Stat Comput (2018) 28:1053–1072
https://doi.org/10.1007/s11222-017-9779-x

Trajectory inference and parameter estimation in stochastic models with temporally aggregated data

Maria Myrto Folia[1] · **Magnus Rattray**[1]

Received: 3 May 2017 / Accepted: 22 September 2017 / Published online: 24 October 2017

Abstract Stochastic models are of fundamental importance in many scientific and engineering applications. For example, stochastic models provide valuable insights into the causes and consequences of intra-cellular fluctuations and inter-cellular heterogeneity in molecular biology. The chemical master equation can be used to model intra-cellular stochasticity in living cells, but analytical solutions are rare and numerical simulations are computationally expensive. Inference of system trajectories and estimation of model parameters from observed data are important tasks and are even more challenging. Here, we consider the case where the observed data are aggregated over time. Aggregation of data over time is required in studies of single cell gene expression using a luciferase reporter, where the emitted light can be very faint and is therefore collected for several minutes for each observation. We show how an existing approach to inference based on the linear noise approximation (LNA) can be generalised to the case of temporally aggregated data. We provide a Kalman filter (KF) algorithm which can be combined with the LNA to carry out inference of system variable trajectories and estimation of model parameters. We apply and evaluate our method on both synthetic and real data scenarios and show that it is able to accurately infer the posterior distribution of model parameters in these examples. We demonstrate how applying standard KF inference to aggregated data without accounting for aggregation will tend to underestimate the process noise and can lead to biased parameter estimates.

✉ Magnus Rattray
magnus.rattray@manchester.ac.uk

[1] Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

Keywords Linear noise approximation · Stochastic systems biology · Time aggregation · Kalman filter

1 Introduction

Stochastic differential equations (SDEs) are used to model the dynamics of processes that evolve randomly over time. SDEs have found a range of applications in finance (e.g. stock markets, Hull 2009), physics (e.g. statistical physics, Gardiner 2004) and biology (e.g. biochemical processes, Wilkinson 2011). Usually, the coefficients (model parameters) of SDEs are unknown and have to be inferred using observations from the systems of interest. Observations are typically partial (e.g. collected at discrete times for a subset of variables), corrupted by measurement noise, and may also be aggregated over time and/or space. Given these observed data, our task is to infer the process trajectory and estimate the model parameters.

A motivating example of stochastic aggregated data comes from biology and more specifically from luminescence bioimaging, where a luciferase reporter gene is used for studying gene expression inside a cell (Spiller et al. 2010). The luminescence intensity emitted from the luciferase experiments is collected from single cells and is integrated over a time period (in certain cases up to 30 min, Harper et al. 2011) and then recorded as a single data point. In this paper, we consider the problem of inferring SDE model parameters given temporally aggregated data of this kind.

Imaging data from single cells are highly stochastic due to the low number of reactant molecules and the inherent stochasticity of cellular processes such as gene transcription or protein translation. The chemical master equation (CME) is widely used to describe the evolution of biochemical reactions inside cells stochastically (Gillespie 1992). Exact

inference with the CME is rare and, even when possible, computationally prohibitive. In Golightly and Wilkinson (2005), the authors perform inference using a diffusion approximation of the CME, resulting in a nonlinear SDE. The linear noise approximation (LNA) (Kampen 2007) has been used as an alternative approximation of the CME which is valid for a sufficiently large system (Komorowski et al. 2009; Fearnhead et al. 2014). According to the LNA, the system is decomposed into a deterministic and a stochastic part. The latter is described by a linear SDE of the following form:

$$\mathrm{d}X_t = a_t X_t \mathrm{d}t + b_t \mathrm{d}W_t, \tag{1}$$

where X_t is a d-dimensional process, a_t is a $(d \times d)$-matrix-valued function, W_t is an m-dimensional Wiener process, and b_t a $(d \times m)$ matrix-valued function.

Given an initial condition $X_0 = c$, Eq. (1) has the following known solution (Arnold 1974):

$$X_t = \Phi_t c + \Phi_t \int_{t_0}^{t} \Phi_s^{-1} b_s \mathrm{d}W_s \,, \tag{2}$$

where Φ_t is the fundamental matrix of the homogeneous equation $\mathrm{d}X_t = a_t X_t \mathrm{d}t$. Note that the right integral in Eq. (2) is a Gaussian process, as it is an integral of a non-random function with respect to W_t (Arnold 1974). If we further assume that the initial condition c is normally distributed or constant, Eq. (2) gives rise to a Gaussian process. Additionally, the solution of a (linear) SDE is a Markov process (Arnold 1974). These properties of linear SDEs (of the form of Eq. (1)) are highly desirable when carrying out inference.

The approaches above do not treat the aggregated nature of luciferase data in a principled way but instead assume that the data are proportional to the quantity of interest at the measurement time (Harper et al. 2011; Komorowski et al. 2009). Here, we build on the work of Komorowski et al. (2009) and Fearnhead et al. (2014) and extend it to the case of aggregated data. Since we are using the LNA, the problem is equivalent to a parameter inference problem for the time integral of a linear SDE as in Eq. (1): $\int_{t_0}^{t} X(u)\mathrm{d}u$. We follow a Bayesian approach, where the likelihood of our model is computed using a continuous-discrete Kalman filter (Särkkä 2006) and parameter inference is achieved using an MCMC algorithm. The paper is structured as follows: we first provide a description of the LNA as an approximation of the CME and introduce the integral of the LNA for treating temporally aggregated observations. We then describe a Kalman filter framework for performing inference with the LNA and its integral. Finally, we apply our method in three different examples. The Ornstein–Uhlenbeck process has been picked as a system where we can study its exact solutions. The Lotka–Volterra model was selected as an example of a nonlinear system with partial observations. The translation inhibition model was used to demonstrate our method with real data.

2 The linear noise approximation and its integral

The CME can be used to model biochemical reactions inside a cell. It is essentially a forward Kolmogorov equation for a Markov process that describes the evolution of a spatially homogeneous biochemical system over time.

Assume a biochemical reaction network consisting of N chemical species $\mathcal{X}_1, \ldots, \mathcal{X}_N$ in a volume Ω and v reactions $R_1, \ldots, R_v$. The usual notation for such a network is given below:

$$\begin{aligned}
R_1 &: p_{11}\mathcal{X}_1 + p_{12}\mathcal{X}_2 + \cdots + p_{1N}\mathcal{X}_N \to q_{11}\mathcal{X}_1 + q_{12}\mathcal{X}_2 + \cdots + q_{1N}\mathcal{X}_N \\
R_2 &: p_{21}\mathcal{X}_1 + p_{22}\mathcal{X}_2 + \cdots + p_{2N}\mathcal{X}_N \to q_{21}\mathcal{X}_1 + q_{22}\mathcal{X}_2 + \cdots + q_{2N}\mathcal{X}_N \\
&\ldots \\
R_v &: p_{v1}\mathcal{X}_1 + p_{v2}\mathcal{X}_2 + \cdots + p_{vN}\mathcal{X}_N \to q_{v1}\mathcal{X}_1 + q_{v2}\mathcal{X}_2 + \cdots + q_{vN}\mathcal{X}_N
\end{aligned}$$

where $\boldsymbol{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_N)^T$ represents the number of chemical species (we assume molecules) and $\boldsymbol{x} = \frac{\boldsymbol{X}}{\Omega}$ is the concentration of molecules. We denote with P the $v \times u$ matrix whose elements are given by p_{ij} and Q the $v \times u$ matrix with elements q_{ij}. We define the *stoichiometry matrix* S as $S = (Q - P)^T$. The probability of a reaction taking place in $[t, t + \mathrm{d}t)$ is given by the vector of reaction rates $h_j(x, \Omega, t)\Omega \mathrm{d}t$.

The probability $p(\boldsymbol{X}, t)$ that the system is in state $\boldsymbol{X}$ at time t is given by the CME:

$$\begin{aligned}
\frac{\mathrm{d}p(\boldsymbol{X}, t)}{\mathrm{d}t} = \Omega \sum_{i=1}^{v} [&h_i(\boldsymbol{X} - S^{(i)}, \Omega, t) p(\boldsymbol{X} - S^{(i)}, t) \\
&- h_i(\boldsymbol{X}, \Omega, t) p(\boldsymbol{X}, t)] \,.
\end{aligned} \tag{3}$$

However, as mentioned before, exact inference with the CME, even when possible, is computationally prohibitive. We use the LNA as an approximation of the CME due to its successful application in Komorowski et al. (2009) and Fearnhead et al. (2014). The state of the system $\boldsymbol{X}$ is expected to have a peak around the macroscopic value of order Ω and fluctuations of order $\Omega^{1/2}$ such that $X_t = \Omega\phi_t + \Omega^{1/2}\xi_t$. This way the system is decomposed to the deterministic part ϕ_t and the stochastic part ξ_t. The LNA arises as a Taylor expansion of the CME in powers of the volume Ω; for a detailed derivation the reader is referred to Kampen (2007) and Elf and Ehrenberg (2003). By collecting terms of order $\Omega^{1/2}$, we obtain the deterministic part of the system, namely

the macroscopic rate equations ϕ_i, where i stands for the ith species:

$$\frac{\mathrm{d}\phi_i}{\mathrm{d}t} = S_i h(\phi_t, \Omega, t) \ . \tag{4}$$

Terms of order Ω^0 give us the stochastic part of the system:

$$d\xi_t = A_t \xi_t \mathrm{d}t + E_t \mathrm{d}W \ , \tag{5}$$

where, $A_t = SF_t$ and $F_{ij} = \frac{\partial h_j(\phi_t, \Omega, t)}{\partial \phi_i(t)}$, while $EE_t{}^T = Sdiag(h(\phi_t, \Omega, t))S^T$. Equation (5) is a linear SDE of the form of Eq. (1). Its solution is a Gaussian Markov process, provided that we have an initial condition that is a constant or a Gaussian random variable. The ordinary differential equations (ODEs) that describe the mean and variance of this Gaussian process are given by Arnold (1974):

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = A_t m_t \ , \tag{6}$$

$$\frac{\mathrm{d}V_t}{\mathrm{d}t} = V_t A_t{}^T + A_t V_t + EE_t{}^T \ . \tag{7}$$

Note that if we set the initial condition of $m_0 = 0$, then Eq. (6) will lead to $m_t = 0$ at all times. We will make the assumption that, at each observation point, m_t is reset to zero since it can be beneficial for inference as discussed in Fearnhead et al. (2014) and Giagos (2010).

In what follows we will assume, without loss of generality, that the volume $\Omega = 1$, i.e. the number of molecules equals the concentration of molecules and thus,

$$X_t = \phi_t + \xi_t \ . \tag{8}$$

Equation (8) is the sum of a deterministic and a Gaussian term; consequently, it will also be normally distributed. By taking its expectation and variance, we have that $X_t|X_0 \sim N(\phi_t + m_t, V_t)$ which, according to the initial condition $m_0 = 0$, leads to $X_t|X_0 \sim N(\phi_t, V_t)$.

We are now interested in the integral of Eq. (8), as this will allow us to model the aggregated data,

$$H_t = \int_{t_0}^{t} X_u \mathrm{d}u = \int_{t_0}^{t} \phi_u \mathrm{d}u + \int_{t_0}^{t} \xi_u \mathrm{d}u = I_t + Q_t \ . \tag{9}$$

The deterministic part of this aggregated process is given by $I(t)$, and the stochastic part is given by $Q(t)$. Subsequently, we have the following ODEs:

$$\frac{\mathrm{d}I_t}{\mathrm{d}t} = \frac{d}{\mathrm{d}t}\int_{t_0}^{t} \phi_t \mathrm{d}u = \phi_t \ , \tag{10}$$

$$\frac{dQ_t}{\mathrm{d}t} = \xi_t \ . \tag{11}$$

Here, Q_t will also follow a Gaussian process (as it is the integral of a Gaussian process) so we need to compute its mean and variance. The ODEs for the mean, variance and $\mathbb{E}[Q_t \xi_t^T]$ are given below; their proofs can be found in "Appendix A.1":

$$\frac{\mathrm{d}E[Q_t]}{\mathrm{d}t} = \mathbb{E}[\xi_t] = 0, \tag{12}$$

$$\frac{d\mathrm{Var}[Q_t]}{\mathrm{d}t} = \mathbb{E}[Q_t \xi_t^T] + \mathbb{E}[\xi_t Q_t^T] \ , \tag{13}$$

$$\frac{d\mathbb{E}[Q_t \xi_t^T]}{dt} = \mathbb{E}[Q_t \xi_t^T] A(t)^T + V_t \ . \tag{14}$$

Note that Q_t is not Markovian since knowledge of its history is not sufficient to determine its current state. However, jointly with ξ_t it forms a bivariate Gaussian Markov process, that is characterised by the following linear SDE:

$$d\begin{bmatrix} \xi_t \\ Q_t \end{bmatrix} = \begin{bmatrix} A_t & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \xi_t \\ Q_t \end{bmatrix} dt + \begin{bmatrix} E_t \\ 0 \end{bmatrix} dW_t \ ,$$
$$\begin{bmatrix} \xi_0 \\ Q_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \ . \tag{15}$$

From Eq. (15) we have that ξ_t, Q_t are jointly Gaussian and, consequently, their marginals are also normally distributed. Thus, according to (9) $H_t|H_0, X_0 \sim N(\mu_t, \Sigma_t)$ with $\mu_t = I_t$ and $\Sigma_t = V[Q_t]$.

3 Kalman filter for the LNA and its integral

The classical filtering problem is concerned with the problem of estimating the state of a linear system given noisy, indirect or partial observations (Kalman 1960). In our case, the state is continuous and is described by Eq. (8) while the observations are collected at discrete time points with or without Gaussian noise. For this reason, we refer to it as the continuous-discrete filtering problem (Jazwinski 1970; Särkkä 2006).

First, we consider the case where observations are taken from the process X_t and not from its integral H_t. In that case, the observation process is given by $y_t = P_t X_t + \epsilon_t$ where $\epsilon_t \sim N(0, R)$ and accounts for technical noise. The observability matrix P_t is used to deal with the partial observability of the system, for example, if we have two species X_1, X_2 and we observe only X_1, $P = [1, 0]^{\mathrm{T}}$.

Following the Kalman filter (KF) methodology, we need to define the following quantities:

- Prior: $p(X_0)$.
- Predictive distribution: $p(X_t|y_{1:t-1})$, where $y_{1:t-1}$ refers to the observations at discrete points up to time $t-1$.
- Posterior or Update distribution: $p(X_t|y_{1:t})$.

The predictive distribution is given by $X_t|y_{1:t-1} \sim N(\mu_{1t}^-, V_t^-)$, where μ_{1t}^- and V_t^- are found by integrating

forward for $[t, t-1]$ Eqs. (4) and (7) initialised at the posterior mean μ_{1t-1} and variance V_{t-1}. In our case, the mean of the stochastic part is initialised at 0, so μ_{1t} corresponds to the deterministic part ϕ_t. By updating the deterministic solution at each observation point, we achieve a better estimate, as the ODE solution can become a poor approximation over long periods of time. The posterior distribution $p(X_t|y_{1:t}) = N(\mu_{1t}, V_t)$ corresponds to the standard posterior distribution of a discrete KF and the updated μ_{1t} and V_t are given in "Appendix A.3". This case has been thoroughly studied in Fearnhead et al. (2014).

We consider now the case where the state X_t is being observed through the integrated process H_t, such that the observation process is given by $y_t = P_t H_t + \epsilon_t$ and $\epsilon_t \sim N(0, R)$. Again, we need to define a prior distribution as well as calculate the predictive and posterior distributions for the system that we are studying.

The predictive distribution of our system is given by $p\left(\begin{bmatrix} X_t \\ H_t \end{bmatrix} | y_{1:t-1}\right) = N\left(\begin{bmatrix} \mu_{1t}^- \\ \mu_{2t}^- \end{bmatrix}, \begin{bmatrix} V_t^- & C_t^{-T} \\ C_t^- & \Sigma_t^- \end{bmatrix}\right)$, where $C_t = \mathbb{E}[Q_t M_t^T]$. For this step, we need to integrate forward the ODEs (4), (10), (7), (13) and (14) with the appropriate initial conditions as seen in Algorithm 1. Note that the integrated process H_t needs to be reset to 0 at each observation point in order to capture correctly the 'area under graph' of the underlying process X_t.

To compute the posterior distribution $p(X_t|y_{1:t})$, we look at the joint distribution of (H_t, X_t, y_t) conditioned on $y_{1:t-1}$:

$$\begin{bmatrix} X_t \\ H_t \\ y_t \end{bmatrix} | y_{1:t-1} \sim N\left(\begin{bmatrix} \mu_{1t}^- \\ \mu_{2t}^- \\ P_t\mu_{2t}^- \end{bmatrix}, \begin{bmatrix} V_t^- & C_t^{-T} & C_t^{-T}P_t^T \\ C_t^- & \Sigma_t^- & \Sigma_t^- P_t^T \\ P_tC_t^- & P_t\Sigma_t^- & P_t\Sigma_t^- P_t^T + R_t \end{bmatrix}\right) \tag{16}$$

By using the lemma in "Appendix A.2" and using the corresponding blocks of the joint distribution (16), we can calculate the posterior mean and variance of $p(X_t|y_{1:t})$:

$$\begin{aligned} \mu_{1t} &= \mu_{1t}^- + P_t C_t^{-T}(P_t\Sigma_t^- P_t^T + R_t)^{-1}(y_t - P_t\mu_{2t}^-)\,, \\ V_t &= V_t^- - P_t C_t^{-T}(P_t\Sigma_t^- P_t^T + R_t)^{-1} P_t C_t^-\,. \end{aligned} \tag{17}$$

Since we are interested in parameter inference, we will need to compute the likelihood $L(\theta)$ of the system, where θ represents the parameter vector of the system:

$$L(\theta) = p(y_1|\theta)\prod_{i=2}^{t} p(y_i|y_{1:i-1}, \theta)\,. \tag{18}$$

The individual terms of the likelihood are given by $p(y_t|y_{1:t-1}) = N(P_t\mu_{2t}^-, P_t\Sigma_t^- P_t^T + R_t)$. Parameter inference is then straightforward either by using a numerical technique such as the Nelder–Mead algorithm to obtain the maximum likelihood (ML) parameters or using a Bayesian method such as a Metropolis-Hastings (MH) algorithm. The general procedure for performing inference using aggregated data is summarised in Algorithm 1.

Algorithm 1 Kalman Filter for the integrated LNA

procedure LIKELIHOOD($y_{1:T}, \theta$)
  Initialisation ($t = 0$) *Set prior* $X_0 \sim N(\mu_{1(t=0)}^-, V_{t=0}^-)$ and $prod \leftarrow 1$.
  Set initial conditions for the system of ODEs $\phi_0 = \mu_{1(t=0)}^-$, $V_0 = V_{t=0}^-$, $\mu_{2(t=0)} = 0$, $\Sigma_0 = 0$, $C_0 = 0$.
  loop:
  Solve the ODEs (4), (7), (10), (13), (14) s.t. the initial conditions for $[t-1, t]$ *to obtain* μ_{1t}^-, V_t^-, μ_t^-, Σ_t^-, C_t^- .
  Calculate $p(y_t|y_{1:t-1}, \theta)$.
  $prod \leftarrow prod * p(y_t|y_{1:t-1}, \theta)$.
  Reset initial conditions according to (17): $\phi_t = \mu_{1t}^- + C_t^{-T}P_t^T(P_t\Sigma_t^- P_t^T + R_t)^{-1}(y_t - P_t\mu_t^-)$, $V_t = V_t^- - P_tC_t^{-T}(P_t\Sigma_t^- P_t^T + R_t)^{-1}P_tC_t^-$, $\mu_t = 0$, $\Sigma_t = 0$, $C_t = 0$.
  Set $t = t + 1$
  if $t < T$ **goto** *loop* .
  Return prod
end procedure

4 The Ornstein–Uhlenbeck process

We first investigate the effect of integration in a one-dimensional, zero-mean OU process of the following form:

$$\mathrm{d}X_t = -\alpha X_t \mathrm{d}t + \sigma \mathrm{d}W_t, \tag{19}$$

where α is the drift or decay rate of the process and σ is the diffusion constant. Both of these parameters are assumed to be unknown, and we will try to infer them using the KF scheme that we have developed.

The OU process is a special case of a linear SDE (Eq. (1)), since its coefficients are time invariant, resulting in a stationary Gaussian–Markov process. Analytical solutions for both the OU and its integral exist (Gillespie 1996) and are presented in "Appendix A.4". The results for the mean m_t and variance V_t of the OU, where $\Delta = t - t_0$, are given below:

$$m_t = m_0 e^{-\alpha\Delta}\,, \tag{20a}$$

$$V_t = e^{-2\alpha\Delta}V_0 + \frac{\sigma^2}{2\alpha}\left(1 - e^{-2\alpha\Delta}\right). \tag{20b}$$

The integral of Eq. (19) is given by $\mathrm{d}Y_t = X_t\mathrm{d}t$, and the mean, variance and covariance are given below,

$$\mathbb{E}[y_t] = \frac{m_0}{\alpha}(1 - e^{-\alpha\Delta}) , \tag{21a}$$

$$\mathrm{Cov}(X_t, Y_t) = \frac{\sigma^2}{2\alpha^2} + \left(-\frac{\sigma^2}{\alpha^2} + \frac{V_0}{\alpha}\right) e^{-\alpha\Delta} + \left(\frac{\sigma^2}{2\alpha^2} - \frac{V_0}{\alpha}\right) e^{-2\alpha\Delta} , \tag{21b}$$

$$\mathrm{Var}[y_t] = \frac{\sigma^2}{\alpha^2}\Delta + \left(\frac{\sigma^2}{2\alpha^3} - \frac{V_0}{\alpha^2}\right)\left(1 - e^{-2\alpha\Delta}\right) + 2\left(-\frac{\sigma^2}{\alpha^3} + \frac{V_0}{\alpha^2}\right)(1 - e^{-\alpha\Delta}) . \tag{21c}$$

We are interested in inferring the parameters α and σ given observations from Y_t at discrete times, where the interval Δ between two observations is constant. We will compare two approaches. First, we will assume that the data come directly from X_t ignoring their aggregated nature and use the standard discrete–continuous KF, referred to as KF1. To make the comparison of this scenario fairer, we will normalise the observations by dividing with Δ, which brings the observation close to an average value of the process, in an attempt to match the observations to data generated from the process X_t. In the second case, we will use the KF on the integrated process in analogy with Algorithm 1, which we will refer to as KF2. The case of inferring the parameters of an OU process using non-aggregated data with an MCMC algorithm has already been studied in Mbalawata et al. (2013).

X_t will reach its stationary distribution after a time of order $\frac{1}{\alpha}$, which is given by $N(0, \frac{\sigma^2}{2\alpha})$ (Gillespie 1992). However, the integrated process Y_t is non-stationary since $\mathrm{Var}[y_t] \to \infty$ as $\Delta \to \infty$. This already shows us that the two processes behave differently.

Since we are going to use the normalised observations from Y_t with KF1, we will take a look at the normalised process $Z_t = \frac{1}{\Delta}Y_t$:

$$\mathbb{E}[z_t] = \mathbb{E}[\frac{1}{\Delta}Y_t] = \frac{1}{\Delta}\mathbb{E}[y_t] = \frac{m_0}{\alpha\Delta}(1 - e^{-\alpha\Delta}) , \tag{22a}$$

$$\mathrm{Var}[z_t] = \mathrm{Var}[\frac{1}{\Delta}Y_t] = \frac{1}{\Delta^2}\mathrm{Var}[y_t] = \frac{\sigma^2}{\alpha^2\Delta} + \frac{1}{\Delta^2}(\frac{\sigma^2}{2\alpha^3} - \frac{V_0}{\alpha^2})(1 - e^{-2\alpha\Delta}) + \frac{2}{\Delta^2}(-\frac{\sigma^2}{\alpha^3} + \frac{V_0}{\alpha^2})(1 - e^{-\alpha\Delta}) . \tag{22b}$$

By taking the limit as $\Delta \to \infty$ in Eq. (22) and using L'Hospital's rule we can show that $\mathbb{E}[z_t] \to 0$ and $\mathrm{Var}[z_t] \to 0$. So, the normalised process is again not approaching the stationary distribution of X_t.

We have generated aggregated data from the integral of an OU process with $\alpha = 4$ and $\sigma = 2$. To simulate data from Y_t, we need to first simulate data from X_t. This can be done in general by discretising the process and using the Euler–Maruyama algorithm. However, in the case of the OU process, we can also use an exact updating formula (see "Appendix A.6"). The aggregated data can then be collected using the discretised form $Y_{t+dt} = Y_t + X_t dt$ or a numerical integration method such as the trapezoidal rule over the indicated integration period. In "Appendix A.12" we have included plots of the OU process and the corresponding aggregated process.

Table 1 Mean posterior ± 1 s.d. for α and σ using a Metropolis-Hastings algorithm

Δ	KF	α	σ
0.1	KF1	3.023 ± 0.235	1.891 ± 0.135
0.5	KF1	1.905 ± 0.141	1.256 ± 0.095
1.0	KF1	1.420 ± 0.102	0.868 ± 0.068
2.0	KF1	1.022 ± 0.075	0.540 ± 0.044
0.1	KF2	4.022 ± 0.295	2.113 ± 0.159
0.5	KF2	4.092 ± 0.335	2.311 ± 0.206
1.0	KF2	3.865 ± 0.368	2.234 ± 0.240
2.0	KF2	3.704 ± 0.513	2.082 ± 0.307

Data were simulated from an OU process with $\alpha = 4$ and $\sigma = 2$

We tested inference using KF1 with normalised data and KF2 with aggregated data. Results of parameter estimation using a standard random walk MH algorithm are presented in Table 1. Improper uniform priors over infinite range have been used on the log parameters, while different time intervals Δ have been considered. For each interval Δ, we have sampled 100 observations from a single trajectory of an OU process with $\alpha = 4$ and $\sigma = 2$ aggregated over the specified Δ. For this example, we have assumed no observation noise. MCMC traceplots of α and σ can be found in "Appendix A.13" (Figs. 6, 7) which indicate a good mixing of the chain and fast convergence. All chains were run for 50K iterations and 30K were discarded as burn-in. To verify the validity of the results, we have run nine more datasets, separately each time. An average over the ten datasets can be found in "Appendix A.7" (Table 5). As we can see, the estimates for KF1 deteriorate for larger Δ. This is expected since the aggregated process diverges further from the OU process as Δ increases. Estimates remain good for KF2 even when Δ is large, although they become more uncertain, as can be witnessed by the increased standard deviations. Filtering results for KF1 and KF2 with aggregated data using the estimated parameter results for $\Delta = 1$ are given in "Appendix A.14".

It is of interest to investigate the inferred stationary variance of the OU process using KF1 and KF2. We have plotted the inferred stationary variances obtained by the MH for both KF1 and KF2 in Fig. 1. The boxplots are obtained using the average of 10 different datasets and correspond again to an OU process with $\alpha = 4$ and $\sigma = 2$, thus giving rise to a stationary variance of $\frac{\sigma^2}{2\alpha} = 0.5$. When using the normalised aggregated data directly with KF1, we infer the wrong sta-

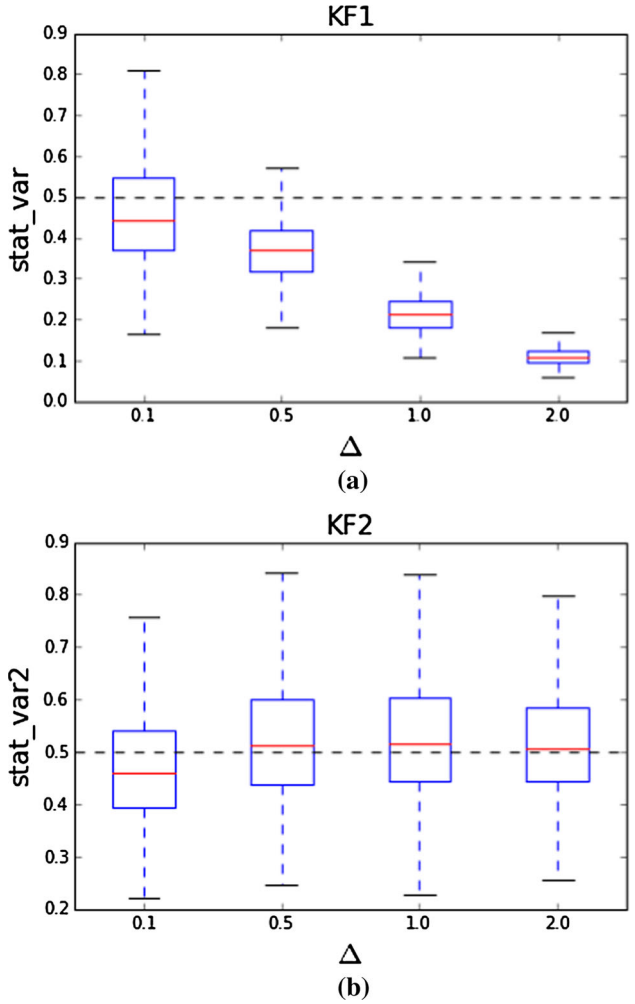


Fig. 1 Boxplots of inferred stationary variance of the OU process for different Δ. The simulated OU process has $\alpha = 4$ and $\sigma = 2$ corresponding to a stationary variance of 0.5, as indicated by the dotted horizontal line. The inferred stationary variance using KF1 tends to zero as Δ grows, but the stationary variance from KF2 is inferred correctly at all Δ. **a** Boxplots of inferred stationary variance for different Δ using KF1. **b** Boxplots of inferred stationary variance for different Δ using KF2

tionary variance of the underlying OU process which tends to zero as Δ becomes larger, consistent with the theoretical results from Eq. (22). Intuitively, we can attribute this behaviour to the fact that aggregated data have relatively smaller fluctuations, so that KF1 will tend to underestimate the process variance.

In this section, we have looked at an example of inferring the parameters of an SDE using aggregated data, and we have found that to obtain accurate results we need to explicitly model the aggregated process. As the observation intervals become larger, there is a greater mismatch between KF1 and KF2. In the next two sections, we will look at examples of more complex stochastic systems that must be approximated by the LNA and compare again inference results using KF1 and KF2.

5 Lotka–Volterra model

We are now going to look at a system of two species that interact with each other according to three reactions

$$X_1 \xrightarrow{\theta_1} 2X_1 \text{ (prey production)} \tag{23a}$$

$$X_1 + X_2 \xrightarrow{\theta_2} 2X_2 \text{ (predator production)} \tag{23b}$$

$$X_2 \xrightarrow{\theta_3} \varnothing \text{ (predator death)} \tag{23c}$$

The model represented by the biochemical reaction network (23) is known as the Lotka–Volterra model, with X_1 representing prey species and X_2 predator species. Although a simple model, it has been used as a reference model (Boys et al. 2008; Fearnhead et al. 2014) since it consists of two species, making it possible to observe it partially through one of the species and also provides a simple example of a nonlinear system.

The LNA can be used to approximate the dynamics and the resulting ODEs can be found in "Appendix A.8". We want to compare parameter estimation results using KF1 and KF2. We collected aggregated data from a Lotka–Volterra model using the Gillespie algorithm. We assumed a known initial population of 10 prey species and 100 predator species. The parameters of the system for producing the synthetic data were set to $(\theta_1, \theta_2, \theta_3) = (0.5, 0.0025, 0.3)$, following (Boys et al. 2008). We have added Gaussian noise with standard deviation set to 3.0, and we assumed that the noise level was known for inference. Our goal was to infer the three parameters $(\theta_1, \theta_2, \theta_3)$ of the system using aggregated observations solely from the predator population.

The Gillespie algorithm was run for 20 min. Data were aggregated and collected every 2 min resulting in 10 observations per sample. To infer the parameters, we assumed that we had 40 independent samples available. Since we assumed independence between the samples, we worked with the product of their likelihoods. In the ideal case of having complete data of a stochastic kinetic model the likelihood is conjugate to an independent gamma prior for the rate constants (Wilkinson 2011). The choice of Ga(2,10) with shape $= 2$ and rate $= 10$ gives a reasonable range for all three parameters and has also been used by Fearnhead et al. (2014). However, in this case the choice of prior is not important as the data dominate the posterior. We have run the same experiment using uninformative exponential priors Exp(10^{-4}) that resulted in equivalent posterior distributions. Since we know that we want all parameters to be positive, we worked with a log transformation. MCMC convergence in this example is relatively slow and adaptive MCMC (Sherlock et al. 2010) was found to speed up convergence (see "Appendix A.9" for details). The adaptive MCMC was run for 30K iterations with 10K regarded as burn-in. The MCMC was initialised at random values sampled from uniform distributions. Parame-

Table 2 Mean posterior ± 1 s.d. for $\theta_1, \theta_2, \theta_3$ using an adaptive MCMC

θ	Ground truth	KF1	KF2
θ_1	0.5	0.480 ± 0.006	0.494 ± 0.005
θ_2	0.0025	$0.0023 \pm 5 \times 10^{-5}$	$0.0025 \pm 5 \times 10^{-5}$
θ_3	0.3	0.243 ± 0.010	0.298 ± 0.010

Data were simulated from a Lotka–Volterra model according to the ground truth values

ter estimation results for all three parameters using adaptive MCMC are shown in Table 2, while Fig. 2 shows histograms of their posterior densities. The ground truth value for each parameter is indicated by a vertical blue line. We can see that only the posterior histograms corresponding to KF2 include the correct estimate for all three parameters in their support. In "Appendix A.15", we have included traceplots of the MCMC runs for all three parameters, where we can see that the adaptive MCMC leads to a fast convergence for both KF1 and KF2. In order to verify the validity of our results, we have run an extra 100 datasets, each consisting of 40 independent samples and obtained point estimates from KF1 and KF2 using the Nelder–Mead algorithm. The results can be found in "Appendix A.10" and agree with our previous conclusion that inference with KF1 gives inaccurate estimates.

Assuming knowledge of the parameter values, we can also use the KF for trajectory inference. In Fig. 3, we demonstrate filtering results for the prey population assuming that we have aggregated data. We simulated a trajectory using $\theta_1 = 0.5, \theta_2 = 0.0025, \theta_3 = 0.3$ and sampled aggregated data every 2 min. Black lines represent the true trajectory of the populations. We see that the inferred credible region with KF1 does not contain the true underlying trajectory in many places. Note that red dots correspond to normalised (aggregated) observations for KF1 and aggregated observations for KF2, so they do not have the same values. In "Appendix A.16", we include filtering results for the unobserved predator population.

6 Translation inhibition model

In this example, we are interested in inferring the degradation rate of a protein from a translation inhibition experiment. We model the translation inhibition experiment by the following set of reactions where R stands for mRNA and P for protein:

$$R \xrightarrow{c_P} R + P \text{ (translation)} \tag{24a}$$

$$P \xrightarrow{d_P P/\Omega} \oslash \text{ (protein degradation)} \tag{24b}$$

The LNA is used, again, as an approximation of the dynamics and the resulting system of ODEs can be found in "Appendix A.11". Before applying our method to real data from this system, we test the performance on synthetic data simulated using the Gillespie algorithm. We simulated 30 time series (corresponding to 30 different cells), assuming

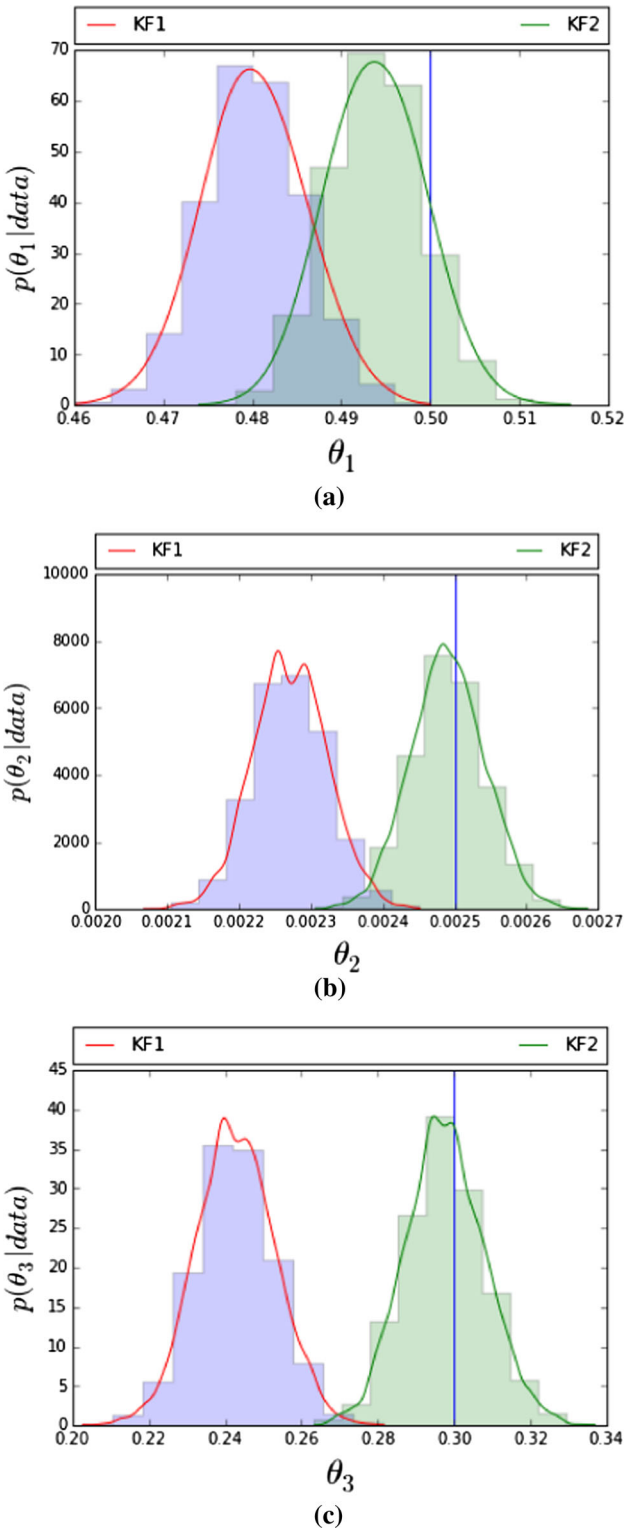


Fig. 2 Posterior densities of $\theta_1, \theta_2, \theta_3$ from aggregated data using KF1 (red histogram) and KF2 (green histogram). **a** Posterior density of θ_1. **b** Posterior density of θ_2. **c** Posterior density of θ_3

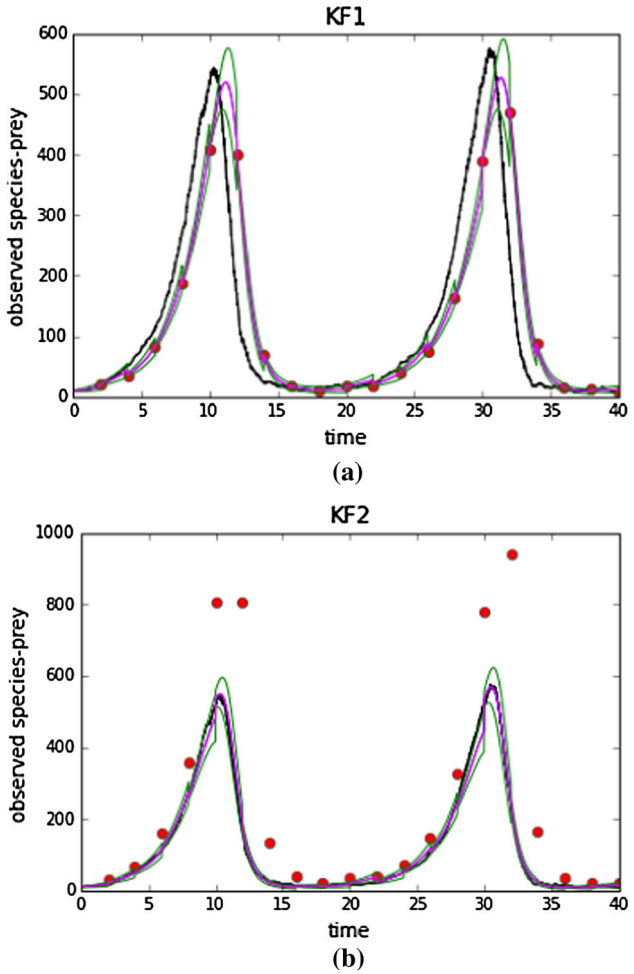


Fig. 3 Filtering results for the prey population. Red dots correspond to aggregated observations for KF2 and normalised observations for KF1. The black line represents the actual process. Purple lines represent the mean estimate and green 1 s.d. . **a** Filtering results for the prey population using KF1. **b** Filtering results for the prey population using KF2

the following values as the ground truth for the kinetic parameters: $c_P = 200$ and $d_P = 0.97$. We further set the initial protein abundance of m_0 to 400 molecules. We have scaled the data by a factor $k = 0.03$, so that they are proportional to the original synthetic data and added Gaussian noise with a variance of $s = 0.1$. For this study, we have assumed that data were integrated over 30 min.

Again we use an adaptive MCMC algorithm (Sherlock et al. 2010). Non-informative exponential priors with mean 10^4 were placed on all parameters. We have adopted the parametarisation used in Komorowski et al. (2009) and Finkenstädt et al. (2013) such as $\widetilde{c}_P = k \cdot c_p$ and $\widetilde{m}_0 = k \cdot m_0$ and worked in the log parameter space. Parameter estimation results for the vector (c_p, d_p, s, k, m_0) using KF1 and KF2 are summarised in Table 3. As we can see, the degradation rates are successfully inferred by both approaches. However, using KF1 leads to an overestimation of m_0 and an underestimation of the noise level s, which corresponds to a smoother process than the underlying one. MCMC traces from both KF1 and KF2 are presented in Fig. 11.

Table 3 Mean posterior $\pm$ 1 s.d. for (c_P, d_P, s, k, m_0) using an adaptive MCMC

c	GT	KF1	KF2
c_p	200	254.152 ± 23.3329	196.9065 ± 25.6251
d_p	0.97	0.9822 ± 0.0364	0.9974 ± 0.0433
s	0.1	0.0349 ± 0.0251	0.0995 ± 0.0093
k	0.03	0.0236 ± 0.0017	0.0312 ± 0.0039
m_0	400	588.9959 ± 44.0205	392.5980 ± 49.0594

Data were simulated from a translation inhibition model according to the ground truth (GT) values

Table 4 Mean posterior $\pm$ 1 s.d. for (c_P, d_P, s, k, m_0) using adaptive MCMC with single cell data obtained from a subset of 11 pituitary cells from a translation inhibition experiment (Harper et al. 2011)

c	KF1	KF2
c_p	217.2987 ± 33.5441	169.9254 ± 43.1153
d_p	1.1020 ± 0.0767	1.2037 ± 0.1046
s	0.0026 ± 0.0026	0.0081 ± 0.0038
k	0.0255 ± 0.0029	0.0373 ± 0.0088
m_0	449.7679 ± 53.9760	278.2987 ± 70.6582

We then applied our model to single cell luciferase data from a subset of 11 pituitary cells (Harper et al. 2011). Parameter estimation results using the same adaptive MCMC are summarised in Table 4. The MCMC was run for 100K iterations out of which 60K were discarded as burn-in. Again, we observe that, using KF1, we get a higher m_0 and a slightly lower noise level s. Posterior histograms of the degradation rates are shown in Fig. 4. A deterministic approach for fitting the data would give a degradation rate of around 1.02 and, as we can see, this value is included in both histograms of Fig. 4. To check convergence using the Gelman–Rubin statistic, we have run 3 different chains with different initialisations. MCMC traces for both KF1 and KF2 are shown in "Appendix A.18" (Fig. 12 and 13) where we can see that the three chains are very close to each other, corresponding to a Gelman–Rubin statistic close to 1.

7 Discussion

We have presented a Bayesian framework for doing inference using aggregated observations from a stochastic process. Motivated by a systems biology example, we chose to use the LNA to approximate the dynamics of the stochastic system, leading to a linear SDE. We then developed a Kalman filter that can deal with integrated, partial and noisy data. We

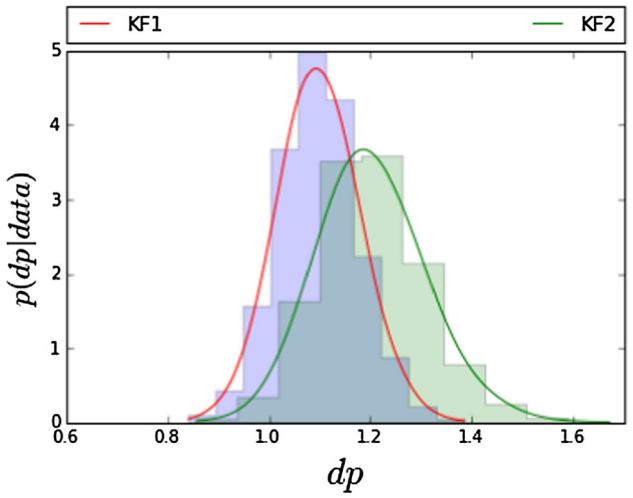


Fig. 4 Posterior histograms of degradation rate using KF1 and KF2

have compared our new inference procedure to the standard Kalman filter which has previously been applied in systems biology applications approximated using the LNA. Overall, we conclude that the aggregated nature of data should be considered when modelling data, as aggregation will tend to reduce fluctuations and therefore the stochastic contribution of the process may be underestimated.

In Sect. 4, we described the different properties of a stochastic process and its integral in the case of the Ornstein–Uhlenbeck process. We showed that one cannot simply treat the integrated observations as proportional to observations coming from the underlying unintegrated process when carrying out inference. As the aggregation time window increases, parameter estimates using this approach become less accurate and the inferred stationary variance of the process is underestimated. In contrast, our modified KF is able to accurately estimate the model parameters and stationary variance of the process.

In Sect. 5, we have demonstrated the ability of our method to give more accurate results in a Lotka–Volterra model given synthetic aggregated data. In Sect. 6, we looked at a real-world application with data from a translation inhibition experiment carried out in single cells. As the LNA depends on its deterministic part, and in a deterministic system integration is dealt with reasonably well using the simple proportionality constant approach, some of the system parameters, such as the degradation rate, can be inferred reasonably well by the standard non-aggregated data approach. However, neglecting the aggregated nature of the data does lead to a significantly larger estimate of the initial population of molecules even in this simple application. This is consistent with our observation that neglecting aggregation will tend to underestimate the scale of fluctuations as it is the number of molecules that determines the size of fluctuations in this example. In models where noise plays a more critical role, e.g. systems with noise-induced oscillations, the effect of parameter misspecification could have more serious consequences on model-based inferences.

Our proposed inference method can deal with the intrinsic noise inside a cell, measurement noise and temporal aggregation. However, cell populations are highly heterogeneous, and cell-to-cell variability has not been considered in our current inference scheme. It would be possible to deal with cell-to-cell variability using a hierarchical model (Finkenstädt et al. 2013) which could be combined with the integrated data Kalman Filter developed here.

All experiments were carried out on a cluster of 64bit Ubuntu machines with an i5-3470 CPU @ 3.20 GHz x 4 processor and 8 GB RAM. All scripts were run in Spyder (Anaconda 2.5.0, Python 2.7.11, Numpy 1.10.4). Code reproducing the results of the experiments can be found on GitHub https://github.com/maria-myrto/inference-aggregated.

Acknowledgements MR was funded by the UK's Medical Research Council (award MR/M008908/1).

A Appendix

A.1 Mean and variance of the integrated process

We start by computing $\mathbb{E}[Q_t]$, i.e. the **mean** of Q_t. We know that:

$$d\xi_t = A_t\xi_t \mathrm{d}t + E_t \mathrm{d}W\,, \tag{25}$$

$$dQ_t = \xi_t \mathrm{d}t \Leftrightarrow Q_{t+dt} = Q_t + \xi_t \mathrm{d}t\,. \tag{26}$$

Averaging Eq. (26), dividing by $\mathrm{d}t$ and letting $\mathrm{d}t \to 0$, gives us:

$$\begin{aligned} \mathbb{E}[Q_{t+\mathrm{d}t}] &= \mathbb{E}[Q_t] + \mathbb{E}[\xi_t]\mathrm{d}t \\ \mathbb{E}[Q_{t+\mathrm{d}t}] - \mathbb{E}[Q_t] &= \mathbb{E}[\xi_t]\mathrm{d}t \\ \frac{dE[Q_t]}{\mathrm{d}t} &= \mathbb{E}[\xi_t] = 0 \end{aligned} \tag{27}$$

The mean of Q_t is set to zero, as we have chosen to use the Restarting LNA.

We now need to compute the **covariance** between Q_t and ξ_t. Again $\mathbb{E}[Q_t] = 0$ and $\mathbb{E}[\xi_t] = 0$ since we are using the Restarting LNA and thus, the covariance is given by $\mathbb{E}[Q_t\xi_t^T]$. For our derivation, we need to use:

$$\xi_{t+\mathrm{d}t}^T = \xi_t^T + \xi_t^T A_t^T \mathrm{d}t + E_t^T \mathrm{d}W_t\,. \tag{28}$$

By multiplying Eqs. (26) and (28) we get:

$$\begin{aligned} Q_{t+\mathrm{d}t}\xi_{t+\mathrm{d}t}{}^T &= (Q_t + \xi_t \mathrm{d}t)(\xi_t^T + \xi_t^T A^T \mathrm{d}t + E_t^T \mathrm{d}W_t) \\ &= Q_t\xi_t^T + Q_t\xi_t^T A_t^T \mathrm{d}t + Q_t E_t^T \mathrm{d}W_t + \\ &\quad + \xi_t\xi_t^T \mathrm{d}t + \xi_t\xi_t^T A_t^T \mathrm{d}t\mathrm{d}t + \xi_t E_t^T \mathrm{d}t\mathrm{d}W_t \,. \end{aligned} \tag{29}$$

Averaging the result (29), retaining terms up to first order in $\mathrm{d}t$, dividing by $\mathrm{d}t$ and letting $\mathrm{d}t \to 0$, we get:

$$\begin{aligned} \mathbb{E}[Q_{t+\mathrm{d}t}\xi_{t+\mathrm{d}t}{}^T] &= \mathbb{E}[Q_t\xi_t^T] + \mathbb{E}[Q_t\xi_t^T]A_t^T \mathrm{d}t \\ &\quad + \mathbb{E}[Q_t \mathrm{d}W_t]E_t^T + \mathbb{E}[\xi_t\xi_t^T]\mathrm{d}t \,, \\ \frac{d\mathbb{E}[Q_t\xi_t^T]}{\mathrm{d}t} &= \mathbb{E}[Q_t\xi_t^T]A(t)^T + \mathbb{E}[\xi_t\xi_t^T] \,, \\ \frac{d\mathbb{E}[Q_t\xi_t^T]}{\mathrm{d}t} &= \mathbb{E}[Q_t\xi_t^T]A(t)^T + V_t \,. \end{aligned} \tag{30}$$

The **variance** of Q_t is given by $\mathrm{Var}[Q_t] = \mathbb{E}[Q_t Q_t^T]$ since $\mathbb{E}[Q_t] = 0$. We have that,

$$\begin{aligned} Q_{t+\mathrm{d}t}\, Q_{t+\mathrm{d}t}{}^T &= (Q_t + \xi_t \mathrm{d}t)(Q_t + \xi_t \mathrm{d}t)^T \,, \\ Q_{t+\mathrm{d}t}\, Q_{t+\mathrm{d}t}{}^T &= Q_t Q_t^T + Q_t\xi_t^T \mathrm{d}t + \xi_t Q_t^T \mathrm{d}t + \xi_t\xi_t^T \mathrm{d}t\mathrm{d}t \,. \end{aligned} \tag{31}$$

By averaging (31), retaining terms up to first order in $\mathrm{d}t$, dividing by $\mathrm{d}t$ and letting $\mathrm{d}t \to 0$, we get:

$$\begin{aligned} \mathbb{E}[Q_{t+\mathrm{d}t}\, Q_{t+\mathrm{d}t}{}^T] &= \mathbb{E}[Q_t Q_t^T] + \mathbb{E}[Q_t\xi_t^T]\mathrm{d}t + \mathbb{E}[\xi_t Q_t^T]\mathrm{d}t \,, \\ \mathbb{E}[Q_{t+\mathrm{d}t}\, Q_{t+\mathrm{d}t}{}^T] - \mathbb{E}[Q_t Q_t^T] &= \mathbb{E}[Q_t\xi_t^T]\mathrm{d}t + \mathbb{E}[\xi_t Q_t^T]\mathrm{d}t \,, \\ \frac{\mathrm{Var}[Q_t]}{\mathrm{d}t} &= \mathbb{E}[Q_t\xi_t^T] + \mathbb{E}[\xi_t Q_t^T] \,. \end{aligned} \tag{32}$$

A.2 Useful Gaussian identities

Let x and y be jointly Gaussian random vectors:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}\right) \tag{33}$$

Then, the marginal and conditional distributions of x (equivalently for y) are, respectively (Bishop 2007):

$$x \sim N(\mu_x, A) \tag{34}$$

$$x|y \sim N(\mu_x + CB^{-1}(y - \mu_y), A - CB^{-1}C^T) \tag{35}$$

A.3 Update equations of a discrete Kalman Filter

Using the Gaussian Identities in A.2 we have

$$\begin{bmatrix} X_i \\ y_i \end{bmatrix} |y_{1:(i-1)} \sim N\left(\begin{bmatrix} m_i \\ Pm_i \end{bmatrix}, \begin{bmatrix} S_i & S_i P^T \\ PS_i & PS_i P^T + R \end{bmatrix}\right) \tag{36}$$

Since we are working with Gaussians, we know that $X_i|y_{1:i} \sim N(m_i^*, S_i^*)$, and the updated m_i^* and S_i^* are given by:

$$\begin{aligned} m_i^* &= m_i + S_i P^T (PS_i P^T + R)^{-1}(y_i - Pm_i) \,, \\ S_i^* &= S_i + S_i P^T (PS_i P^T + R)^{-1} PS_i \,. \end{aligned} \tag{37}$$

A.4 Analytical solutions for the OU process and its integral

Given an OU process of the following form:

$$\mathrm{d}X_t = -\alpha X_t \mathrm{d}t + \sigma \mathrm{d}W_t \tag{38}$$

we can derive its solution according to the general theory for linear SDEs. Since the solution is a Gaussian process, we will only need to define its mean and variance which are given by Eqs. (6, 7). All the ODEs in this case are first-order linear ODEs with constant coefficients, so using for example an integrating factor, we can derive the following solution for an ODE of the form $\frac{\mathrm{d}x}{\mathrm{d}t} + ax = g(t), \quad x(t=0) = x_0$:

$$x_t = e^{-a(t-t_0)}x_0 + \int_{t_0}^{t} e^{-a(t-\tau)} g(\tau) d\tau. \tag{39}$$

For the mean we get from Eq. (6):

$$\begin{aligned} &\frac{\mathrm{d}m_t}{\mathrm{d}t} = -\alpha m_t, \quad m_{t_0} = m_0 \Rightarrow \\ &m_t = m_0 e^{-a(t-t_0)} \end{aligned} \tag{40}$$

For the variance we have the following:

$$\begin{aligned} &\frac{\mathrm{d}V_t}{\mathrm{d}t} = -2\alpha V_t + \sigma^2, \quad V_{t_0} = V_0 \Rightarrow \\ &V_t = e^{-2\alpha(t-t_0)}V_0 + \int_{t_0}^{t} e^{-2\alpha(t-\tau)}\sigma^2 d\tau \Rightarrow \\ &V_t = e^{-2\alpha(t-t_0)}V_0 + \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha(t-t_0)}) \end{aligned} \tag{41}$$

For the solution of the integrated OU process $\mathrm{d}Y_t/\mathrm{d}t = X_t$, we need to calculate its mean, covariance and variance given by Eqs. (12), (13) and (14). The initial conditions for these ODEs will be set to 0, since at each observation point the integrated process starts from 0. For clarity, we will use the results A, B, C from "Appendix A.5".

First we find the mean:

$$\begin{aligned} &\frac{d\mathbb{E}_t}{\mathrm{d}t} = m_t = m_0 e^{-\alpha(t-t_0)}, \quad \mathbb{E}(t_0) = 0 \Rightarrow \\ &\mathbb{E}_t = \int_{t_0}^{t} m_0 e^{-\alpha(\tau-t_0)} d\tau \overset{A}{\Rightarrow} \\ &\mathbb{E}_t = \frac{m_0}{\alpha}(1 - e^{-\alpha(t-t_0)}) \end{aligned} \tag{42}$$

Table 5 Average of mean posterior ± 1 s.d. over 10 different datasets for α and σ using a Metropolis–Hastings algorithm

Δ	KF	α	σ
0.1	KF1	3.081 ± 0.258	1.670 ± 0.209
0.5	KF1	1.956 ± 0.153	1.199 ± 0.125
1.0	KF1	1.493 ± 0.112	0.799 ± 0.088
2.0	KF1	1.064 ± 0.090	0.485 ± 0.046
0.1	KF2	4.171 ± 0.417	1.974 ± 0.208
0.5	KF2	4.121 ± 0.377	2.068 ± 0.257
1.0	KF2	4.123 ± 0.445	2.083 ± 0.283
2.0	KF2	4.208 ± 0.783	2.091 ± 0.371

Data were simulated from an OU process with $\alpha = 4$ and $\sigma = 2$

For the covariance, we first calculate from Eq. (13):

$$\begin{aligned}
&\frac{d\mathbb{E}[X_tY_t]}{\mathrm{d}t} = -\alpha\mathbb{E}[X_tY_t] + \mathbb{E}[{X_t}^2], \quad \mathbb{E}[X_0Y_0] = 0 \Rightarrow \\
&\mathbb{E}[X_tY_t] = \int_{t_0}^{t} \mathbb{E}[{X_t}^2] e^{-\alpha(t-\tau)} d\tau \overset{\mathbb{E}[{X_t}^2]=V_t+{m_t}^2}{\Rightarrow} \\
&\mathbb{E}[X_tY_t] = \int_{t_0}^{t} \left(\left(m_0^2 - \frac{\sigma^2}{2\alpha} + V_0 \right) e^{-2\alpha(\tau-t_0)} \right. \\
&\qquad \left. + \frac{\sigma^2}{2\alpha} \right) e^{-\alpha(t-\tau)} d\tau \overset{A,C}{\Rightarrow} \\
&\mathbb{E}[X_tY_t] = \frac{\sigma^2}{2\alpha^2}(1 - e^{-\alpha(t-t_0)}) + \\
&\frac{1}{\alpha}\left(m_0^2 - \frac{\sigma^2}{2\alpha} + V_0 \right)(e^{-\alpha(t-t_0)} - e^{-2\alpha(t-t_0)})
\end{aligned} \tag{43}$$

Now the covariance can be calculated from:

$$\begin{aligned}
&\mathrm{Cov}(X_t, Y_t) = \mathbb{E}[X_tY_t] - m_t\mathbb{E}_t \Rightarrow \\
&\mathrm{Cov}(X_t, Y_t) = \\
&= \frac{\sigma^2}{2\alpha^2} + \left(-\frac{\sigma^2}{\alpha^2} + \frac{V_0}{\alpha} \right) e^{-\alpha(t-t_0)} + \left(\frac{\sigma^2}{2\alpha^2} - \frac{V_0}{\alpha} \right) e^{-2\alpha(t-t_0)}
\end{aligned} \tag{44}$$

For the variance, we need to calculate:

$$\begin{aligned}
&\frac{d\mathbb{E}[{Y_t}^2]}{\mathrm{d}t} = 2\mathbb{E}[X_tY_t], \quad \mathbb{E}[{Y_0}^2] = 0 \Rightarrow \\
&\mathbb{E}[{Y_t}^2] = 2\int_{t_0}^{t} \mathbb{E}[X_\tau Y_\tau] d\tau \overset{(43),B}{\Rightarrow} \\
&\mathbb{E}[{Y_t}^2] = \frac{{m_0}^2}{\alpha^2}(1 - 2e^{-\alpha(t-t_0)} + e^{-2\alpha(t-t_0)})
\end{aligned} \tag{45}$$

Now we can derive the variance:

$$\begin{aligned}
&\mathrm{Var}[y_t] = \mathbb{E}[{Y_t}^2] - {\mathbb{E}_t}^2 \Rightarrow \\
&\mathrm{Var}[y_t] = \frac{\sigma^2}{\alpha^2}(t - t_0) + \left(\frac{\sigma^2}{2\alpha^3} - \frac{V_0}{\alpha^2}\right)(1 - e^{-2\alpha(t-t_0)}) \\
&\qquad + 2\left(-\frac{\sigma^2}{\alpha^3} + \frac{V_0}{\alpha^2}\right)(1 - e^{-\alpha(t-t_0)})
\end{aligned} \tag{46}$$

A.5 Frequently used integrals for part (A.4)

$$A = \int_{t_0}^{t} e^{-\alpha(\tau-t_0)} d\tau = \frac{1}{\alpha}(1 - e^{-\alpha(t-t_0)}) \tag{47}$$

$$B = \int_{t_0}^{t} e^{-2\alpha(\tau-t_0)} d\tau = \frac{1}{2\alpha}(1 - e^{-2\alpha(t-t_0)}) \tag{48}$$

$$\begin{aligned}
C &= \int_{t_0}^{t} e^{-\alpha(t-\tau)} e^{-2\alpha(\tau-t_0)} d\tau \\
&= \frac{1}{\alpha}(e^{-\alpha(t-t_0)} - e^{-2\alpha(t-t_0)})
\end{aligned} \tag{49}$$

A.6 Exact updating formula of OU process

The OU process $\mathrm{d}X_t = -\alpha X_t \mathrm{d}t + \sigma \mathrm{d}W_t$ admits an exact update formula given by Gillespie (1992):

$$X_{t+\mathrm{d}t} = X_t e^{-\alpha \mathrm{d}t} + \sqrt{\sigma^2 \frac{1}{2\alpha} e^{-2\alpha dt}} N(0, 1), \tag{50}$$

A.7 Average over 10 datasets—OU example

See Table 5.

A.8 LNA for Lotka–Volterra model

The Lotka–Volterra model (23) gives rise to the stoichiometry matrix,

$$S = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \tag{51}$$

with transition rates,

$$h(X) = \begin{bmatrix} \theta_1 X_1 \\ \theta_2 X_1 X_2 \\ \theta_3 X_2 \end{bmatrix}. \tag{52}$$

Table 6 Nelder–Mead results for $\theta_1, \theta_2, \theta_3$. The median values across 100 datasets are shown in the third and fourth column for KF1 and KF2, respectively

θ	Ground truth	KF1 Median[LQ,UQ]	KF2 Median[LQ,UQ]
θ_1	0.5	0.48160 [0.47770,0.48651]	0.49746 [0.49278,0.50122]
θ_2	0.0025	0.00227 [0.00222,0.00232]	0.00248[0.00244,0.00254]
θ_3	0.3	0.24773 [0.23927,0.25797]	0.30047[0.29320,0.31061]

Lower and upper quartiles are shown in brackets

The following matrices need to be computed:

$$F = \begin{bmatrix} \theta_1 & 0 \\ \theta_2 y_2 & \theta_2 y_1 \\ 0 & \theta_3 \end{bmatrix}, \tag{53}$$

$$SF^T = A = \begin{bmatrix} \theta_1 - \theta_2 y_2 & -\theta_2 y_1 \\ \theta_2 y_2 & \theta_2 y_1 - \theta_3 \end{bmatrix}, \tag{54}$$

$$Sdiag(h(y_t))S^T = EE^T = \begin{bmatrix} \theta_1 y_1 + \theta_2 y_1 y_2 & -\theta_2 y_1 y_2 \\ -\theta_2 y_1 y_2 & \theta_2 y_1 y_2 + \theta_3 y_2 \end{bmatrix}, \tag{55}$$

The macroscopic rate equations are now given by:

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = \theta_1 y_1 - \theta_2 y_1 y_2 \tag{56}$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}t} = \theta_2 y_1 y_2 - \theta_3 y_2 \tag{57}$$

For the diffusion terms, we only need to compute the variance of the resulting Gaussian process since we restart the stochastic part at each observation point in accordance with (Fearnhead et al. 2014).

$$\begin{aligned} \frac{\mathrm{d}V}{\mathrm{d}t} = VA^T + EE^T + AV = \\ \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{bmatrix} \theta_1 - \theta_2 y_2 & \theta_2 y_2 \\ -\theta_2 y_1 & \theta_2 y_1 - \theta_3 \end{bmatrix} + \\ + \begin{bmatrix} \theta_1 y_1 + \theta_2 y_1 y_2 & -\theta_2 y_1 y_2 \\ -\theta_2 y_1 y_2 & \theta_2 y_1 y_2 + \theta_3 y_2 \end{bmatrix} + \\ + \begin{bmatrix} \theta_1 - \theta_2 y_2 & -\theta_2 y_1 \\ \theta_2 y_2 & \theta_2 y_1 - \theta_3 \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \end{aligned} \tag{58}$$

V is a symmetric matrix so $V_{12} = V_{21}$. So:

$$\frac{\mathrm{d}V_{11}}{\mathrm{d}t} = 2V_{11}(\theta_1 - \theta_2 y_2) - 2V_{12}\theta_2 y_1 + \theta_2 y_1 y_2 + \theta_1 y_1 \tag{59}$$

$$\frac{\mathrm{d}V_{12}}{\mathrm{d}t} = V_{12}(\theta_2 y_1 - \theta_3 + \theta_1 - \theta_2 y_2) + V_{11}\theta_2 y_2 - \theta_2 y_1 V_{22} - \theta_2 y_1 y_2 \tag{60}$$

$$\frac{\mathrm{d}V_{22}}{\mathrm{d}t} = 2V_{22}(\theta_2 y_1 - \theta_3) + 2V_{12}\theta_2 y_2 + \theta_2 y_1 y_2 + \theta_3 y_2 \tag{61}$$

The integrated process $\mathrm{d}Y_t = X_t \mathrm{d}t$ follows Eqs. (10),(13), (14). The deterministic part is given by:

$$\frac{\mathrm{d}I_1}{\mathrm{d}t} = y_1 \,, \quad \frac{\mathrm{d}I_2}{\mathrm{d}t} = y_2 \,. \tag{62}$$

The ODEs for its integrated variance and covariance with the underline process X_t are given below, where $\mathrm{Cov}(YX^T) = C_t = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$ and $\mathrm{Var}(Y) = G_t$:

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} \theta_1 - \theta_2 y_2 & \theta_2 y_2 \\ -\theta_2 y_1 & \theta_2 y_1 - \theta_3 \end{bmatrix} + \begin{bmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{bmatrix}, \tag{63}$$

such as,

$$\frac{\mathrm{d}C_{11}}{\mathrm{d}t} = (\theta_1 - \theta_2 y_2)C_{11} - \theta_2 y_1 C_{12} + V_{11} \tag{64}$$

$$\frac{\mathrm{d}C_{12}}{\mathrm{d}t} = \theta_2 y_2 C_{11} + (\theta_2 y_1 - \theta_3)C_{12} + V_{12} \tag{65}$$

$$\frac{\mathrm{d}C_{21}}{\mathrm{d}t} = (\theta_1 - \theta_2 y_1)C_{21} - \theta_2 y_1 C_{22} + V_{12} \tag{66}$$

$$\frac{\mathrm{d}C_{22}}{dt} = \theta_2 y_2 C_{21} + (\theta_2 y_1 - \theta_3)C_{22} + V_{22} \tag{67}$$

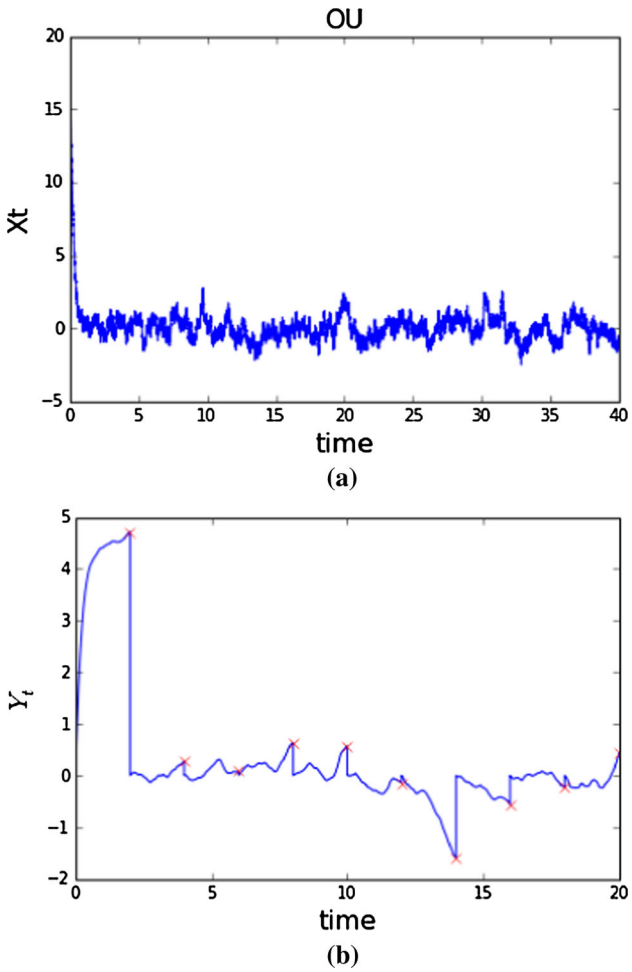
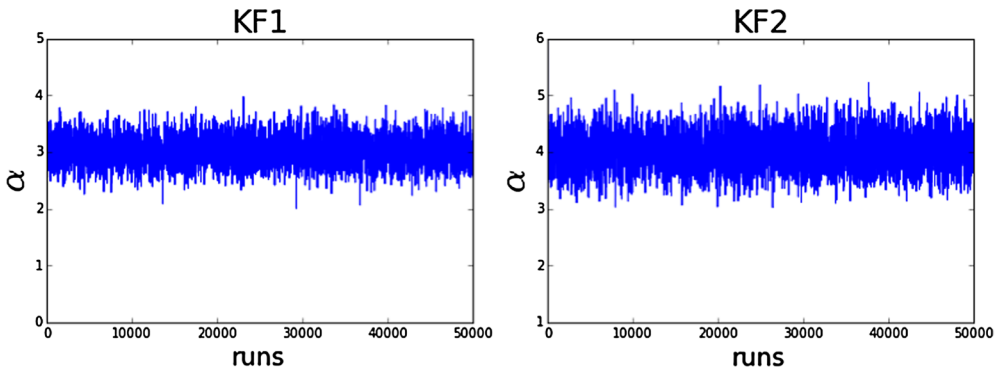
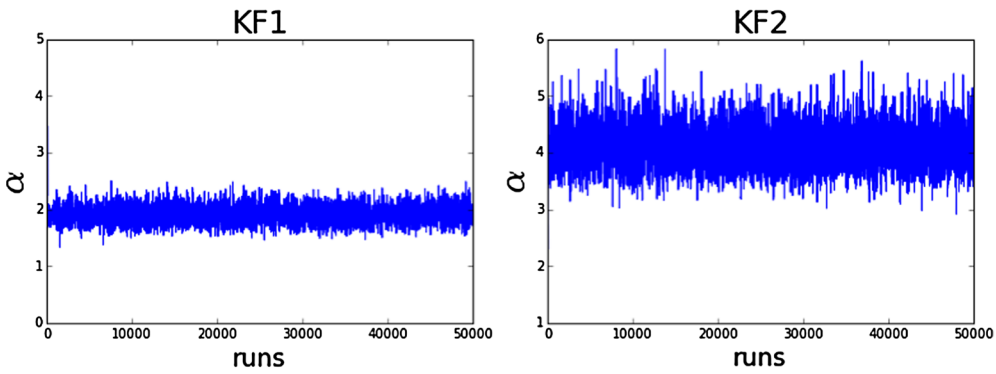


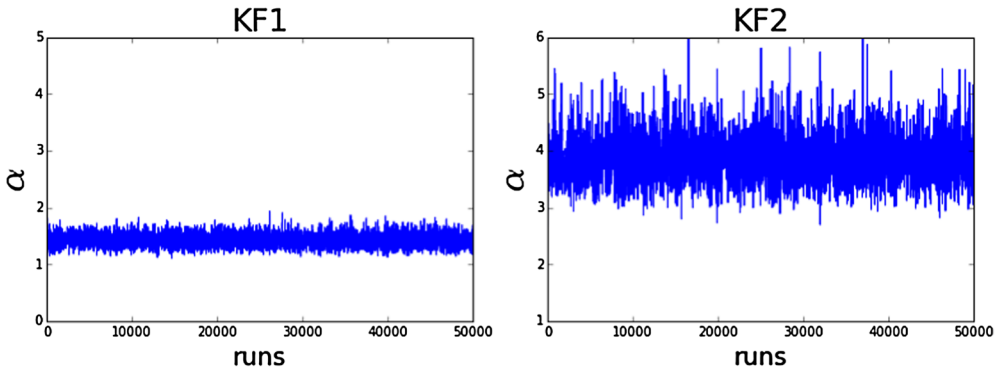
Fig. 5 Simulated trajectories from an OU process with $\alpha = 4$ and $\sigma = 2$, along with the corresponding aggregated process with an integration period of 2 min. For the aggregated process, we assumed observations every 2 min, which are indicated by red crosses. **a** OU process. **b** Aggregated process



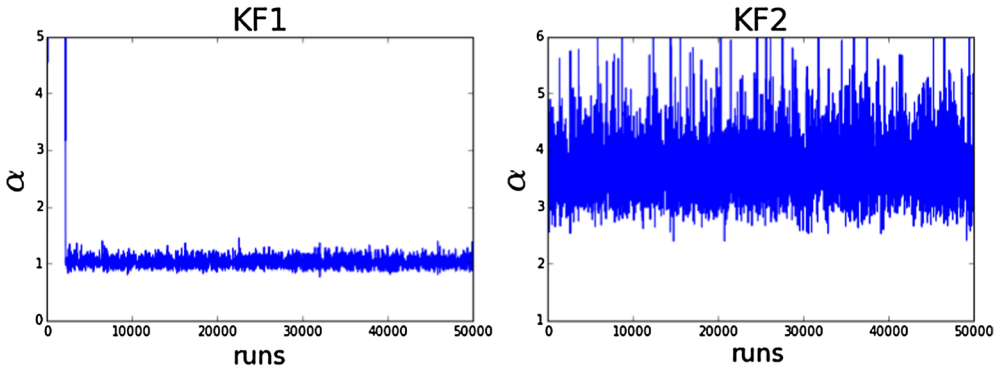
KF1, $\Delta = 0.1$. KF2, $\Delta = 0.1$.



KF1, $\Delta = 0.5$. KF2, $\Delta = 0.5$.

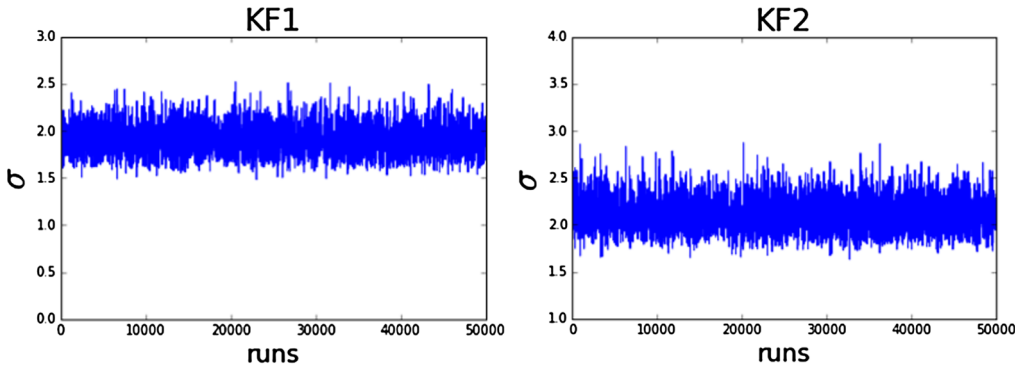


KF1, $\Delta = 1.0$. KF2, $\Delta = 1.0$.

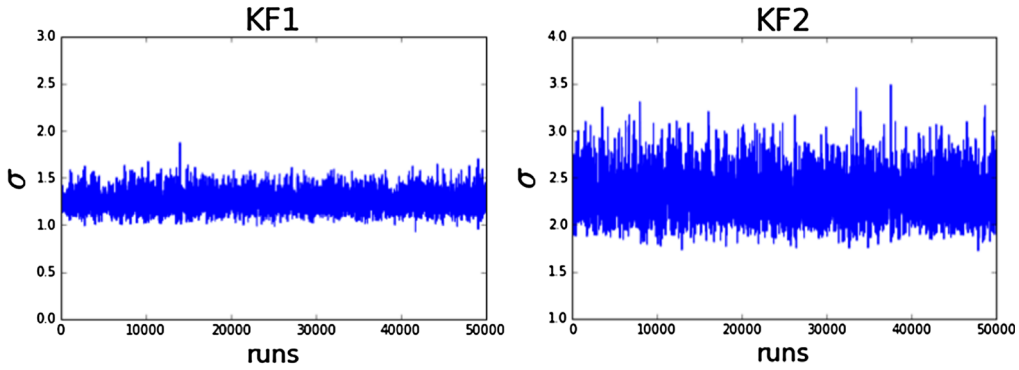


KF1, $\Delta = 2.0$. KF2, $\Delta = 2.0$.

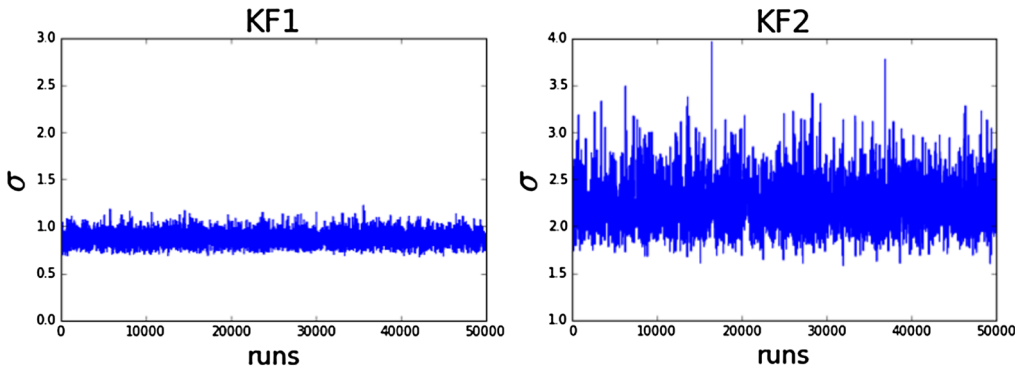
Fig. 6 MCMC traces of the posterior of α using a random walk MH for both KF1 and KF2. Ground truth for $\alpha = 4$



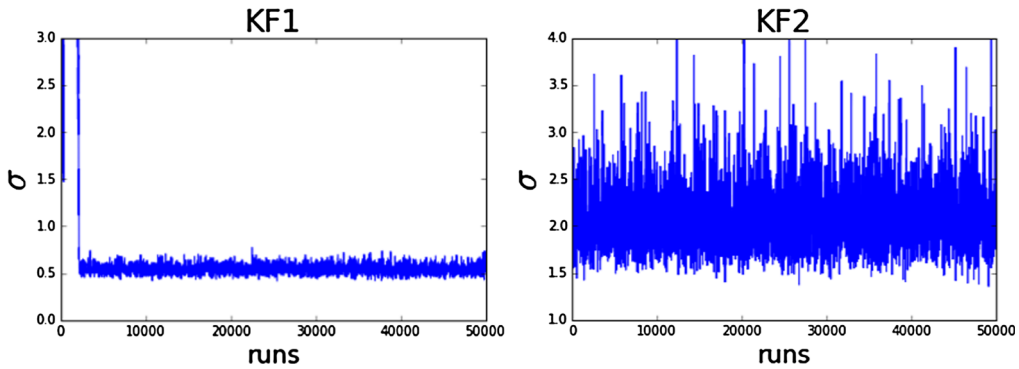
KF1, $\Delta = 0.1$. KF2, $\Delta = 0.1$.



KF1, $\Delta = 0.5$. KF2, $\Delta = 0.5$.



KF1, $\Delta = 1.0$. KF2, $\Delta = 1.0$.



KF1, $\Delta = 2.0$. KF2, $\Delta = 2.0$.

Fig. 7 MCMC traces of the posterior of σ using a random walk MH for both KF1 and KF2. Ground truth for $\sigma = 2$

$$\frac{\mathrm{d}G}{\mathrm{d}t} = C_t + {C_t}^T \tag{68}$$

A.9 Adaptive MCMC

According to the specific adaptive MH (Sherlock et al. 2010), the new state θ^* is sampled from a mixture of Gaussians:

$$\theta^* = \begin{cases} N(\theta_t, \Sigma_0), & \text{w.p. } \delta \\ N(\theta_t, \lambda\Sigma_t), & \text{w.p. } 1-\delta \end{cases} \tag{69}$$

Σ_t corresponds to the sampled variance up to iteration t and is estimated after enough samples have been accepted. The parameter $\delta \in (0, 1)$ and is defined by the user, we have chosen a value of 0.05. The scaling factor λ can either be fixed (Roberts and Rosenthal 2009) or be tuned (Sherlock et al. 2010). This algorithm targets an acceptance rate of ≈ 0.3.

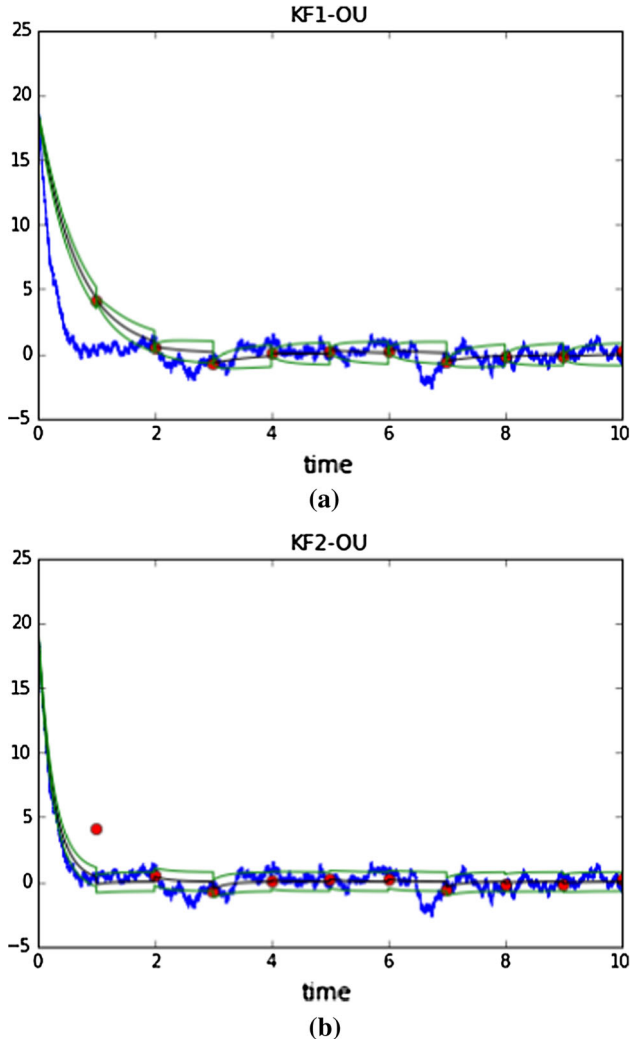


Fig. 8 Filtering results from KF1 and KF2 for an OU process with $\alpha = 4.0$ and $\sigma = 2.0$ (blue trace) using aggregated data over an integration period of $\Delta = 1.0$. Black lines correspond to the posterior mean estimate and green lines to 1 s.d.. For inference, we used the estimated parameters from A.7. **a** KF1 ($\Delta = 1.0$). **b** KF2 ($\Delta = 1.0$). (Color figure online)

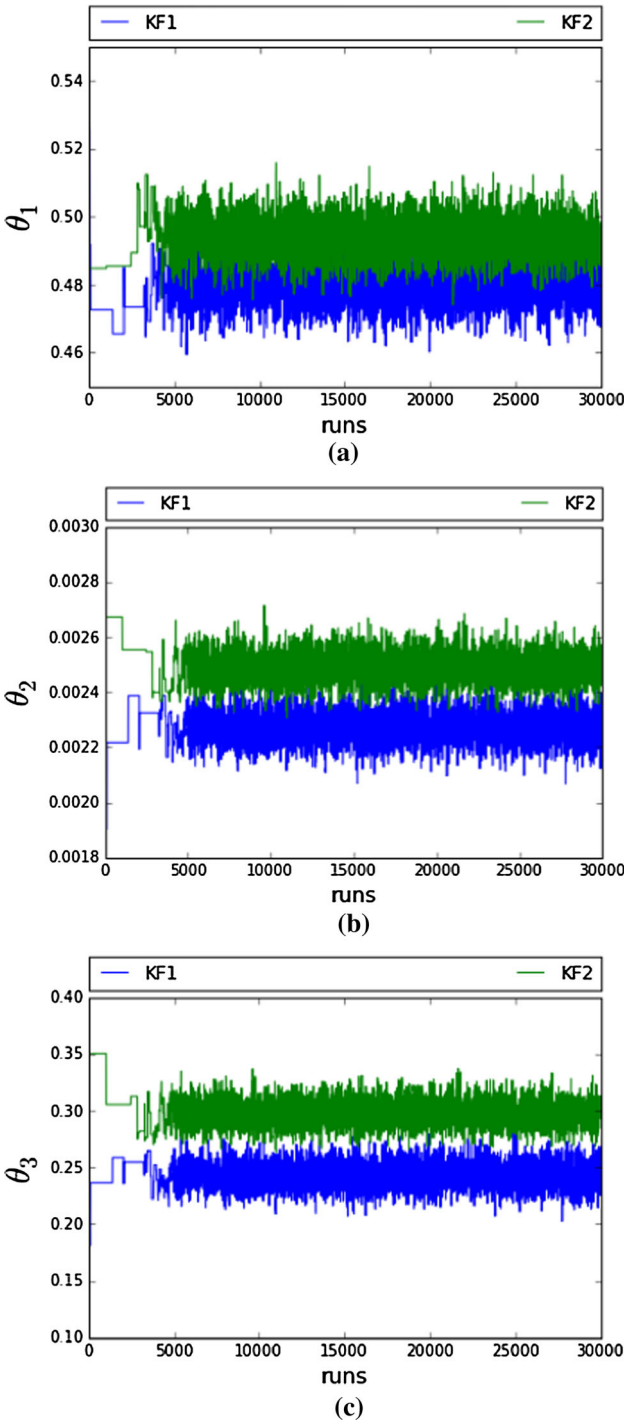


Fig. 9 MCMC traceplots for the LV experiment using an adaptive MCMC algorithm. **a** MCMC traceplots for θ_1. **b** MCMC traceplots for θ_2. **c** MCMC traceplots for θ_3

A.10 Nelder Mead results for LV model

See Table 6.

A.11 LNA for translation inhibition model

The following model is being assumed where R and P stand for the (numbers of) gene mRNA and protein, respectively:

$$R \xrightarrow{c_P} R + P \text{ (translation)} \tag{70a}$$

$$P \xrightarrow{d_P P/\Omega} \oslash \text{ (protein degradation)} \tag{70b}$$

The above equations result in the following stoichiometry matrix:

$$S = \begin{bmatrix} 1 & -1 \end{bmatrix}, \tag{71}$$

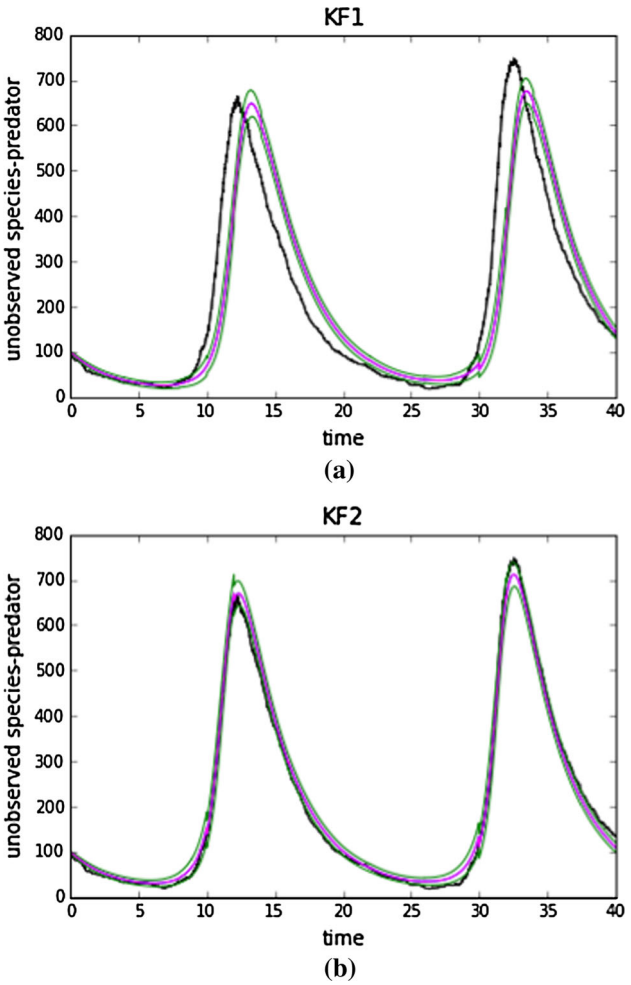


Fig. 10 Filtering results for the predator population. Red dots correspond to aggregated observations, and the black line represents the actual process. Purple lines represent the mean estimate and green 1 s.d. . **a** Filtering results for the predator population using KF1. **b** Filtering results for the predator population using KF2

and the transition rates are :

$$h(x,t) = \begin{bmatrix} c_P \\ d_P p \end{bmatrix}, \tag{72}$$

The required matrices are calculated below:

$$F = \begin{bmatrix} 0 & d_P \end{bmatrix}, \tag{73}$$

$$SF^T = A = \begin{bmatrix} -d_P \end{bmatrix}, \tag{74}$$

$$\text{Sdiag}(h(y_t,\theta))S^T = EE^T = \begin{bmatrix} c_P + d_P p \end{bmatrix}, \tag{75}$$

The deterministic part is now given by:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = c_P - d_P p \tag{76}$$

The stochastic part is given by the (restarting) LNA where we have dropped the dependency of M_t, V_t from time:

$$\mathrm{d}M_p = -d_p M_p \mathrm{d}t + \sqrt{c_P + d_P p}\mathrm{d}W_t \tag{77}$$

resulting in the following ODE for the stochastic variance:

$$\frac{\mathrm{d}V_p}{\mathrm{d}t} = -2d_p V_p + c_P + d_P p \tag{78}$$

For the integrated process, we get the following according to Eqs. (10),(13) and (14) (Fig. 5). First the deterministic part is given by:

$$\frac{\mathrm{d}I_p}{\mathrm{d}t} = p, \tag{79}$$

The stochastic part is given by:

$$\frac{\mathrm{d}Q_p}{\mathrm{d}t} = M_p, \tag{80}$$

resulting in the following ODEs for its integrated variance and covariance with the unintegrated process:

$$\frac{dCov(Q_p M_p^T)}{\mathrm{d}t} = -d_p Cov(Q_p M_p^T) + V_p \tag{81}$$

$$\frac{d\mathrm{Var}(Q_p)}{\mathrm{d}t} = 2Cov(Q_p M_p^T) \tag{82}$$

A.12 OU and aggregated OU process

See Fig. 5.

A.13 OU traceplots

Figures 6 and 7 show samples of the OU parameters during the MCMC runs.

A.14 Filtering results for OU process using aggregated data

See Fig. 8.

A.15 MCMC traces from LV experiment

See Fig. 9.

A.16 Filtering results for the predator population

See Fig. 10.

A.17 MCMC traces for Translation inhibition example with synthetic data

See Fig. 11.

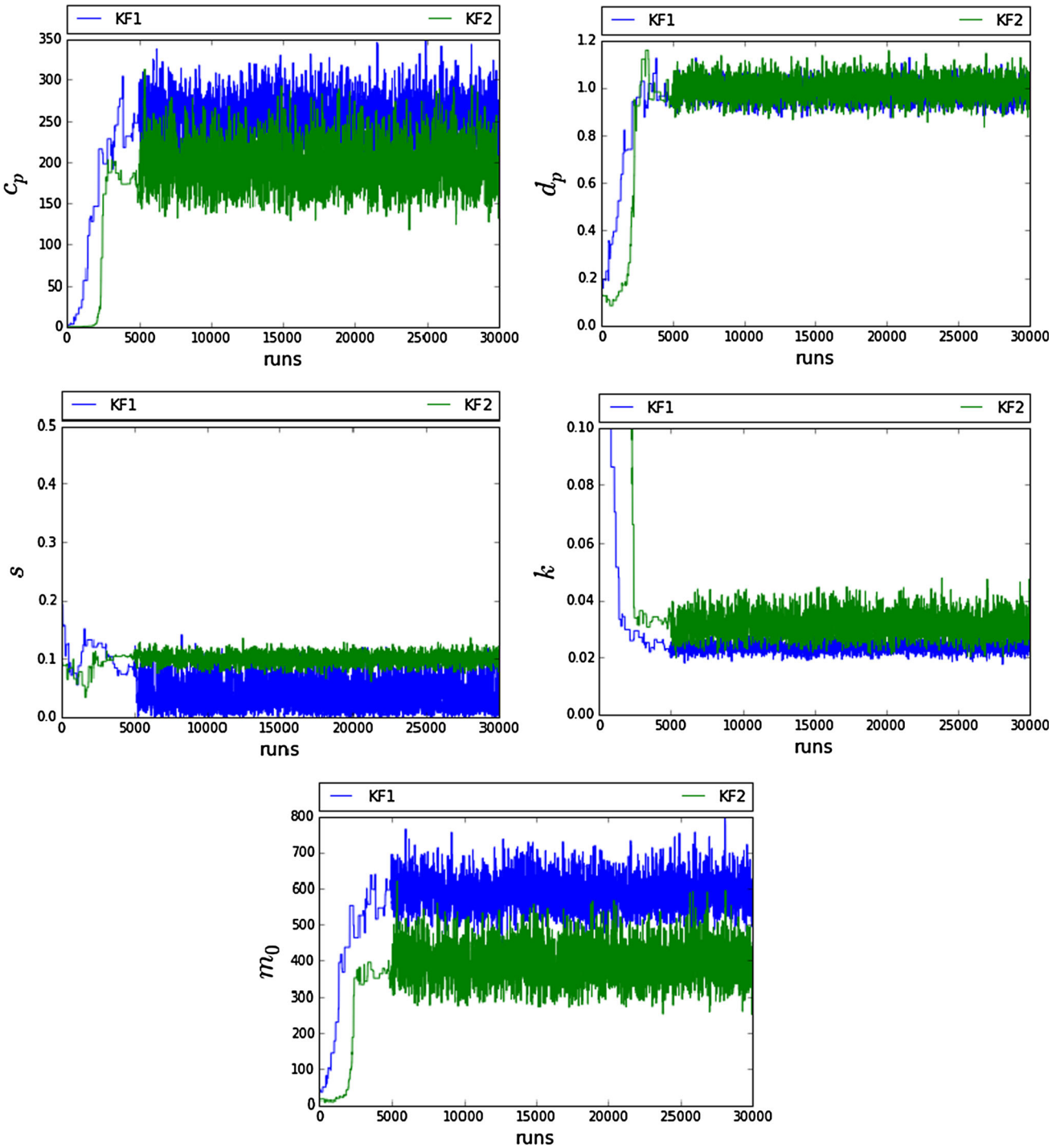


Fig. 11 Adaptive MCMC traces of the posterior vector (c_p, d_p, s, k, m_0) using synthetic data with KF1 and KF2

A.18 MCMC traces for translation inhibition example with single cell data

See Figs. 12 and 13

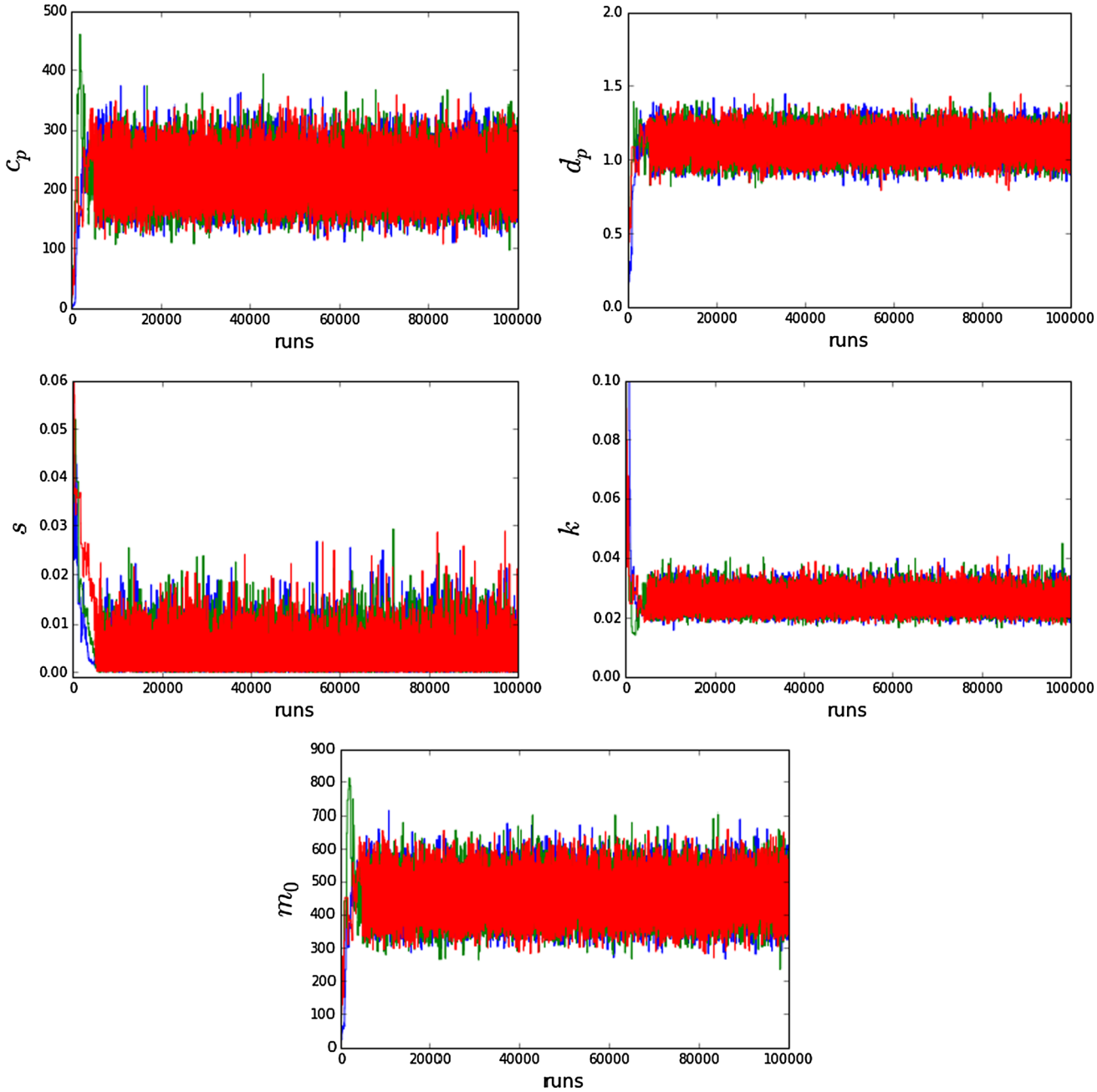


Fig. 12 Three MCMC chains of the posterior vector (c_p, d_p, s, k, m_0) using single cell data with KF1

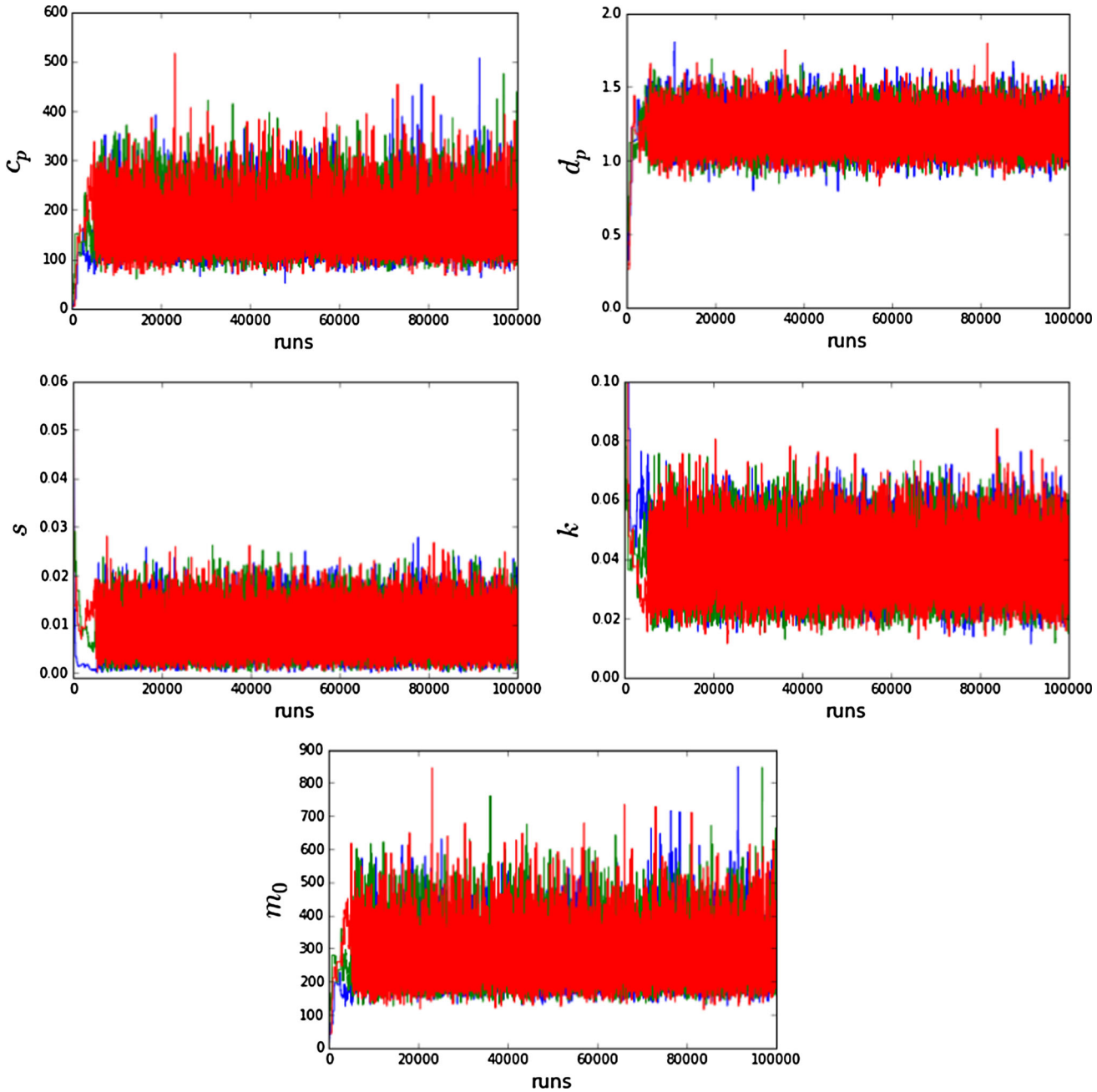


Fig. 13 Three MCMC chains of the posterior vector (c_p, d_p, s, k, m_0) using single cell data with KF2

References

Arnold, L.: Stochastic Differential Equations Theory and Applications. [S.l.]. Wiley, Hoboken (1974)

Bishop, C.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, New York (2007)

Boys, R., Wilkinson, D., Kirkwood, T.: Bayesian inference for a discretely observed stochastic kinetic model. Stat. Comput. **18**(2), 125–135 (2008)

Elf, J., Ehrenberg, M.: Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. Genome Res. **13**(11), 2475–2484 (2003)

Fearnhead, P., Giagos, V., Sherlock, C.: Inference for reaction networks using the linear noise approximation. Biometrics **70**(2), 457–466 (2014)

Finkenstädt, B., Woodcock, D.J., Komorowski, M., Harper, C.V., Davis, J.R.E., White, M.R.H., Rand, D.A.: Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: an application to single cell data. Ann. Appl. Stat. **7**(4), 1960–1982 (2013)

Gardiner, C.: Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences (Springer Series in Synergetics), 3rd edn. Springer, Berlin (2004)

Giagos, V.: Inference for auto-regulatory genetic networks using diffusion process approximations. Ph.D. Thesis, Lancaster University (2010)

Gillespie, D.T.: Markov Processes : An Introduction for Physical Scientists. Academic Press, Boston (1992)

Gillespie, D.T.: A rigorous derivation of the chemical master equation. Phys. A Stat. Mech. Appl. **188**, 404–425 (1992)

Gillespie, D.T.: Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral. Phys. Rev. E **54**, 2084–2091 (1996). doi:10.1103/physreve.54.2084

Golightly, A., Wilkinson, D.J.: Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics **61**(3), 781–788 (2005)

Harper, C.V., Finkenstädt, B., Woodcock, D.J., Friedrichsen, S., Semprini, S., Ashall, L., Spiller, D.G., Mullins, J.J., Rand, D.A., Davis, J.R.E., White, M.R.H.: Dynamic analysis of stochastic transcription cycles. PLoS Biol **9**(4), e1000607 (2011)

Hull, J.: Futures and Other Derivatives. Options, Options, Futures and Other Derivatives. Pearson/Prentice Hall, Upper Saddle River (2009)

Jazwinski, A.H.: Stochastic Processes and Filtering Theory. Academic Press, Cambridge (1970)

Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Basic Eng. **82**(1), 35–45 (1960)

Komorowski, M., Finkenstädt, B., Harper, C.V., Rand, D.A.: Bayesian inference of biochemical kinetic parameters using the linear noise approximation. BMC Bioinform. **10**(1), 1–10 (2009)

Mbalawata, I.S., Särkkä, S., Haario, H.: Parameter estimation in stochastic differential equations with Markov chain Monte Carlo and non-linear Kalman filtering. Comput. Stat. **28**(3), 1195–1223 (2013)

Roberts, G.O., Rosenthal, J.S.: Examples of adaptive MCMC. J. Comput. Graph. Stat. **18**(2), 349–367 (2009)

Särkkä, S.: Recursive Bayesian inference on stochastic differential equations. Ph.D. Thesis, Helsinki University of Technology (2006)

Sherlock, C., Fearnhead, P., Roberts, G.O.: The random walk metropolis: linking theory and practice through a case study. Stat. Sci. **25**(2), 172–190 (2010)

Spiller, D.G., Wood, C.D., Rand, D.A., White, M.R.H.: Measurement of single-cell dynamics. Nature **465**(7299), 736–745 (2010)

van Kampen, N.: Stochastic Processes in Physics and Chemistry, 3rd edn. Elsevier, Amsterdam (2007)

Wilkinson, D.J.: Stochastic modelling for systems biology. In: Chapman & Hall/CRC Mathematical and Computational Biology, 2nd edn. Taylor & Francis (2011)