead, B., Girolami, M., Lawrence, N.: Accelerating Bayesian inference over nonlinear differential equations with gaussian processes. Adv. Neural Inf. Process. Syst. **21** (2008)

Dondelinger, F., Husmeier, D., Rogers, S., Filippone, M.: Ode parameter inference using adaptive gradient matching with gaussian processes. In: Artificial Intelligence and Statistics, pp. 216–228 (2013). PMLR

Barber, D., Wang, Y.: Gaussian processes for Bayesian estimation in ordinary differential equations. In: International Conference on Machine Learning, pp. 1485–1493 (2014)

Ghosh, S., Dasmahapatra, S., Maharatna, K.: Fast approximate Bayesian computation for estimating parameters in differential equations. Stat. Comput. **27**(1), 19–38 (2017)

Lazarus, A., Husmeier, D., Papamarkou, T.: Multiphase MCMC sampling for parameter inference in nonlinear ordinary differential equations. In: International Conference on Artificial Intelligence and Statistics, pp. 1252–1260 (2018)

Wenk, P., Gotovos, A., Bauer, S., Gorbach, N.S., Krause, A., Buhmann, J.M.: Fast Gaussian process based gradient matching for parameter identification in systems of nonlinear ODEs. In: International Conference on Artificial Intelligence and Statistics, pp. 1351–1360 (2019)

Yang, S., Wong, S.W.K., Kou, S.C.: Inference of dynamic systems from noisy and sparse data via manifold-constrained gaussian processes. Proc. Natl. Acad. Sci. (2021). https://doi.org/10.1073/pnas.2020397118

Wenk, P., Gotovos, A., Bauer, S., Gorbach, N.S., Krause, A., Buhmann, J.M.: Fast gaussian process based gradient matching for parameter identification in systems of nonlinear odes. In: Chaudhuri, K., Sugiyama, M. (eds.) Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 89, pp. 1351–1360. PMLR, (2019)

Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian monte Carlo methods. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **73**(2), 123–214 (2011)

Skilling, J.: Bayesian solution of ordinary differential equations. In: Maximum Entropy and Bayesian Methods, pp. 23–37. Springer, (1992)

Tronarp, F., Kersting, H., Särkkä, S., Hennig, P.: Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective. Stat. Comput. **29**(6), 1297–1315 (2019)

Krämer, N., Schmidt, J., Hennig, P.: Probabilistic numerical method of lines for time-dependent partial differential equations. In: International Conference on Artificial Intelligence and Statistics, pp. 625–639 (2022). PMLR

Diaconis, P., Holmes, S., Shahshahani, M.: Sampling from a manifold. Advances in modern statistical theory and applications: a Festschrift in honor of Morris L. Eaton, Vol. 10, pp. 102–125 (2013)

Wu, H.: Statistical methods for HIV dynamic studies in AIDS clinical trials. Stat. Methods Med. Res. **14**(2), 118–134 (2005). https://doi.org/10.1191/0962280205sm390oa

Li, L., Brown, M., Lee, K., Gupta, S.: Estimation and inference for a spline-enhanced population pharmacokinetic model. Biometrics **58**(3), 601–11 (2002). https://doi.org/10.1111/j.0006-341x.2002.00601.x

Huang, Y., Liu, H., Wu, D.: Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. Biometrics **62**(2), 413–423 (2006). https://doi.org/10.1111/j.1541-0420.2005.00447.x

Cao, J., Huang, J.Z., Wu, H.: Penalized nonlinear least squares estimation of time-varying parameters in ordinary differential equations. J. Comput. Graph. Stat. **21**(1), 42–56 (2012)

Pei, S., Shaman, J.: Initial simulation of SARS-CoV2 spread and intervention effects in the continental us. MedRxiv (2020)

Shaman, J., Karspeck, A.: Forecasting seasonal outbreaks of influenza. Proc. Natl. Acad. Sci. **109**(50), 20425–20430 (2012)

Schmidt, J., Krämer, N., Hennig, P.: A probabilistic state space model for joint inference from differential equations and data. Adv. Neural Inf. Process. Syst. **34**, 12374–12385 (2021)

Pokern, Y., Stuart, A.M., van Zanten, J.H.: Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDES. Stoch. Process. Their Appl. **123**(2), 603–628 (2013)

Papaspiliopoulos, O., Pokern, Y., Roberts, G.O., Stuart, A.M.: Nonparametric estimation of diffusions: a differential equations approach. Biometrika **99**(3), 511–531 (2012)

Hairer, M., Stuart, A.M., Voss, J.: Signal Processing Problems on Function Space: Bayesian Formulation, Stochastic PDEs and Effective MCMC Methods. Oxford University Press, Oxford (2011)

Cotter, S.L., Dashti, M., Stuart, A.M.: Approximation of Bayesian inverse problems for PDEs. SIAM J. Numer. Anal. **48**(1), 322–345 (2010)

Chen, Y., Hosseini, B., Owhadi, H., Stuart, A.M.: Solving and learning nonlinear PDEs with gaussian processes. J. Comput. Phys. **447**, 110668 (2021)

Bayarri, M., Berger, J., Liu, F.: Modularization in Bayesian analysis, with emphasis on analysis of computer models. Bayesian Anal. **4**(1), 119–150 (2009)

Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)

Neal, R.M.: MCMC using Hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G., Meng, X.L. (eds.) Handbook of Markov Chain Monte Carlo. Handbooks of Modern Statistical Methods, pp. 113–162. CRC Press, Boca Raton (2011)

Yang, W., Kandula, S., Huynh, M., Greene, S.K., Van Wye, G., Li, W., Chan, H.T., McGibbon, E., Yeung, A., Olson, D., et al.: Estimating the infection fatality risk of COVID-19 in New York city, March 1-May 16, 2020. MedRxiv (2020)

Hethcote, H.W.: The mathematics of infectious diseases. SIAM Rev. **42**(4), 599–653 (2000)

Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., Wang, C.: Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. Nature **584**(7821), 420–424 (2020)

Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. **20**(5), 533–534 (2020)

Mwalili, S., Kimathi, M., Ojiambo, V., Gathungu, D., Mbogo, R.: SEIR model for COVID-19 dynamics incorporating the environment and social distancing. BMC Res. Notes **13**(1), 1–5 (2020)

Goel, N.S., Maitra, S.C., Montroll, E.W.: On the Volterra and other nonlinear models of interacting populations. Rev. Modern Phys. **43**(2), 231 (1971)

Donald, D.B., Stewart Anderson, R.: Resistance of the prey-to-predator ratio to environmental gradients and to biomanipulations. Ecology **84**(9), 2387–2394 (2003)

Chen, J., Wu, H.: Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. J. Am. Stat. Assoc. **103**(481), 369–384 (2008). https://doi.org/10.1198/016214507000001382

Huang, Y., Rosenkranz, S.L., Wu, H.: Modeling HIV dynamics and antiviral response with consideration of time-varying drug exposures, adherence and phenotypic sensitivity. Math. Biosci. **184**(2), 165–186 (2003)

Cuenod, C.-A., Favetto, B., Genon-Catalot, V., Rozenholc, Y., Samson, A.: Parameter estimation and change-point detection from dynamic contrast enhanced MRI data using stochastic differential equations. Math. Biosci. **233**(1), 68–76 (2011)

Xun, X., Cao, J., Mallick, B., Maity, A., Carroll, R.J.: Parameter estimation of partial differential equation models. J. Am. Stat. Assoc. **108**(503), 1009–1020 (2013)

Kou, S.C., Xie, X.S.: Generalized Langevin equation with fractional gaussian noise: subdiffusion within a single protein molecule. Phys. Rev. Lett. **93**(18), 180603 (2004)