Markov Chain Monte Carlo for Continuous-Time Switching Dynamical Systems

Lukas Köhs [1] **Bastian Alt** [1] **Heinz Koeppl** [1]

Abstract

Switching dynamical systems are an expressive model class for the analysis of time-series data. As in many fields within the natural and engineering sciences, the systems under study typically evolve continuously in time, it is natural to consider continuous-time model formulations consisting of switching stochastic differential equations governed by an underlying Markov jump process. Inference in these types of models is however notoriously difficult, and tractable computational schemes are rare. In this work, we propose a novel inference algorithm utilizing a Markov Chain Monte Carlo approach. The presented Gibbs sampler allows to efficiently obtain samples from the exact continuous-time posterior processes. Our framework naturally enables Bayesian parameter estimation, and we also include an estimate for the diffusion covariance, which is oftentimes assumed fixed in stochastic differential equations models. We evaluate our framework under the modeling assumption and compare it against an existing variational inference approach.

1. Introduction

A wide range of natural and engineered systems are naturally modeled as continuous-time stochastic processes. The model state space depends on the system at hand; while modeling approaches often focus on either a discrete or continuous state description, a great variety of systems involve both discrete and continuous components, where a discrete, switching process influences the dynamics of some continuous quantity. In the biological sciences in particular, one is often dealing with systems with hybrid state space structure: In neuroscience, for instance, the brain is commonly assumed to adopt different states, e.g., depending on the environment or own actions, such as *eyes opened* versus *eyes closed*, eliciting qualitatively different electrophysiological dynamics (Weng et al., 2020). Likewise, in cellular biology, the state of genetic toggle switches drives continuous measurable quantities such as a protein concentrations (Tian & Burrage, 2006), and the stochastic conformational gating of ion channels (Bressloff, 2020) determines the ion current passing through (Anderson et al., 2015). Further examples include engineering applications, such as the safety of air traffic under potential system failures (Lygeros & Prandini, 2010); the control of the power distribution in an electrical grid in different connectivity modes (Střelec et al., 2012); and analyses of exchange rates or stock returns depending on market states in econometrics (Azzouzi & Nabney, 1999).

A versatile framework to analyze systems of this kind are stochastic hybrid systems (SHS), which have a long history in control, statistics and machine learning (Davis, 1984; Hu et al., 2000; Engell et al., 2003; Davis, 2018; Cassandras & Lygeros, 2018). SHS can be defined in various ways; In this paper, we focus on fully stochastic, continuous-time systems described as switching stochastic differential equations (SSDEs), in which a discrete Markov jump process (MJP) drives a subordinate stochastic differential equation (SDE) (Mao & Yuan, 2006).

The problem of inference has been treated extensively in both SDEs and MJPs. Classical exact results include the well-known Kalman and Wonham filters (Bain & Crisan, 2008), as well as respective smoothing extensions, such as the Rauch-Tung-Striebel (RTS) smoother (Särkkä, 2013; Van Handel, 2007). Also, a great variety of approximate solutions have been worked out, building both on sampling schemes (Doucet et al., 2000; Rao & Teh, 2013) and variational inference approaches (Opper & Sanguinetti, 2007; Wildner & Koeppl, 2021).

Inference frameworks for SSDEs are scarce, however. In switching linear dynamical system (SLDS), a discrete-time analog of SSDEs, inference methods exist utilizing exact sampling (Oh et al., 2005; Fox et al., 2008; Linderman et al., 2017); likewise, due to the computational intractability of the full posteriors (Murphy, 2012), approximate solutions such as the Gaussian sum filter (Alspach & Sorenson,

[1]Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt, Darmstadt, Germany. Correspondence to: Heinz Koeppl <heinz.koeppl@tu-darmstadt.de>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1972), the switching Kalman filter (Böker & Lunze, 2002), or—more recently—variational neural network approaches (Johnson et al., 2016) have been put forward. While a variational approach to SSDE inference has been proposed recently (Köhs et al., 2021), a framework for exact inference is lacking to the best of our knowledge. Here, we present a Gibbs sampling scheme, allowing to efficiently sample from the exact posterior SDE and MJP processes. To this end, we derive the posterior process equations and present a backward-forward/forward-backward (BFFB)-sweeping algorithm to deal with the intricacies of stochastic integration. We then combine this with a Bayesian treatment of the model parameters, for which we obtain full posterior distributions. An implementation of the proposed framework is publicly available.[‡]

2. Model

2.1. Mathematical Background

The switching dynamical systems we consider consist of three joint stochastic processes (i) a continuous-time switching process $Z := \{Z(t)\}_{t\geq 0}$, (ii) a continuous-time subordinated diffusion process $Y := \{Y(t)\}_{t\geq 0}$, and (iii) an observation process $X := \{X_i\}_{i\in\mathbb{N}}$ at discrete time points $\{t_i\}_{i\in\mathbb{N}}$. A realization of this system is shown in Fig. 1.

The *switching process* Z, at the top of the hierarchy is given as a Markov jump process (MJP) (Norris, 1997) freely evolving in time t. An MJP, with $Z(t) \in \mathcal{Z} \subseteq \mathbb{N}$, is a time-continuous Markov process on a countable state space $\mathcal{Z}$ which is fully characterized by (i) an initial probability distribution $p(z_0) := \mathsf{P}(Z(0) = z_0),\ \forall z_0 \in \mathcal{Z}$, and (ii) the transition rate function defined for $z' \in \mathcal{Z} \setminus z$ as

$$\Lambda(z, z', t) := \lim_{h\searrow 0} h^{-1}\mathsf{P}(Z(t+h) = z' \mid Z(t) = z) \quad (1)$$

and the *exit rate* $\Lambda(z, z, t) := -\sum_{z'\in\mathcal{Z}\setminus z} \Lambda(z, z', t)$.

The *subordinated diffusion process* Y is a continuous-valued process with $Y(t) \in \mathcal{Y} \subseteq \mathbb{R}^n$, depending on the freely evolving MJP Z. This yields a switching stochastic differential equation (SSDE) (Mao & Yuan, 2006) defined in an Itô sense as

$$\mathrm{d}Y(t) = f(Z(t), Y(t), t)\,\mathrm{d}t + Q(Z(t), Y(t), t)\,\mathrm{d}W(t), \quad (2)$$

with drift function $f : \mathcal{Z}\times\mathcal{Y}\times\mathbb{R}_{\geq 0} \to \mathcal{Y}$, the invertible dispersion $Q : \mathcal{Z}\times\mathcal{Y}\times\mathbb{R}_{\geq 0} \to \mathbb{R}^{n\times n}$ determining the noise covariance as $D(z, y, t) := Q(z, y, t)Q^\top(z, y, t)$ with $y \in \mathcal{Y}$, $z \in \mathcal{Z}$. The diffusion is driven by the n-dimensional standard Wiener process W, and the distribution of the initial value is given by the density $p(y_0 \mid z_0)$. The difference to a conventional SDE consists in the $Z(t)$-dependence of f and Q; for an accessible introduction to SDEs, see, e.g., Särkkä & Solin (2019). Given a realization of the MJP, the SSDE in Eq. (2) can hence be equivalently interpreted as a concatenation of individual SDEs determined by the switching process Z.

The MJP Z and the diffusion Y together constitute the latent *hybrid process* $\{Z(t), Y(t)\}_{t\geq 0}$. To avoid ambiguity, we denote the discrete value $Z(t)$ as the *mode* and the continuous value $Y(t)$ as the *state* of the system. Furthermore, we will use upper-case letters $Z(t)$ to refer to random variables and lower-case letters $z(t)$ to refer to respective realizations throughout the paper. For any time interval $[0, T]$, the hybrid process induces a measure P on the space Ω_T of all possible paths $\omega_T := \left(z_{[0,T]}, y_{[0,T]}\right)$, where $y_{[0,T]} := \{y(t)\}_{t\in[0,T]}, z_{[0,T]} := \{z(t)\}_{t\in[0,T]}$ (Çınlar, 2011); that is, for any event $\mathcal{A}$ in the Borel σ-algebra of paths, we can formally find its associated probability by integration,

$$\begin{aligned}\mathsf{P}\left(\left(Z_{[0,T]}, Y_{[0,T]}\right) \in \mathcal{A}\right) &= \int_{\mathcal{A}} \mathsf{P}\left(\left(Z_{[0,T]}, Y_{[0,T]}\right) \in \mathrm{d}\omega\right) \\ &\equiv \int_{\mathcal{A}} \mathrm{d}\mathsf{P}(\omega_T).\end{aligned} \quad (3)$$

Though it is not sensible to define a density for the path-measure in Eq. (3) as there does not exist a Lebesgue measure for an infinite uncountable number of random variables, we note that time-point-wise this quantity admits a probability density function $p(z, y, t)$,

$$\begin{aligned}\mathsf{E}\left[\varphi(Z(t), Y(t), t)\right] &= \int_\Omega \varphi(Z(t), Y(t), t)\mathrm{d}\mathsf{P}(\omega_T) \\ &= \sum_{z\in\mathcal{Z}} \int_{\mathcal{Y}} \varphi(z, y, t)p(z, y, t)\,\mathrm{d}y,\end{aligned} \quad (4)$$

where $\varphi : \mathcal{Y}\times\mathcal{Z}\times\mathbb{R}_{\geq 0} \to \mathbb{R}$ is an arbitrary test function. This density evolves over time according to the hybrid master equation (HME)

$$\partial_t p(y, z, t) = [\mathcal{L}p](y, z, t) \quad (5)$$

with initial distribution $p(y_0, z_0, 0) = p(z_0)p(y_0 \mid z_0)$ and $\mathcal{L} = \mathcal{T} + \mathcal{F}$,

$$\begin{aligned}[\mathcal{T}\varphi](y, z, t) &:= \sum_{z'\in\mathcal{Z}} \Lambda(z', z, t)\varphi(y, z', t) \\ [\mathcal{F}\varphi](y, z, t) &:= -\sum_{i=1}^{n} \partial_{y_i}\left\{f_i(y, z, t)\varphi(y, z, t)\right\} \\ &\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{y_i}\partial_{y_j}\{D_{ij}(y, z, t)\varphi(y, z, t)\}\end{aligned}$$

with φ again an arbitrary test function, see, e.g. (Pawula, 1967; Köhs et al., 2021). Unfortunately, a general, analytical solution to the HME does not exist. Numerical solution

[‡]https://git.rwth-aachen.de/bcs/projects/lk/mcmc-ct-sds.git

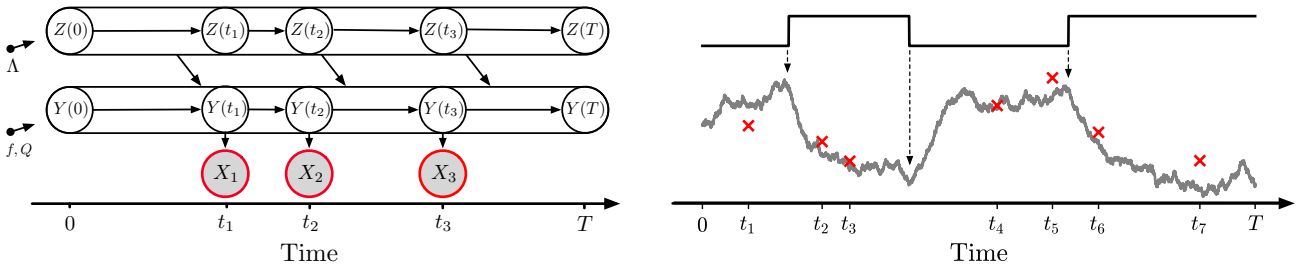

Figure 1. Hybrid process model. Left: adaptation of a probabilistic graphical model. Right: sketch of a realization. A two-state MJP $Z(t)$ (top, see Eq. (1)), governed by rates Λ and freely evolving in the interval $t \in [0, T]$, controls the dynamics of the SSDE $Y(t) \mid Z(t)$ (bottom, gray line), cf. Eq. (2), via drift and diffusion f and Q. From these latent continuous dynamics, only sparse and noisy observations (red crosses) obtained at irregularly-spaced time points $t_1, t_2, \ldots$ are available for inference. Vertical arrows indicate the Z-transitions.

methods such as finite elements may be applied, which however—suffering from the curse of dimensionality—can only be used in very low-dimensional settings, and even in low dimensions are non-trivial to adapt to the given partial differential equation (PDE) (Grossmann et al., 2007). Sampling trajectories from $\{Y(t), Z(t)\}_{t\in[0,T]}$ is straightforward, on the other hand: A realization $z_{[0,T]}$ of the discrete process Z can be simulated by utilizing the Doob-Gillespie algorithm (Doob, 1945). Given this trajectory $z_{[0,T]}$, the diffusion Y can be simulated using, e.g., an Euler-Maruyama or stochastic Runge-Kutta method (Kloeden & Platen, 1992).

Finally, the *observation process* X consists of a countable set of observed data points $\{X_i\}_{i\in\mathbb{N}}$, with $X_i \in \mathcal{X}$, at times $\{t_i\}_{i\in\mathbb{N}}$. The ith data point X_i is generated conditional on the diffusion process Y as $X_i \sim p(x_i \mid Y(t_i) = y_i)$. In general, the observation space $\mathcal{X}$ can be either discrete, $\mathcal{X} \subseteq \mathbb{N}^l$, or continuous, $\mathcal{X} \subseteq \mathbb{R}^l$. As we provide a *continuous-time* description for the latent processes while observations are generated at *discrete* time points, our model belongs to the class of continuous-discrete models, on which a large body of literature exists in the field of stochastic filtering (Maybeck, 1982; Daum, 1984; Särkkä & Solin, 2019). We emphasize that this model formulation is of great practical relevance, as data is often recorded at discrete time points, while the system of interest evolves continuously in time, see, e.g., (Cassandras & Lafortune, 2009).

2.2. Modeling Assumptions

The background presented so far was general to any hybrid processes. For the remainder of the paper, we consider a time-homogeneous prior switching process Z with rate function $\Lambda(z, z', t) = \Lambda(z, z')$ and we parametrize the initial distributions as $p(z_0) = \mathrm{Cat}(z_0 \mid \pi_{z_0})$, π_{z_0} a vector of individual entries $\pi^i_{z_0} \in [0, 1], \sum_i \pi^i_{z_0} = 1$. Furthermore, we focus on mode-dependent linear time-invariant stochastic systems, i.e.,

$$f(z, y, t) = f(z, y) = A(z)y + b(z), \tag{6}$$

where the affine drift function is parameterized for each mode $z \in \mathcal{Z}$ by $A(z) \in \mathbb{R}^{n\times n}$ and $b(z) \in \mathbb{R}^n$. We define the shorthands $\Gamma(z) := [A(z), b(z)] \in \mathbb{R}^{n\times n+1}$ and $\bar{y} := \left[y^\top, 1_n^\top\right]^\top \in \mathbb{R}^{n+1}$, where 1_n is the n-dimensional all-ones vector. This parametrization yields a linear system in the variable $\bar{y}$, i.e., $A(z)y + b(z) = \Gamma(z)\bar{y}$. For the mode-dependent linear time-invariant stochastic system, we assume a time-homogeneous and state-independent dispersion, i.e., $Q(z, y, t) = Q(z)$ and we assume a Gaussian initial distribution, which we parameterize as $p(y_0) = \mathcal{N}(y_0 \mid \mu_0, \Sigma_0)$.

Lastly, we follow a linear observation model for the data as $X_i = Y_i + \zeta$, with the zero-mean Gaussian noise $\zeta \sim \mathcal{N}(\zeta \mid 0, \Sigma_x)$ and observation covariance Σ_x. Hence, the observation likelihood for the ith data point X_i is given as

$$X_i \sim \mathcal{N}(x_i \mid y_i, \Sigma_x). \tag{7}$$

3. Inference

For inference, we take on a Bayesian view similar to discrete-time filtering and smoothing (Särkkä, 2013). We consider a set $x_{[1,N]} := \{x_i\}_{i=1}^N$ of N observations obtained at time points $0 \le t_1, ..., t_N \le T$. We are interested in computing a path-wise posterior distribution over the latent hybrid process $\{Z(t), Y(t)\}$ on the interval $[0, T]$ and all of its parameters Θ given the observed data, i.e.,

$$\mathsf{P}\left(\left(Z_{[0,T]}, Y_{[0,T]}, \Theta\right) \in \cdot \mid x_{[1,N]}\right). \tag{8}$$

As is common in Bayesian inference, computing Eq. (8) is intractable as (i) integration over a high-dimensional parameter space is hard and (ii) computing the required posterior distribution over the latent hybrid process

$$\mathsf{P}\left(\left(Z_{[0,T]}, Y_{[0,T]}\right) \in \mathcal{A} \mid x_{[1,N]}, \theta\right) = \int_{\mathcal{A}} \mathrm{d}\mathsf{P}(\omega_T \mid x_{[1,N]}, \theta),$$

is infeasible. This can be seen by noting that the posterior distribution over the latent hybrid process can be equivalently expressed via the *smoothing distribution*, that is,

the time point-wise posterior marginal density $p(z, y, t \mid x_{[1,N]}, \theta)$, cf. Eq. (4). For this quantity, an exact evolution equation can be derived (Köhs et al., 2021). While the above posterior can hence be computed in principle, this is not a feasible option even for toy systems as it requires the solution of two coupled PDEs of type (5) with discrete and continuous components.

To overcome these issues, we propose a blocked Gibbs sampler, where the switching process $Z_{[0,T]}$, the diffusion process $Y_{[0,T]}$ and the parameters Θ are sampled in turn from the complete conditional measures, i.e., the measures conditioned on each other and the data $x_{[1,N]}$:

$$Y_{[0,T]} \sim \mathsf{P}(Y_{[0,T]} \in \cdot \mid z_{[0,T]}, x_{[1,N]}, \theta), \tag{9}$$
$$Z_{[0,T]} \sim \mathsf{P}(Z_{[0,T]} \in \cdot \mid y_{[0,T]}, x_{[1,N]}, \theta), \tag{10}$$
$$\Theta \sim \mathsf{P}(\Theta \in \cdot \mid z_{[0,T]}, y_{[0,T]}, x_{[1,N]}). \tag{11}$$

Hence, this scheme yields samples from the desired joint posterior distribution in Eq. (8).

The path-wise quantities Eqs. (9) and (10) can be shown to each describe conditional Markov processes. In principle, we desire to sample from these processes in a manner akin to the forward- and backward-recursion in traditional discrete-time hidden Markov models (HMMs) (Barber, 2012). In contrast to the discrete-time case, however, we can not obtain such recursions by naively carrying out Bayes' rule using densities, as it is not possible to define sensible probability densities on the path-space Ω, cf. Section 2. Sampling from the full conditional measures is hence non-trivial. Drawing on results from filtering and smoothing theory, we derive in the following the evolution equations for Eqs. (9) and (10) on the process level, allowing us to generate the desired samples. As Eq. (9) and Eq. (10) both depend on the same Brownian motion instance W, complications arise due to the asymmetry of the Itô integral which are not found in discrete time. To circumvent these issues, all stochastic integrations are carried out forward in time. This results in a backward-forward/forward-backward (BFFB) scheme, where the time direction of the actual path simulation is reversed between the posterior Z- and Y-paths.

We additionally sample the model parameters from the full conditional distribution Eq. (11). By the use of conjugate prior distributions, Eq. (11) yields closed-form distributions for all parameters. For the dispersion Q, we do not obtain the posterior directly, but utilize a Metropolis-adapted Langevin scheme, ensuring numerical stability (Besag et al., 1995).

We emphasize that inference schemes for this versatile class of processes are rare: to the best of our knowledge, the only framework available is a recent variational approach (Köhs et al., 2021), which however needs to make strong approximating assumptions and does not provide Bayesian parameter estimates. In this approach, the exact smoothing density $p(z, y, t \mid x_{[1,N]})$ is approximated by a simple mixture of Gaussian processes (GPs), which is unable to resolve non-stationary diffusion dynamics.

3.1. Gibbs Step: Sampling the Conditional Diffusion Process $Y_{[0,T]}$

To sample from the full conditional diffusion path measure

$$Y_{[0,T]} \sim \mathsf{P}(Y_{[0,T]} \in \cdot \mid z_{[0,T]}, x_{[1,N]}, \theta), \tag{12}$$

we first acknowledge that by conditioning on the process $z_{[0,T]}$ the generative model presented in Section 2 reduces to a SSDE with a fixed switching path. We can interpret this SSDE (2) as a temporal sequence of individual SDEs, or, equivalently, as we assume all drift functions Eq. (6) to be affine, as one SDE with an explicit time-dependence of the parameters; more concretely, we have a drift function

$$f(z(t), y) = A(z(t))y + b(z(t)) \equiv f(y, t) \tag{13}$$

and the dispersion $Q(z(t), y, t) = Q(z(t)) \equiv Q(t)$. Hence, the path measure in Eq. (12) is governed by the solution of an SDE process conditioned on the data $x_{[1,N]}$. For a conventional, i.e., non-switching, SDE,

$$\mathrm{d}Y(t) = f(Y(t), t)\,\mathrm{d}t + Q(t)\,\mathrm{d}W(t), \tag{14}$$

it is well known that the posterior process conditioned on some data $x_{[1,N]}$ can in turn via Doob's h-transform (Doob, 1984; Rogers & Williams, 2000) be expressed as an SDE with a modified drift function $\tilde{f}(Y(t), t)$:

$$\begin{aligned} \mathrm{d}Y(t) &= \tilde{f}(Y(t), t)\,\mathrm{d}t + Q(t)\,\mathrm{d}W(t), \\ \tilde{f}(Y(t), t) &= f(Y(t), t) + D(t)\partial_y \log \beta(Y(t), t), \end{aligned} \tag{15}$$

where the noise covariance is given by $D(t) = Q(t)Q^\top(t)$. The reverse-filtered likelihood, an analogue to the backward messages in discrete-time HMMs,

$$\beta(y, t) \coloneqq p(x_{[k,N]} \mid y, t),\ t_k \geq t, \tag{16}$$

has to fulfill the Kolmogorov backward equation (KBE) (Särkkä, 2013)

$$\begin{aligned} \partial_t \beta(y, t) = &- \sum_{i=1}^{n} f_i(y, t)\partial_{y_i}\beta(y, t) \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}(t)\partial_{y_i}\partial_{y_j}\beta(y, t), \end{aligned} \tag{17}$$

with the end-point condition at $t = T$ as $\beta(y, T) = 1$.

Under the linearity assumption, the KBE (17) can be evaluated in closed form, yielding a backward Kalman-type

filter. Note that these results are known, e.g., from results on smoothing for nonlinear diffusions (Mider et al., 2021). We provide the derivations in Appendix A.1 for completeness. In between observations, we find

$$\beta(y,t) = \mathcal{N}\left(x_k, ..., x_N \mid F(t)y + m(t), \Sigma(t)\right),\ t_k \geq t, \tag{18}$$

where $F(t) \in \mathbb{R}^{(N-k)n\times n}$, $m(t) \in \mathbb{R}^{(N-k)n}$, and $\Sigma(t) \in \mathbb{R}^{(N-k)n\times(N-k)n}$ are determined by a set of ordinary differential equations (ODEs). Note, however, that the support of this distribution increases with each observation incorporated. By re-interpreting the Gaussians as distributions over y rather than x, this can be decomposed as

$$\log\beta(y,t) = -c(t) - \frac{1}{2}y^\top I(t)y + a(t)^\top y, \tag{19}$$

where $c(t), I(t)$ and $a(t)$ depend on the parameters of Eq. (18). Importantly, (i) these parameters are fixed in size, $c(t) \in \mathbb{R}$, $I(t) \in \mathbb{R}^{n\times n}$, and $a(t) \in \mathbb{R}^n$; and (ii) we do not require the normalizer $c(t)$, as Eq. (15) only depends on the gradient

$$\partial_y \log\beta(y,t) = -I(t)y + a(t). \tag{20}$$

For these parameters, the KBE yields a continuous-time backward analogue to the discrete-time information filter (Stengel, 1994),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}I(t) &= -A(t)^\top I(t) - I(t)A(t) + I(t)D(t)I(t),\\ \frac{\mathrm{d}}{\mathrm{d}t}a(t) &= -A(t)^\top a(t) + I(t)D(t)a(t) + I(t)b(t).\end{aligned} \tag{21}$$

At the observation times, the ODE solutions are subject to the usual reset conditions (Kushner, 1964)

$$I(t_i) = \Sigma_x^{-1} + I(t_i^+), \quad a(t_i) = \Sigma_x^{-1}x_i + a(t_i^+), \tag{22}$$

where $a(t_i^+) \coloneqq \lim_{h\searrow 0} a(t_i + h)$. Note that the ODEs Eq. (21) together with the reset conditions constitute a set of *impulsive* ODEs (Samoilenko & Perestyuk, 1995). Having computed $\partial_y \log\beta(y,t)$ backward from $t = T$ to $t = 0$, we can then straightforwardly simulate the SDE (15) forward in time. Hence, we can simulate the SDE

$$\begin{aligned}\mathrm{d}Y(t) = \big\{\big[A(z(t)) - Q(z(t))Q^\top(z(t))I(t)\big]\,Y(t)\\ +b(z(t)) + a(t)\big\}\,\mathrm{d}t + Q(z(t))\,\mathrm{d}W(t),\end{aligned} \tag{23}$$

which yields samples from the full conditional distribution in Eq. (12). For the initial value of the full-conditional path measure, we have

$$Y(0) \mid Z_{[0,T]}, X_{[1,N]}, \Theta \sim \mathcal{N}\left(y_0 \mid \bar{\mu}, \bar{\Sigma}\right),$$

where we show in Appendix A.2 that

$$\bar{\mu} = \bar{\Sigma}(\Sigma_0^{-1}\mu_0 + a(0)), \quad \bar{\Sigma} = \left(\Sigma_0^{-1} + I(0)\right)^{-1}. \tag{24}$$

3.2. Gibbs Step: Sampling the Conditional Switching Process $Z_{[0,T]}$

With the simulated SSDE path $y_{[0,T]}$, we aim to sample from the switching full conditional path measure,

$$Z_{[0,T]} \sim \mathsf{P}(Z_{[0,T]} \in \cdot \mid y_{[0,T]}, x_{[1,N]}, \theta),$$

which due to the Markovian structure described in Section 2 reduces to

$$Z_{[0,T]} \sim \mathsf{P}(Z_{[0,T]} \in \cdot \mid y_{[0,T]}, \theta). \tag{25}$$

Notice that this setting is qualitatively different from the continuous-discrete smoothing problem in Section 3.1, as the observations now consist in a full path instead of a finite set of points. In principle, we would like to pursue a similar approach as in the preceding section, i.e., a backward-filtering, forward-sampling scheme based on a set of filtering ODEs akin to Eq. (21). Appropriate theoretical results exist for pure SDE systems with state-independent drift as *robust* filters and smoothers (Pardoux, 1980; Van Handel, 2007). The state-dependence of the SSDE drift function Eq. (2) does however not admit a robust ODE formulation of the reverse-filtered likelihood analogous to Eq. (21) (Davis, 1979; Crisan et al., 2013). Instead, this quantity will depend on a stochastic integral with respect to the same Brownian motion W that generated the forward diffusion process. Roughly speaking, when trying to obtain the sought-after reverse quantity starting from $t = T$, one would have to integrate with respect to a process that is known for every $t' < t$ and hence "looks into the future". It is not obvious whether it is possible to find such a reverse process, but see, e.g., (Nualart & Pardoux, 1988; Nualart, 2006) on anticipative stochastic calculus.

To circumvent these peculiarities, we resort to the formulation of the filter as a stochastic integral. We adopt a reversed approach, where we first compute the filtering distribution forward in time and subsequently simulate backwards: we thereby can jointly solve all occurring stochastic integrals in the *same time direction* and hence compute both the path $y_{[0,T]}$ and the filtering distribution simultaneously. We first compute the filtering density $p_f(z,t)$, which follows from the conditional path measure $\mathsf{P}(Z_{[0,t]} \in \mathrm{d}z_{[0,t]} \mid y_{[0,t]})$ as an expectation:

$$\begin{aligned}p_f(z,t) &= \mathsf{E}\left[\mathbb{1}(z(t) = z) \mid y_{[0,t]}\right]\\ &= \int \mathbb{1}(z(t) = z)\mathsf{P}(Z_{[0,t]} \in \mathrm{d}z_{[0,t]} \mid y_{[0,t]}).\end{aligned} \tag{26}$$

This path measure is expressible via the measure of a standard Brownian motion W, $\mathsf{P}(W_{[0,t]} \in \mathrm{d}y_{[0,t]})$, utilizing Girsanov's theorem (Øksendal, 2003)

$$\begin{aligned}\mathsf{P}\left(Z_{[0,t]} \in \mathrm{d}z_{[0,t]} \mid y_{[0,t]}\right) \propto G(z_{[0,t]}, y_{[0,t]})\cdot\\ \mathsf{P}(W_{[0,t]} \in \mathrm{d}y_{[0,t]})\mathsf{P}\left(Z_{[0,t]} \in \mathrm{d}z_{[0,t]}\right)\end{aligned} \tag{27}$$

with the Radon-Nikodym derivative between the conditional and the Brownian motion measure,

$$G(z_{[0,t]}, y_{[0,t]}) := \frac{\mathrm{d}\mathsf{P}(Y_{[0,t]} \in \mathrm{d}y_{[0,t]} \mid z_{[0,t]})}{\mathrm{d}\mathsf{P}(W_{[0,t]} \in \mathrm{d}y_{[0,t]})} \tag{28}$$
$$= \exp\left\{ \int_0^t f^\top(z(s), y(s)) D^{-1}(z(s)) \mathrm{d}y(s) - \frac{1}{2} \int_0^t f^\top(z(s), y(s)) D^{-1}(z(s)) f(z(s), y(s)) \mathrm{d}s \right\}.$$

Computing a stochastic differential equation for Eq. (26) utilizing Itô's lemma yields a Kushner-Stratonovich SDE describing its time evolution (Del Moral & Penev, 2017)

$$\mathrm{d}p_f(z,t) = \sum_{z' \in \mathcal{Z}} \Lambda(z', z, t) p_f(z', t) \mathrm{d}t + p_f(z,t) \cdot (f(y,z) - \bar{f}(y,t))^\top D^{-1}(z)(\mathrm{d}y(t) - \bar{f}(y,t)\mathrm{d}t) \tag{29}$$

with $\bar{f}(y,t) = \sum_{z' \in \mathcal{Z}} f(z', y) p_f(z', t)$; for a detailed mathematical derivation, see Appendix B. Note that for $f(y,z) = f(z)$, we recover the classical Wonham filter (Wonham, 1964).

It is a known result that the smoothing density $p_s(z,t) := p(z, t \mid y_{[0,T]})$ admitted by the sought-after switching path measure in Eq. (25) can be expressed as a backward master equation via the forward filtered density $p_f(z,t) := p(z, t \mid y_{[0,t]})$ (Anderson & Rhodes, 1983; Van Handel, 2007). The backward master equation reads

$$\frac{\mathrm{d}}{\mathrm{d}t} p_s(z,t) = -\sum_{z' \in \mathcal{Z}} \tilde{\Lambda}(z', z, t) p_s(z', t), \tag{30}$$

with the end-point condition $p_s(z,T) = p_f(z,T)$. The backward rates in Eq. (30) are defined as

$$\tilde{\Lambda}(z, z', t) := \lim_{h \searrow 0} h^{-1} \mathsf{P}(Z(t-h) = z' \mid Z(t) = z),$$

and determined by the filtering density,

$$\tilde{\Lambda}(z', z, t) = \frac{p_f(z', t)}{p_f(z, t)} \Lambda(z, z'), \tag{31}$$

with the usual exit rates

$$\tilde{\Lambda}(z, z, t) := -\sum_{z' \in \mathcal{Z} \setminus z} \tilde{\Lambda}(z, z', t).$$

This allows us to backward-sample a new path $z_{[0,T]}$ after forward-filtering via Eq. (29). To simulate from the conditional switching process Z with time-dependent rates Eq. (31), we utilize the thinning algorithm, which generalizes the standard Doob-Gillespie algorithm (Lewis & Shedler, 1979).

Altogether, generating one full sample path of the system requires solving the ODEs (21) and the SDE (29) as well as generating the respective paths $y_{[0,T]}, z_{[0,T]}$. Generally, the computational cost depends on the solver (determining, e.g., whether adaptive step sizes are used); the individual function evaluations scale as $\mathcal{O}(n^3)$ for $I(t)$, $\mathcal{O}(n^2)$ for $a(t)$ and $\mathcal{O}(|\mathcal{Z}|^2 + |\mathcal{Z}|n^2)$ for $p_f(t)$. Note that by utilizing the BFFB scheme, we can sample $y_{[0,T]}$ while simultaneously computing Eq. (29), saving one full pass through the trajectory compared to an approach where both $y_{[0,T]}$ and $z_{[0,T]}$ would be computed in a backward-forward manner.

Algorithm 1 BFFB Gibbs Sampling for Continuous-Time Switching Dynamical Systems

Input: observation data $\{t_i, x_i\}_{i=1,\dots,N}$
Initialize $z^0_{[0,T]}, y^0_{[0,T]}, \theta^0$
for i = 0, ..., NumSamples **do**
  Given $z^i_{[0,T]}$, compute $\partial_y \log \beta$ using Eq. (21)
  Given $z^i_{[0,T]}$, sample $y^{i+1}_{[0,T]}$ according to (23)
  Given $y^{i+1}_{[0,T]}$, compute p_f using Eq. (29)
  Given $y^{i+1}_{[0,T]}$, sample $z^{i+1}_{[0,T]}$ according to Eq. (31)
  Given $z^{i+1}_{[0,T]}, y^{i+1}_{[0,T]}$, sample model parameters θ^{i+1}
end for

3.3. Gibbs Step: Sampling the Parameters Θ

Our framework naturally lends itself to Bayesian parameter estimation. In the following, we specify the used prior distributions over the model parameters and provide the resulting full conditionals one at a time. With this model definition, the prior parameter distributions are conjugate to the respective likelihoods, ensuring tractability. The posterior distributions hence are found by updating the distribution hyperparameters. Due to space constraints, we omit the densities of the used distributions as well as the mathematical details of the updates and state here only the results, but provide all definitions and derivations in Appendix C. For a comprehensive overview over (conjugate) distributions, see, e.g., (Gelman et al., 2013).

Initial Conditions We impose a Dirichlet prior with hyper-parameter α_{z_0} on the initial MJP state distribution parameter π_{z_0}, resulting in

$$p(\pi_{z_0} \mid z_{[0,T]}) = \mathrm{Dir}(\pi_{z_0} \mid \alpha_{z_0} + \delta_{z(0)}), \tag{32}$$

with the point mass $\delta_{z(0)}$ on the value $z(0)$. Note that we suppress all variables in the conditioning set that π_{z_0} is conditionally independent of. We shall keep with this notation in all following update equations.

On the SSDE initial distribution parameters μ_0 and Σ_0 we place a Normal-inverse-Wishart (NIW) prior, with hyper-

parameters $(\eta, \lambda, \Psi, \kappa)$, yielding

$$p(\mu_0, \Sigma_0 \mid y_{[0,T]}) = \mathcal{NIW}\left(\mu_0, \Sigma_0 \,\middle|\, \tilde{\eta}, \tilde{\lambda}, \tilde{\Psi}, \tilde{\kappa}\right) \quad (33)$$

with

$$\begin{aligned} \tilde{\eta} &= \frac{\lambda\eta + y(0)}{\lambda + 1}, \quad \tilde{\lambda} = \lambda + 1, \quad \tilde{\kappa} = \kappa + 1, \\ \tilde{\Psi} &= \left(\Psi^{-1} + \frac{\lambda}{\lambda+1}(y(0) - \eta)(y(0) - \eta)^\top\right)^{-1}. \end{aligned} \quad (34)$$

MJP Rates We assume the prior rates to be given by a Gamma distribution: Introducing the shorthand $\Lambda_{zz'} := \Lambda(z, z')$,

$$p\left(\Lambda_{zz'}\right) = \mathrm{Gam}(\Lambda_{zz'} \mid s, r) \quad (35)$$

with the shape $s \in \mathbb{R}_{>0}$ and rate parameter $r \in \mathbb{R}_{>0}$. Characterizing the simulated path $z_{[0,T]}$ via (i) the *sojourn times* $\{\tau_k\}$ between jumps, and (ii) the *state sequence* $\{z_k\}$, $k = 0, ..., K$, allows to express the path likelihood $p(z_{[0,T]} \mid \Lambda_{zz'})$ via the observed transitions $N_{zz'} = \sum_{k=0}^{K-1} \mathbb{1}(z_k = z \wedge z_{k+1} = z')$ and cumulative sojourn times $T_z = \sum_{k=0}^{K} \mathbb{1}(z_k = z)\tau_k$, yielding a posterior Gamma distribution for the rates,

$$p(\Lambda_{zz'} \mid z_{[0,T]}) = \mathrm{Gam}(\Lambda_{zz'} \,|\, s + N_{zz'}, r + T_z). \quad (36)$$

SDE Drift Parameters In the following, we utilize as above the shorthand $\Gamma_z := \Gamma(z)$. The SSDE parameters Γ_z are specified via a Matrix-Normal (MN) prior

$$p(\Gamma_z) = \mathcal{MN}\left(\Gamma_z \mid M_z, D_z, K_z\right), \quad (37)$$

where $D_z = D(z)$ is the SSDE covariance. Expressing the conditional Y-posterior via the Radon-Nikodym derivative $G(z_{[0,T]}, y_{[0,T]})$, c.f. Eq. (28), we can interpret G as the likelihood of the drift parameters, $G(z_{[0,T]}, y_{[0,T]}) = G(z_{[0,T]}, y_{[0,T]} \mid \{\Gamma_z\})$,

$$p(\Gamma_z \,|\, z_{[0,T]}, y_{[0,T]}) \propto G(z_{[0,T]}, y_{[0,T]} \,|\, \{\Gamma_z\}) p(\Gamma_z). \quad (38)$$

We show in Appendix C that the prior is conjugate with respect to G when $y_{[0,T]}$ is simulated utilizing an Euler-Maruyama solver. Accordingly,

$$p(\Gamma_z \,|\, y_{[0,T]}, z_{[0,T]}) = \mathcal{MN}\left(\Gamma_z \,|\, \tilde{M}_z, D_z, \tilde{K}_z\right). \quad (39)$$

See Appendix C for the hyperparameter update equations. Note that this prior does not guarantee stability of the individual modes, as it does not impose any constraints on the eigenvalue spectrum of the sub-matrices $A(z)$. However, it is known that for switching systems, global stability of the system does not require strict intra-mode stability (Mao, 1999). Conditioned on data from a stable mode, the posterior will in any event be likely peaked around stable matrices.

SDE Dispersion The dispersion $Q(z)$ has a special role among the model parameters, as together with the used time step-size, it determines the accuracy of the SDE solver. While a posterior dispersion can be derived in the same vein as for the drift parameters Eq. (38), the resulting posteriors may cause the solver to become unstable if Q becomes too large. One may approach this issue by utilizing adaptive step-size solvers (Kloeden & Platen, 1992). For simplicity, we instead apply a Metropolis-adapted Langevin sampling scheme (Roberts & Rosenthal, 1998) for $Q(z) = Q_z$, applying the usual shorthand. We simulate an SDE on the space of dispersion matrices with step-size $0 < \xi \ll 1$ and $\varepsilon \sim \mathcal{N}(0, \mathbb{I}_n)$ (Roberts & Rosenthal, 1998),

$$Q_z^* = Q_z + \xi \partial_{Q_z} \log p(Q_z \mid y_0, y_h, \ldots, y_{Lh}) + \sqrt{2\kappa}\varepsilon. \quad (40)$$

Here, $p(Q_z \mid \{y_{hl}\})$ is the approximation of $p(Q_z \mid y_{[0,T]}) \propto G(z_{[0,T]}, y_{[0,T]}) p(Q_z)$ on the SDE simulation time grid, $t_0 = 0, t_1 = h, ..., t_L = Lh = T$ with time-step h. In practice, we parameterize Q_z via

$$D_z = Q_z Q_z^\top \sim \mathcal{IW}(D_z \mid \Psi_{D_z}, \lambda_{D_z}). \quad (41)$$

Note that as shown in Appendix C, the approximate density $p(Q_z \mid \{y_{hl}\})$ is equivalent to a product of $L - 1$ Gaussian transition distributions $\mathcal{N}(y_l \mid y_{l-1}, D_z h)$. We then utilize a Metropolis rejection scheme with acceptance probability

$$A(Q, Q^*) = \frac{p(Q_z^* \mid y_{[0,T]}) q(Q \mid Q^*)}{p(Q_z \mid y_{[0,T]}) q(Q^* \mid Q)}, \quad (42)$$

where q denotes the (Gaussian) proposal density induced by Eq. (40). Note that we expect slow mixing of the sampler with respect to Q, which is due to the fact that the measures of two diffusion processes with different dispersions are singular with respect to each other (Shephard & Pitt, 1997). As potential solutions to this peculiar issue are quite involved, an in-depth analysis of this issue is outside the scope of the present study, but see , e.g., (Shephard & Pitt, 1997; Golightly & Wilkinson, 2008). For our purposes, we are satisfied with a tractable and numerically stable posterior and hence defer respective extensions to future works.

Observation Covariance Lastly, we impose an inverse-Wishart (IW) prior on the observation covariance Σ_x, resulting in $p(\Sigma_x \mid x_{[1,N]}) = \mathcal{IW}(\Sigma_x \mid \tilde{\Psi}_x, \tilde{\lambda}_x)$. We summarize the full Gibbs sampling algorithm in Algorithm 1.

4. Results

We first verify the method on data generated under the modeling assumptions and compare with the existing variational framework (Köhs et al., 2021). Subsequently, we use the Gibbs sampler to infer the latent expression states of an inducible gene expression system. In all experiments, the

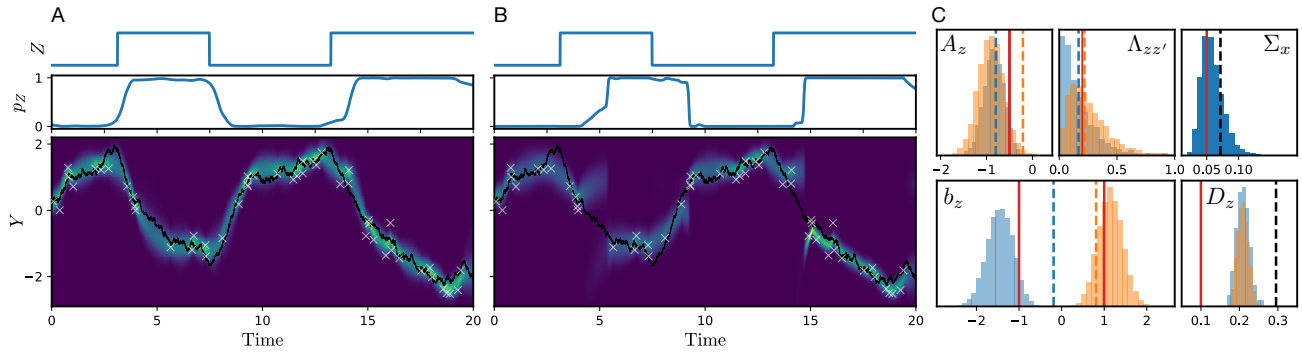

Figure 2. Model validation on synthetic data and comparison with variational results. A: Results of the MCMC method. Top: ground-truth switching trajectory $z_{[0,T]}$. Middle: empirical posterior $p(z, t \mid x_{[1,N]})$. Brighter colors indicate higher probability density. Bottom: respective posterior $p(y, t \mid x_{[1,N]})$. Black solid line: ground-truth latent trajectory $y_{[0,T]}$. White crosses: observations. $N_{\text{samples}} = 10000$. B: Results of the variational method (Köhs et al., 2021). Middle: variational posterior $q(z, t \mid x_{[1,N]})$. Bottom: variational posterior $q(y, t \mid x_{[1,N]})$. C: Parameter estimates of the drift parameters $A(z), b(z)$, cf. Eq. (6), the MJP rates $\Lambda(z, z')$, the SDE covariance $D(z)$ and the observation covariance Σ_x. Red lines: true values. Blue and orange shading indicates the two modes $z = 1, 2$ where applicable. Dashed lines: variational point estimates.

hyperparameters are set empirically; for other options, see, e.g. (Casella, 2001). We initialize the Gibbs sampler in the same way to start at reasonable parameter values such as to achieve fast burn-in. All hyperparameters are provided in Appendix D.

4.1. Verification on Ground-Truth Data

1D System We test the method on synthetic data from a one-dimensional two-mode switching dynamical system as specified in Section 2.2. Our Gibbs sampling scheme is able to faithfully recover the latent ground-truth trajectories; both $z_{[0,T]}$ and $y_{[0,T]}$ are reproduced with high fidelity, see Fig. 2. In particular, we note that the smooth relaxation from one set-point to the other upon a switch of the Z-process is accurately reconstructed. For comparison, we run the variational method (Köhs et al., 2021) on the same observations $x_{[1,N]}$. This method returns approximate posterior marginal densities $q(z, t \mid x_{[1,N]})$ and $q(y, t \mid x_{[1,N]})$. The variational framework fails to capture the non-stationary transition periods upon Z-switches and results in bimodal posterior marginals, exhibiting a "gap" in the posterior density $q(y, t \mid x_{[1,N]})$, which is furthermore reflected in delayed transitions of $q(z, t \mid x_{[1,N]})$. Furthermore, we obtain accurate Bayesian parameter estimates. Note that all posteriors except the diffusion covariance D_z cover the ground truth very well. While the distribution of the latter is in total closer to the ground truth than the variational estimate, we observe slow mixing in this parameter, as discussed in Section 3.3.

2D System To demonstrate our method on more complex problems, we additionally apply it to a 2D problem, where the continuous dynamics are given by two counter-rotating "swirls". As in the 1D case, we find that the ground truth paths $z_{[0,T]}$ and $y_{[0,T]}$ are faithfully recovered, but we do observe that the oscillatory behavior of the diffusion components is reflected in the Z-reconstruction, Fig. 3.

4.2. Inference of Gene-Switching Dynamics

We use our framework to infer the switching dynamics of an inducible gene system measured in-house. We expressed an inducible green fluorescent protein (GFP) in the eucaryotic model organism *Saccharomyces cerevisiae*. Using a microfluidic platform, gene expression can be induced at arbitrary time points by a chemical control signal utilizing β-estradiol (Hofmann et al., 2019). Upon induction, expression of the GFP-encoding gene is initiated. We measure the amount of GFP fluorescence over time using fluorescence microscopy. The dynamics of transcription and translation is commonly modeled by switching SDEs with rate parameters depending on the stochastic promoter state of the gene ("on" vs. "off") (Ocone et al., 2013). We aim to infer the latent stochastic promoter state, the GFP level and the rate parameters from a set of noisy microscopy measurements. Although we have no ground truth available, we can rationalize the inferred promoter activity shown in the upper panel of Fig. 4 in the context of the available inducer control signal. Upon addition of β-estradiol to the medium, a certain delay is incurred through molecular diffusion until the promoter gets activated. Similarly, upon removal of the chemical from the medium, promoter deactivation is governed by the export of the chemical from the cell through diffusion. Along the same lines, the inferred GFP level shown in the lower panel of Fig. 4 is in line with the elementary processes of ribonucleic acid (RNA) transcription and protein translation that result in a delayed activation and

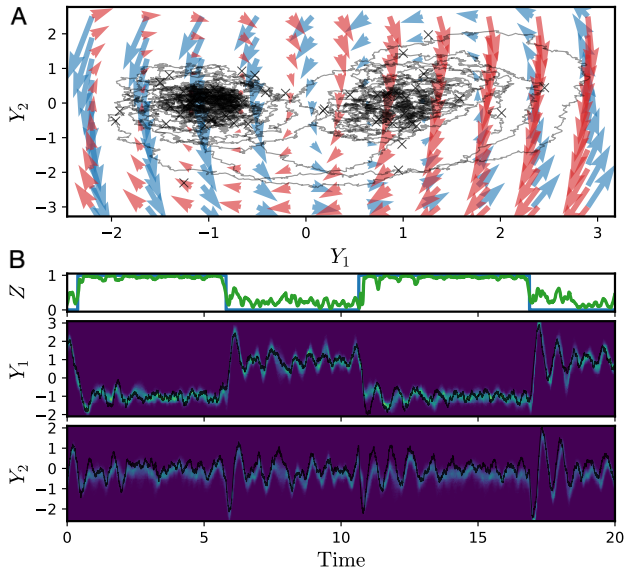

Figure 3. Model validation on synthetic 2D data. A: phase space representation of the flow (6) of both modes (red and blue arrows) revolving around $[-1, 0]^\top$ and $[1, 0]^\top$. Black solid line: ground-truth trajectory $y_{[0,T]}$. Crosses: observed data points. Note that for the sake of clarity, most observations are not shown. B: Top: ground-truth trajectory $z_{[0,T]}$ (blue) and empirical posterior $p(z, t \mid x_{[1,N]})$ (green). Middle and Bottom: components of the posterior $p(y, t \mid x_{[1,N]})$. Black solid line as above.

deactivation on the protein level with respect to the promoter state. The estimates of the rate parameters are provided in Appendix D.

5. Discussion

We presented, to the best of our knowledge, the first tractable sampling-based path-space inference scheme for discretely observed continuous-time switching dynamical systems. This enables accurate Bayesian posterior inference in settings where existing variational inference methods (Köhs et al., 2021) fail. We derived a blocked Gibbs sampler, where we generate sample paths from the exact full conditionals for the switching and diffusion components. Additionally, we sample the posterior parameters of the system by the use of conjugate updates. In future work, an exciting direction is to extend the method to non-linear settings, where, e.g., samples from our method could serve as a proposal distribution in a particle smoother setting (Klaas et al., 2006; Mider et al., 2021). Further, to allow for more expressive observation likelihoods, the latent hybrid process $\{Z(t), Y(t)\}$ could be combinded with recent neural network approaches to continuous-time processes (Li et al., 2020).

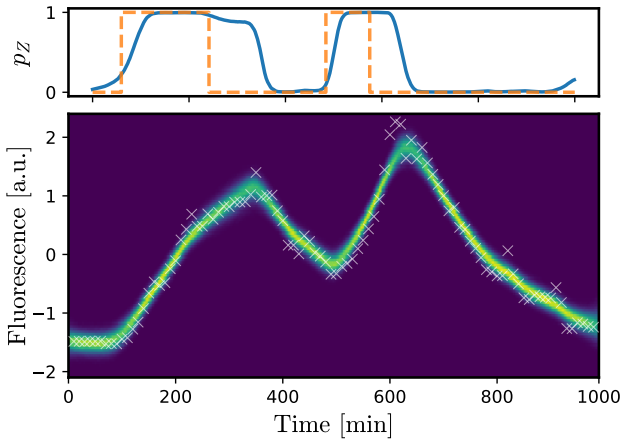

Figure 4. Inference of promoter states for an inducible gene expression system. Top: empirical posterior $p(z, t \mid x_{[1,N]})$ (blue) and chemical control (orange; "off", state 1, and "on", state 2). Note that this does not directly correspond to the promoter-"on" and promoter-"off" state, as a the inducer has to diffuse in and out the cell and its nucleus. Bottom: empirical posterior $p(y, t \mid x_{[1,N]})$. White crosses: observed data. $N_{\text{samples}} = 10000$.

Acknowledgements

We thank Tim Prangemeier for providing the gene switching data and for helpful discussions and the anonymous reviewers for their useful comments and suggestions. This work has been funded by the European Research Council (ERC) within the CONSYN project, grant agreement number 773196, and by the German Research Foundation (DFG) as part of the project B4 within the Collaborative Research Center (CRC) 1053 – MAKI.

References

Alspach, D. and Sorenson, H. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE transactions on automatic control*, 17(4):439–448, 1972.

Anderson, B. D. O. and Rhodes, I. B. Smoothing algorithms for nonlinear finite-dimensional systems. *Stochastics: An International Journal of Probability and Stochastic Processes*, 9(1-2):139–165, 1983.

Anderson, D. F., Ermentrout, B., and Thomas, P. J. Stochastic representations of ion channel kinetics and exact stochastic simulation of neuronal dynamics. *Journal of computational neuroscience*, 38(1):67–82, 2015.

Azzouzi, M. and Nabney, I. T. Modelling financial time series with switching state space models. In *Proceedings of the IEEE/IAFE 1999 Conference on Computational Intelligence for Financial Engineering (CIFEr)*, pp. 240–249. IEEE, 1999.

Bain, A. and Crisan, D. *Fundamentals of stochastic filtering*, volume 60. Springer Science & Business Media, 2008.

Barber, D. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

Besag, J., Green, P., Higdon, D., and Mengersen, K. Bayesian computation and stochastic systems. *Statistical science*, pp. 3–41, 1995.

Böker, G. and Lunze, J. Stability and performance of switching kalman filters. *International Journal of Control*, 75 (16-17):1269–1281, 2002.

Bressloff, P. C. Switching diffusions and stochastic resetting. *Journal of Physics A: Mathematical and Theoretical*, 53 (27):275003, 2020.

Casella, G. Empirical Bayes Gibbs sampling. *Biostatistics*, 2(4):485–500, 2001.

Cassandras, C. G. and Lafortune, S. *Introduction to discrete event systems*. Springer Science & Business Media, 2009.

Cassandras, C. G. and Lygeros, J. *Stochastic hybrid systems*. CRC Press, 2018.

Çınlar, E. *Probability and Stochastics*. Graduate Texts in Mathematics. Springer New York, 2011.

Crisan, D., Diehl, J., Friz, P. K., and Oberhauser, H. Robust filtering: correlated noise and multidimensional observation. *The Annals of Applied Probability*, 23(5):2139–2160, 2013.

Daum, F. E. Exact finite dimensional nonlinear filters for continuous time processes with discrete time measurements. In *The 23rd IEEE Conference on Decision and Control*, pp. 16–22. IEEE, 1984.

Davis, M. Pathwise solutions and multiplicative functionals in nonlinear filtering. In *1979 18th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes*, volume 2, pp. 176–181. IEEE, 1979.

Davis, M. H. Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):353–376, 1984.

Davis, M. H. A. *Markov models & optimization*. Routledge, 2018.

Del Moral, P. and Penev, S. *Stochastic Processes: From Applications to Theory*. Chapman and Hall/CRC, 2017.

Doob, J. L. Markoff chains–denumerable case. *Transactions of the American Mathematical Society*, 58(3):455–473, 1945.

Doob, J. L. *Classical potential theory and its probabilistic counterpart*, volume 549. Springer, 1984.

Doucet, A., Godsill, S., and Andrieu, C. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

Engell, S., Frehse, G., and Schnieder, E. *Modelling, analysis and design of hybrid systems*, volume 279. Springer, 2003.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. Nonparametric Bayesian learning of switching linear dynamical systems. *Advances in Neural Information Processing Systems*, 21:457–464, 2008.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis*. CRC press, 2013.

Golightly, A. and Wilkinson, D. J. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3): 1674–1693, 2008.

Grossmann, C., Roos, H.-G., and Stynes, M. *Numerical treatment of partial differential equations*. Springer, 2007.

Hofmann, A., Falk, J., Prangemeier, T., Happel, D., Köber, A., Christmann, A., Koeppl, H., and Kolmar, H. A tightly regulated and adjustable crispr-dcas9 based and gate in yeast. *Nucleic acids research*, 47(1):509–520, 2019.

Hu, J., Lygeros, J., and Sastry, S. Towards a theory of stochastic hybrid systems. In *International Workshop on Hybrid Systems: Computation and Control*, pp. 160–173. Springer, 2000.

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29:2946–2954, 2016.

Klaas, M., Briers, M., De Freitas, N., Doucet, A., Maskell, S., and Lang, D. Fast particle smoothing: If i had a million particles. In *Proceedings of the 23rd international conference on Machine learning*, pp. 481–488, 2006.

Kloeden, P. E. and Platen, E. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

Köhs, L., Alt, B., and Koeppl, H. Variational inference for continuous-time switching dynamical systems. *Advances in Neural Information Processing Systems*, 34, 2021.

Kushner, H. J. On the differential equations satisfied by conditional probablitity densities of Markov processes, with applications. *Journal of the Society for Industrial*

and Applied Mathematics, Series A: Control, 2(1):106–119, 1964.

Lewis, P. W. and Shedler, G. S. Simulation of nonhomogeneous Poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

Li, X., Wong, T.-K. L., Chen, R. T., and Duvenaud, D. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.

Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., and Paninski, L. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial Intelligence and Statistics*, pp. 914–922. PMLR, 2017.

Lygeros, J. and Prandini, M. Stochastic hybrid systems: a powerful framework for complex, large scale applications. *European Journal of Control*, 16(6):583–594, 2010.

Mao, X. Stability of stochastic differential equations with Markovian switching. *Stochastic processes and their applications*, 79(1):45–67, 1999.

Mao, X. and Yuan, C. *Stochastic differential equations with Markovian switching*. Imperial college press, 2006.

Maybeck, P. S. *Stochastic models, estimation, and control*. Academic press, 1982.

Mider, M., Schauer, M., and van der Meulen, F. Continuous-discrete smoothing of diffusions. *Electronic Journal of Statistics*, 15(2):4295 – 4342, 2021.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Norris, J. R. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

Nualart, D. *The Malliavin calculus and related topics*, volume 1995. Springer, 2006.

Nualart, D. and Pardoux, É. Stochastic calculus with anticipating integrands. *Probability theory and related fields*, 78(4):535–581, 1988.

Ocone, A., Millar, A. J., and Sanguinetti, G. Hybrid regulatory models: a statistically tractable approach to model regulatory network dynamics. *Bioinformatics*, 29(7):910–916, 2013.

Oh, S. M., Rehg, J. M., Balch, T., and Dellaert, F. Data-driven MCMC for learning and inference in switching linear dynamic systems. Georgia Institute of Technology, 2005.

Øksendal, B. *Stochastic differential equations*. Springer, 2003.

Opper, M. and Sanguinetti, G. Variational inference for Markov jump processes. *Advances in neural information processing systems*, 20:1105–1112, 2007.

Pardoux, E. Stochastic partial differential equations and filtering of diffusion processes. *Stochastics*, 3(1-4):127–167, 1980.

Pawula, R. F. Generalizations and extensions of the Fokker-Planck-Kolmogorov equations. *IEEE Transactions on Information Theory*, 13(1):33–41, 1967.

Rao, V. and Teh, Y. W. Fast MCMC sampling for Markov jump processes and extensions. *Journal of Machine Learning Research*, 14(11), 2013.

Roberts, G. O. and Rosenthal, J. S. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Rogers, L. C. G. and Williams, D. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 2. Cambridge university press, 2000.

Samoilenko, A. and Perestyuk, M. *Impulsive differential equations*. World scientific, 1995.

Särkkä, S. *Bayesian filtering and smoothing*. Number 3. Cambridge University Press, 2013.

Särkkä, S. and Solin, A. *Applied stochastic differential equations*. Cambridge University Press, 2019.

Shephard, N. and Pitt, M. K. Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3): 653–667, 1997.

Stengel, R. F. *Optimal control and estimation*. Courier Corporation, 1994.

Střelec, M., Macek, K., and Abate, A. Modeling and simulation of a microgrid as a stochastic hybrid system. In *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, pp. 1–9. IEEE, 2012.

Tian, T. and Burrage, K. Stochastic models for regulatory networks of the genetic toggle switch. *Proceedings of the national Academy of Sciences*, 103(22):8372–8377, 2006.

Van Handel, R. *Filtering, stability, and robustness*. PhD thesis, California Institute of Technology, 2007.

Weng, Y., Liu, X., Hu, H., Huang, H., Zheng, S., Chen, Q., Song, J., Cao, B., Wang, J., Wang, S., et al. Open eyes and closed eyes elicit different temporal properties of brain functional networks. *NeuroImage*, 222:117230, 2020.

Wildner, C. and Koeppl, H. Moment-based variational inference for stochastic differential equations. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1918–1926. PMLR, 2021.

Wonham, W. M. Some applications of stochastic differential equations to optimal nonlinear filtering. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 2(3):347–369, 1964.

Markov Chain Monte Carlo for Continuous-Time Switching Dynamical Systems

— Supplementary Material —

A. Sampling the Conditional Diffusion Process

A.1. Derivation of the backward continuous-discrete Kalman filter

The backward distribution $p(x_N \mid y, t)$ between the Nth and $N-1$th observation is given by the Kolmogorov backward equation (KBE), which reads

$$\partial_t p(x_N \mid y, t) = -\mathcal{A}^\dagger p(x_N \mid y, t). \tag{43}$$

Consider the linear dynamical case, here, the adjoint operator is given by

$$\mathcal{A}^\dagger(\cdot) = (\nabla_y(\cdot))^\top (A(t)y + b(t)) + \frac{1}{2}\operatorname{tr}\left(D\nabla_y\nabla_y^\top(\cdot)\right). \tag{44}$$

Hence,

$$\partial_t p(x_N \mid y, t) = -(\nabla_y p(x_N \mid y, t))^\top (A(t)y + b(t)) - \frac{1}{2}\operatorname{tr}\left(D\nabla_y\nabla_y^\top p(x_N \mid y, t)\right), \tag{45}$$

where we assume the end-point condition

$$p(x_N \mid y, T) = \mathcal{N}(x_N \mid Fy, \Sigma). \tag{46}$$

We write the KBE component-wise as

$$\partial_t p(x_N \mid y, t) = -\sum_{k,l} \partial_{y_k} p(x_N \mid y, t) A_{kl}(t) y_l - \sum_k \partial_{y_k} p(x_N \mid y, t) b_k(t) - \frac{1}{2}\sum_{k,l} \partial_{y_k}\partial_{y_l} p(x_N \mid y, t) D_{kl}, \tag{47}$$

We use the multivariate Fourier transform

$$\hat{f}(u) = \mathcal{F}\{f(y)\} = \int f(y) e^{-iu^\top y}\,\mathrm{d}y. \tag{48}$$

Note the following rules

$$\begin{aligned}
&\mathcal{F}\{f(Ly)\} = \frac{1}{|F|}\hat{f}(F^{-\top}u)\\
&\mathcal{F}\{\nabla_y f(y)\} = iu\hat{f}(u)\\
&\mathcal{F}\{y\nabla_y f(y)\} = i\nabla_u\hat{f}(u).
\end{aligned} \tag{49}$$

Therefore, we find $\partial_t \hat{p}(x_N \mid u,t)$ as

$$\begin{aligned}
\partial_t \hat{p}(x_N \mid u,t) &= -\sum_{k,l} i\partial_{u_l} \mathcal{F}\{\partial_{y_k} p(x_N \mid y,t) A_{kl}(t)\} - \sum_k i u_k \hat{p}(x_N \mid u,t) b_k(t) - \frac{1}{2}\sum_{k,l} u_k u_l \hat{p}(x_N \mid u,t) D_{kl} \\
&= -\sum_{k,l} i\partial_{u_l} \{i u_k \hat{p}(x_N \mid u,t) A_{kl}(t)\} - \sum_k i u_k \hat{p}(x_N \mid u,t) b_k(t) - \frac{1}{2}\sum_{k,l} u_k u_l \hat{p}(x_N \mid u,t) D_{kl} \\
&= \sum_{k,l} \partial_{u_l} u_k \hat{p}(x_N \mid u,t) A_{kl}(t) + \sum_{k,l} u_k \partial_{u_l} \hat{p}(x_N \mid u,t) A_{kl}(t) - \sum_k i u_k \hat{p}(x_N \mid u,t) b_k(t) \\
&\qquad - \frac{1}{2}\sum_{k,l} i u_k i u_l \hat{p}(x_N \mid u,t) D_{kl} \\
&= \sum_{k,l} \mathbb{1}(k=l) \hat{p}(x_N \mid u,t) A_{kl}(t) + \sum_{k,l} u_k \partial_{u_l} \hat{p}(x_N \mid u,t) A_{kl}(t) - \sum_k i u_k \hat{p}(x_N \mid u,t) b_k(t) \\
&\qquad + \frac{1}{2}\sum_{k,l} u_k u_l \hat{p}(x_N \mid u,t) D_{kl}.
\end{aligned} \tag{50}$$

This can be written in vector form as

$$\partial_t \hat{p}(x_N \mid u,t) = \operatorname{tr}(A(t)) \hat{p}(x_N \mid u,t) + (\nabla_u \hat{p}(x_N \mid u,t))^\top A^\top(t) u - i u^\top b(t) \hat{p}(x_N \mid u,t) + \frac{1}{2} u^\top D u \hat{p}(x_N \mid u,t). \tag{51}$$

We define the characteristic curve as

$$\frac{\mathrm{d}}{\mathrm{d}t} u(t) = -A^\top(t) u(t), \tag{52}$$

with end-point boundary $u(T) = u_T$. The formal solution is

$$u(t) = \Phi^\top(t,T) u_T. \tag{53}$$

Hence, we have the following properties

$$\begin{aligned}
u_T &= \Phi^{-\top}(t,T) u(t) \\
\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t,T) &= -\Phi(t,T) A(t).
\end{aligned} \tag{54}$$

We build the total derivative of $\hat{p}(x_N \mid u,t)$ at $u = u(t)$ as

$$\frac{\mathrm{d}}{\mathrm{d}t} \hat{p}(x_N \mid u(t),t) = (\nabla_u \hat{p}(x_N \mid u(t),t))^\top \frac{\mathrm{d}}{\mathrm{d}t} u(t) + \partial_t \hat{p}(x_N \mid u(t),t). \tag{55}$$

Plugging in the Fourier transformed backward Fokker-Planck quantity $\partial_t \hat{p}(x_N \mid u(t),t)$ and the characteristic ODE $\frac{\mathrm{d}}{\mathrm{d}t} u(t)$, we find

$$\frac{\mathrm{d}}{\mathrm{d}t} \hat{p}(x_N \mid u(t),t) = \left(\operatorname{tr}(A(t)) - i u^\top(t) b(t) + \frac{1}{2} u^\top(t) D u(t) \right) \hat{p}(x_N \mid u(t),t). \tag{56}$$

This yields the solution

$$\hat{p}(x_N \mid u(t),t) = \exp\left(-\int_t^T \operatorname{tr}(A(s))\,\mathrm{d}s - i\left(-\int_t^T u^\top(s) b(s)\,\mathrm{d}s \right) - \frac{1}{2}\int_t^T u^\top(s) D u(s)\,\mathrm{d}s \right) \hat{p}(x_N \mid u(T),T). \tag{57}$$

For the end-point condition in Fourier space $\hat{p}(x_N \mid u(T),T)$, we compute

$$\hat{p}(x_N \mid u(T),T) = \hat{p}(x_N \mid u_T,T), \tag{58}$$

where we compute the Fourier transform as

$$\begin{aligned}
\mathcal{F}\{p(x_N \mid y,T)\} = \mathcal{F}\{\mathcal{N}(x_N \mid Fy, \Sigma)\} &= \mathcal{F}\{\mathcal{N}(Fy \mid x_N, \Sigma)\} \\
&= \frac{1}{|F|} \exp\left(-i \left(F^{-\top} u\right)^\top x_N - \frac{1}{2} \left(F^{-\top} u\right)^\top \Sigma F^{-\top} u \right),
\end{aligned} \tag{59}$$

where we assume F is quadratic and invertible. Hence we have

$$\hat{p}(x_N \mid u(T), T) = \hat{p}(x_N \mid u_T, T) = \frac{1}{|F|} \exp\left(-i \left(F^{-\top} u_T\right)^\top x_N - \frac{1}{2} \left(F^{-\top} u_T\right)^\top \Sigma F^{-\top} u_T\right). \tag{60}$$

Using the end-point condition for $\hat{p}(x_N \mid u(t), t)$, we have

$$\begin{aligned}\hat{p}(x_N \mid u(t), t) = \frac{1}{|F|} \exp\Bigg\{ &- \int_t^T \operatorname{tr}\left(A(s)\right) \mathrm{d}s - i \left(\left(F^{-\top} u_T\right)^\top x_N - \int_t^T u^\top(s) b(s) \,\mathrm{d}s\right) \\ &- \frac{1}{2} \left(\left(F^{-\top} u_T\right)^\top \Sigma F^{-\top} u_T + \int_t^T u^\top(s) D u(s) \,\mathrm{d}s\right)\Bigg\}.\end{aligned} \tag{61}$$

Using the formal solution $u(t) = \Phi^\top(t, T) u_T$ we have

$$\begin{aligned}\hat{p}(x_N \mid u(t), t) &= \frac{1}{|F|} \exp\Bigg\{ - \int_t^T \operatorname{tr}\left(A(s)\right) \mathrm{d}s - i \left(\left(F^{-\top} u_T\right)^\top x_N - u_T^\top \int_t^T \Phi(s, T) b(s) \,\mathrm{d}s\right) \\ &\quad - \frac{1}{2} \left(\left(F^{-\top} u_T\right)^\top \Sigma F^{-\top} u_T + u_T^\top \int_t^T \Phi(s, T) D \Phi^\top(s, T) \,\mathrm{d}s\, u_T\right)\Bigg\} \\ &= \frac{1}{|F|} \exp\Bigg\{ - \int_t^T \operatorname{tr}\left(A(s)\right) \mathrm{d}s - i u_T^\top F^{-1} \left(x_N - F \int_t^T \Phi(s, T) b(s) \,\mathrm{d}s\right) \\ &\quad - \frac{1}{2} u_T^\top F^{-1} \left(\Sigma + F \int_t^T \Phi(s, T) D \Phi^\top(s, T) \,\mathrm{d}s\, F^\top\right) F^{-\top} u_T\Bigg\}.\end{aligned} \tag{62}$$

Using $u_T = \Phi^{-\top}(t, T) u(t) \Longleftrightarrow u_T^\top = u^\top(t) \Phi^{-1}(t, T)$, we find

$$\begin{aligned}\hat{p}(x_N \mid u(t), t) = \frac{1}{|F|} \exp\Bigg\{ &- \int_t^T \operatorname{tr}\left(A(s)\right) \mathrm{d}s - i u^\top(t) \Phi^{-1}(t, T) F^{-1} \left(x_N - F \int_t^T \Phi(s, T) b(s) \,\mathrm{d}s\right) \\ &- \frac{1}{2} u^\top(t) \Phi^{-1}(t, T) F^{-1} \left(\Sigma + F \int_t^T \Phi(s, T) D \Phi^\top(s, T) \,\mathrm{d}s\, F^\top\right) F^{-\top} \Phi^{-\top}(t, T) u(t)\Bigg\} \\ = \frac{1}{|F|} \exp\Bigg\{ &- \int_t^T \operatorname{tr}\left(A(s)\right) \mathrm{d}s - i \left(\left(F \Phi(t, T)\right)^{-\top} u(t)\right)^\top \left(x_N - F \int_t^T \Phi(s, T) b(s) \,\mathrm{d}s\right) \\ &- \frac{1}{2} \left(\left(F \Phi(t, T)\right)^{-\top} u(t)\right)^\top \left(\Sigma + F \int_t^T \Phi(s, T) D \Phi^\top(s, T) \,\mathrm{d}s\, F^\top\right) \left(F \Phi(t, T)\right)^{-\top} u(t)\Bigg\}.\end{aligned} \tag{63}$$

Utilizing the inverse Fourier transform yields

$$\begin{aligned}p(x_N \mid y, t) &= \frac{|\Phi(t, T)|}{\exp\left(\int_t^T \operatorname{tr}\left(A(s)\right) \mathrm{d}s\right)} \mathcal{N}\left(F \Phi(t, T) y \,\middle|\, x_N - F \int_t^T \Phi(s, T) b(s) \,\mathrm{d}s, \right. \\ &\qquad\qquad \left. \Sigma + F \int_t^T \Phi(s, T) D \Phi^\top(s, T) \,\mathrm{d}s\, F^\top\right) \\ &= \frac{|\Phi(t, T)|}{\exp\left(\int_t^T \operatorname{tr}\left(A(s)\right) \mathrm{d}s\right)} \mathcal{N}\left(x_N \,\middle|\, F \Phi(t, T) y + F \int_t^T \Phi(s, T) b(s) \,\mathrm{d}s, \right. \\ &\qquad\qquad \left. \Sigma + F \int_t^T \Phi(s, T) D \Phi^\top(s, T) \,\mathrm{d}s\, F^\top\right).\end{aligned} \tag{64}$$

Jacobi's formula yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}|\Phi(t,T)| &= |\Phi(t,T)|\,\mathrm{tr}\,\left(\Phi^{-1}(t,T)\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t,T)\right)\\ &= |\Phi(t,T)|\,\mathrm{tr}\,\left(\Phi^{-1}(t,T)\left(-\Phi(t,T)A(t)\right)\right)\\ &= |\Phi(t,T)|\,\mathrm{tr}\,\left(-A(t)\right).\end{aligned} \tag{65}$$

Therefore, we have for the solution

$$|\Phi(t,T)| = \exp\left(-\int_t^T \mathrm{tr}\,\left(-A(s)\right)\,\mathrm{d}s\right)\Phi(T,T) \tag{66}$$

and $\Phi(T,T) = \mathbb{I}$. Consequently,

$$|\Phi(t,T)| = \exp\left(\int_t^T \mathrm{tr}\,\left(A(s)\right)\,\mathrm{d}s\right), \tag{67}$$

and

$$p(x_N \mid y,t) = \mathcal{N}\left(x_N \,\middle|\, F\Phi(t,T)y + F\int_t^T \Phi(s,T)b(s)\,\mathrm{d}s, \Sigma + F\int_t^T \Phi(s,T)D\Phi^\top(s,T)\,\mathrm{d}s\,F^\top\right). \tag{68}$$

Let $F(t) = F\Phi(t,T)$; computing the time derivative, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F(t) &= F\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t,T)\\ &= F\left(-\Phi(t,T)A(t)\right)\\ \iff \frac{\mathrm{d}}{\mathrm{d}t}F(t) &= -F(t)A(t),\end{aligned} \tag{69}$$

with end point condition $F(T) = F$. Let now $m(t) = F\int_t^T \Phi(s,T)b(s)\,\mathrm{d}s = \int_t^T F(s)b(s)\,\mathrm{d}s$. By differentiation (utilizing Leibniz' integral rule) we find

$$\frac{\mathrm{d}}{\mathrm{d}t}m(t) = -F(t)b(t), \tag{70}$$

with boundary condition $m(T) = 0$. Let further $\Sigma(t) = \Sigma + F\int_t^T \Phi(s,T)D\Phi^\top(s,T)\,\mathrm{d}s\,F^\top = \Sigma + \int_t^T F(s)DF^\top(s)\,\mathrm{d}s$. We find analogously

$$\frac{\mathrm{d}}{\mathrm{d}t}\Sigma(t) = -F(t)DF^\top(t), \tag{71}$$

with boundary condition $\Sigma(T) = \Sigma$.

Summarizing, we have

$$p(x_N \mid y,t) = \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right), \tag{72}$$

with

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}F(t) = -F(t)A(t) \qquad \text{with} \qquad F(T) = F,\\ &\frac{\mathrm{d}}{\mathrm{d}t}m(t) = -F(t)b(t) \qquad \text{with} \qquad m(T) = 0,\\ &\frac{\mathrm{d}}{\mathrm{d}t}\Sigma(t) = -F(t)DF^\top(t) \quad \text{with} \qquad \Sigma(T) = \Sigma.\end{aligned} \tag{73}$$

A.1.1. The proof for non-invertible F.

In the proof we assumed that F is invertible. However, the solution also holds for general matrices F, which can be shown by plugging in the solution $p(x_N \mid y,t) = \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)$ in the backward Fokker-Planck equation

$$\partial_t p(x_N \mid y,t) = -\sum_{k,l}\partial_{y_k}p(x_N \mid y,t)A_{kl}(t)y_l - \sum_k \partial_{y_k}p(x_N \mid y,t)b_k(t) - \frac{1}{2}\sum_{k,l}\partial_{y_k}\partial_{y_l}p(x_N \mid y,t)D_{kl}. \tag{74}$$

This yields the PDE

$$\begin{aligned}\partial_t \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) = &-\sum_{k,l} \partial_{y_k} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) A_{kl}(t) y_l \\ &- \sum_k \partial_{y_k} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) b_k(t) - \frac{1}{2}\sum_{k,l} \partial_{y_k}\partial_{y_l} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) D_{kl}.\end{aligned} \tag{75}$$

We compute the partial derivatives $\partial_t \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)$, $\partial_{y_k} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)$ and $\partial_{y_k}\partial_{y_l} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)$. First, note the following

$$\partial_\theta \mathcal{N}(x \mid a, A) = \mathcal{N}(x \mid a, A)\left(-h^\top(\partial_\theta x) + h^\top(\partial_\theta a) - \frac{1}{2}\operatorname{tr}\left(A^{-1}\partial_\theta A\right) + \frac{1}{2}h^\top(\partial_\theta A)h\right), \tag{76}$$

with $h = A^{-1}(x-a)$. This yields

$$\begin{aligned}\partial_t \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) = \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)\Big[&h^\top\left(\partial_t F(t)y + \partial_t m(t)\right) \\ &-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(t)\partial_t\Sigma(t)\right) + \frac{1}{2}h^\top\partial_t\Sigma h\Big],\end{aligned} \tag{77}$$

$$\partial_{y_k} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) = \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)\left[h^\top F_{\cdot k}(t)\right], \tag{78}$$

$$\begin{aligned}\partial_{y_k}\partial_{y_l} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) &= \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)\left[h^\top F_{\cdot k}(t)\right]\left[h^\top F_{\cdot l}(t)\right] \\ &\quad + \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)\partial_{y_l}\left\{h^\top F_{\cdot k}(t)\right\} \\ &= \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)\left[h^\top F_{\cdot k}(t)h^\top F_{\cdot l}(t) - L_{\cdot l}^\top(t)\Sigma^{-1}(t)F_{\cdot k}(t)\right],\end{aligned} \tag{79}$$

with $h = \Sigma^{-1}(t)(x_N - F(t)y - m(t))$. Inserting these equations into the KBE yields

$$\begin{aligned}&h^\top\left(\partial_t F(t)y + \partial_t m(t)\right) - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(t)\partial_t\Sigma(t)\right) + \frac{1}{2}h^\top\partial_t\Sigma h \\ &= -\sum_{k,l} h^\top F_{\cdot k,l}(t)A_{kl}(t)y_l - \sum_k h^\top F_{\cdot k}(t)b_k(t) - \frac{1}{2}\sum_{k,l}\left(h^\top F_{\cdot k}(t)h^\top F_{\cdot l}(t) - F_{\cdot l}^\top(t)\Sigma^{-1}(t)F_{\cdot k}(t)\right)D_{kl}\end{aligned} \tag{80}$$

By utilizing vector notation, we have

$$\begin{aligned}&h^\top\left[\partial_t F(t)y + \partial_t m(t)\right] - \frac{1}{2}\operatorname{tr}\left\{\left(\Sigma^{-1}(t) - hh^\top\right)\partial_t\Sigma(t)\right\} \\ &= h^\top\left[-F(t)A(t)y - F(t)b(t)\right] - \frac{1}{2}\operatorname{tr}\left\{\left(\Sigma^{-1}(t) - hh^\top\right)\left[-F^\top(t)DF(t)\right]\right\}\end{aligned} \tag{81}$$

By comparing coefficients, we find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F(t) &= -F(t)A(t) \qquad \text{with} \qquad F(T) = F \\ \frac{\mathrm{d}}{\mathrm{d}t}m(t) &= -F(t)b(t) \qquad \text{with} \qquad m(T) = 0 \\ \frac{\mathrm{d}}{\mathrm{d}t}\Sigma(t) &= -F(t)DF^\top(t) \quad \text{with} \qquad \Sigma(T) = \Sigma,\end{aligned} \tag{82}$$

where we find the end-point conditions by comparing the end-point boundary as

$$\mathcal{N}\left(x_N \mid F(T)y + m(T), \Sigma(T)\right) = \mathcal{N}\left(x_N \mid Fy, \Sigma\right). \tag{83}$$

Reset conditions Starting at the end-point $t = T$, we consider the last observation X_{N-1} at time point t_{N-1}. We have, due to the Markov property,

$$\begin{aligned}\beta(y, t_{N-1}) = p(x_N, x_{N-1} \mid y, t_{N-1}) &= p(x_N \mid y, t_{N-1})p(x_{N-1} \mid x_N, y, t_{N-1}) \\ &= p(x_N \mid y, t_{N-1})p(x_{N-1} \mid y, t_{N-1}) \\ &= \beta(y, t_{N-1}^+)p(x_{N-1} \mid y, t_{N-1}),\end{aligned} \tag{84}$$

where $\beta(y, t_{N-1}^+) := \lim_{h \searrow 0} \beta(y, t_{N-1} + h)$. As we assume a Gaussian observation likelihood, we have, due to the Gaussian properties,

$$\beta(y, t_{N-1}) = \mathcal{N}\left(x_N \mid F(t_{N-1}^+)y + m(t_{N-1}^+), \Sigma(t_{N-1}^+)\right) \mathcal{N}\left(x_{N-1} \mid y, \Sigma_x\right) \tag{85}$$

$$= \mathcal{N}\left(x_{N-1}, x_N \mid F(t_{N-1})y + m(t_{N-1}), \Sigma(t_{N-1})\right) \tag{86}$$

with reset parameters

$$F(t_{N-1}) = \begin{pmatrix} \mathbb{I}_n \\ F(t_{N-1}^+) \end{pmatrix} \in \mathbb{R}^{2n \times n}, \tag{87}$$

$$m(t_{N-1}) = \begin{pmatrix} 0_n \\ m(t_{N-1}^+) \end{pmatrix} \in \mathbb{R}^{2n}, \tag{88}$$

$$M(t_{N-1}) = \begin{pmatrix} \Sigma_x & 0 \\ 0 & \Sigma(t_{N-1}^+) \end{pmatrix} \in \mathbb{R}_{\text{psd}}^{2n \times 2n}, \tag{89}$$

where 0_n denotes the n-dimensional all-zeros vector and $\mathbb{I}_n$ is the n-dimensional identity matrix.

A.1.2. Information filter parameterization

The above backward filter has the property that its support increases upon every incorporated observation, which is computationally disadvantageous. Notice, however, that the contribution of the backward filter to the drift of the posterior SDE is fixed in size; for convenience, we re-state the SDE:

$$\mathrm{d}Y(t) = \left(f(Y(t), t) + D(Z(t))\nabla \log \beta(Y(t), t)\right) \mathrm{d}t + Q(Z(t))\mathrm{d}W(t),$$

where we notice that

$$\nabla \log \beta(y, t) = -\frac{1}{2}\nabla (x(t) - F(t)y(t) - m(t))^\top \Sigma^{-1}(t)(x(t) - F(t)y(t) - m(t)) \tag{90}$$

$$= F(t)^\top \Sigma^{-1}(t)(x(t) - m(t)) - F(t)^\top \Sigma^{-1}(t)F(t)y(t) \tag{91}$$

where the gradient was taken with respect to y. Defining

$$\begin{aligned} \nu(t) &:= F(t)^\top \Sigma^{-1}(t)(x(t) - m(t)), \\ M(t) &:= F(t)^\top \Sigma^{-1}(t)F(t), \end{aligned} \tag{92}$$

one can compute straightforwardly the respective time derivatives, where we use the notation $\dot{f} = \frac{\mathrm{d}}{\mathrm{d}t} f$ for conciseness,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} M(t) &= \dot{F}(t)^\top \Sigma^{-1}(t)F(t) + F(t)^\top \dot{\Sigma}^{-1}(t)F(t) + F(t)^\top \Sigma^{-1}(t)\dot{F}(t) \\ &= -A(t)^\top M(t) + F(t)^\top \dot{\Sigma}^{-1}(t)F(t) - M(t)A(t) \\ &= -A(t)^\top M(t) - F(t)^\top \Sigma^{-1}(t)\dot{\Sigma}(t)\Sigma^{-1}(t)F(t) - M(t)A(t) \\ &= -A(t)^\top M(t) + F(t)^\top \Sigma^{-1}(t)F(t)DF(t)^\top \Sigma^{-1}(t)F(t) - M(t)A(t) \\ &= -A(t)^\top M(t) + M(t)DM(t) - M(t)A(t), \end{aligned} \tag{93}$$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \nu(t) &= \dot{F}(t)^\top \Sigma^{-1}(t)(x(t) - m(t)) + F(t)^\top \dot{\Sigma}^{-1}(t)(x(t) - m(t)) - F(t)^\top \Sigma^{-1}(t)\dot{m}(t) \\ &= -A(t)^\top \nu(t) + M(t)D\nu(t) + M(t)b(t). \end{aligned} \tag{94}$$

The reset conditions for the information filter given in the main paper follow directly by comparing Eq. (87) and Eq. (92).

A.2. Posterior SSDE Initial Condition

The posterior initial distribution $p(y_0 \mid x_{[1,N]})$ is found as

$$\begin{aligned} p(y_0 \mid x_{[1,N]}, -) &\propto \beta(y_0, 0) p(y_0 \mid \mu_0, \Sigma_0) \\ &\propto \exp\left\{-c(0) - \frac{1}{2} y_0^\top I(0) y_0 + a(0)^\top y_0\right\} \mathcal{N}(y_0 \mid \mu_0, \Sigma_0) \\ &\propto \mathcal{N}\left(y_0 \mid \bar{\mu}, \bar{\Sigma}\right) \end{aligned} \tag{95}$$

with

$$\bar{\mu} = \bar{\Sigma}(\Sigma_0^{-1}\mu_0 + a(0)), \quad \bar{\Sigma} = \left(\Sigma_0^{-1} + I(0)\right)^{-1}. \tag{96}$$

B. Derivation of the Wonham-type filter

In the following, we derive the filtering density $p_f(z,t)$. We follow the treatment in (Del Moral & Penev, 2017), but see also, e.g., (Van Handel, 2007). For convenience, we re-iterate the time evolution of the hybrid system as

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} p(z,t) &= \sum_{z' \in \mathcal{Z}} \Lambda(z', z, t) p(z', t) \quad \forall z' \in \mathcal{Z}, \\ \mathrm{d}Y(t) &= f(Y(t), Z(t))\mathrm{d}t + Q(Z(t))\mathrm{d}W(t), \end{aligned} \tag{97}$$

where the dispersion $Q(z)Q(z)^\top = D(z)$. We are interested in the conditional path measure

$$\begin{aligned} \mathsf{P}\left(Z_{[0,t]} \in \mathrm{d}z_{[0,t]} \mid y_{[0,t]}\right) &\propto \mathsf{P}\left((Z_{[0,t]}, Y_{[0,t]}) \in \mathrm{d}(z_{[0,t]}, y_{[0,t]})\right) \\ &= \mathsf{P}\left(Y_{[0,t]} \in \mathrm{d}y_{[0,t]} \mid z_{[0,t]}\right) \mathsf{P}\left(Z_{[0,t]} \in \mathrm{d}z_{[0,t]}\right) \\ &= G(z_{[0,t]}, y_{[0,t]}) \mathsf{P}(W_{[0,t]} \in \mathrm{d}y_{[0,t]}) \mathsf{P}\left(Z_{[0,t]} \in \mathrm{d}z_{[0,t]}\right) \end{aligned} \tag{98}$$

where the last equality is due to Girsanov's theorem (Øksendal, 2003) with the Radon-Nikodym derivative

$$G(z_{[0,t]}, y_{[0,t]}) := \frac{\mathrm{d}\mathsf{P}_Z}{\mathrm{d}\mathsf{P}_W} = \exp\left\{F(t)\right\}, \tag{99}$$

where we used the subscripts to indicate the conditional posterior and the Brownian motion measures and defined the shorthand

$$F(t) = \int_0^t \mathrm{d}F(t) := \int_0^t f^\top(z(s), y(s)) D(z(s))^{-1} \mathrm{d}y(s) - \frac{1}{2}\int_0^t f^\top(z(s), y(s)) D(z(s))^{-1} f(z(s), y(s)) \mathrm{d}s, \tag{100}$$

the posterior measure can be expressed via the Kallianpur-Striebel formula,

$$\mathsf{P}\left(Z_{[0,t]} \in \mathrm{d}z_{[0,t]} \mid y_{[0,t]}\right) = \frac{G(z_{[0,t]}, y_{[0,t]}) \mathsf{P}(Z_{[0,t]} \in \mathrm{d}z_{[0,t]})}{\int G(z'_{[0,t]}, y_{[0,t]}) \mathsf{P}(Z_{[0,t]} \in \mathrm{d}z'_{[0,t]})}. \tag{101}$$

Our quantity of interest follows as an expectation,

$$p_f(z,t) = \mathsf{E}\left[\mathbb{1}(Z(t) = z)\right] = \int \mathbb{1}(Z(t) = z) \mathsf{P}\left(Z_{[0,t]} \in \mathrm{d}z_{[0,t]} \mid y_{[0,t]}\right), \tag{102}$$

where we use the subscript f to indicate the *filtering* distribution. Using Eq. (101), we can restate this in terms of the prior measure,

$$p_f(z,t) = \frac{\int \mathbb{1}(Z(t) = z) G(z_{[0,t]}, y_{[0,t]}) \mathsf{P}(Z_{[0,t]} \in \mathrm{d}z_{[0,t]})}{\int G(z'_{[0,t]}, y_{[0,t]}) \mathsf{P}(Z_{[0,t]} \in \mathrm{d}z'_{[0,t]})} = \frac{\mathsf{E}\left[\mathbb{1}(Z(t) = z) G(z_{[0,t]}, y_{[0,t]})\right]}{\mathsf{E}\left[G(z_{[0,t]}, y_{[0,t]})\right]}. \tag{103}$$

Here and in the following, the expectation operator $\mathsf{E}[\cdot]$ refers to the expectation over the prior measure $P(Z_{[0,t]} \in \mathrm{d}z_{[0,t]})$. Using the Itô calculus chain rule, we compute

$$\begin{aligned}\mathrm{d}G(z_{[0,t]}, y_{[0,t]}) &= \exp\{F(t)\}\mathrm{d}F(t) + \frac{1}{2}\exp\{F(t)\}\mathrm{d}F(t)\mathrm{d}F(t) \\ &= G(z_{[0,t]}, y_{[0,t]})\mathrm{d}F(t) + \frac{1}{2}G(z_{[0,t]}, y_{[0,t]})\mathrm{d}F(t)\mathrm{d}F(t) \\ &= G(z_{[0,t]}, y_{[0,t]})f(z(t), y(t))^\top D(z(t))^{-1}\mathrm{d}y(t).\end{aligned} \tag{104}$$

and find for the unnormalized quantity

$$\tilde{p}_f(z,t) := \mathsf{E}\left[\mathbb{1}(Z(t) = z)G(z_{[0,t]}, y_{[0,t]})\right] \tag{105}$$

the expression

$$\begin{aligned}\mathrm{d}\tilde{p}_f(z,t) &= \mathsf{E}\left[\mathrm{d}\big(\mathbb{1}(Z(t) = z)G(z_{[0,t]}, y_{[0,t]})\big)\right] \\ &= \mathsf{E}\left[\mathrm{d}\mathbb{1}(Z(t) = z)G(z_{[0,t]}, y_{[0,t]}) + \mathbb{1}(Z(t) = z)\mathrm{d}G(z_{[0,t]}, y_{[0,t]}) + \mathrm{d}\mathbb{1}(Z(t) = z)\mathrm{d}G(z_{[0,t]}, y_{[0,t]})\right].\end{aligned} \tag{106}$$

Now, noticing that

$$\mathsf{E}\left[\mathrm{d}\mathbb{1}(Z(t) = z)\right] = \mathsf{E}\left[\sum_{z' \in \mathcal{Z}} \Lambda(z, z', t)\mathbb{1}(Z(t) = z')\mathrm{d}t\right] \tag{107}$$

and accordingly, acknowledging that $\mathrm{d}t \cdot \mathrm{d}y(t) = 0$,

$$\begin{aligned}\mathrm{d}\tilde{p}_f(z,t) &= \mathsf{E}\left[\mathrm{d}\mathbb{1}(Z(t) = z)G(z_{[0,t]}, y_{[0,t]}) + \mathbb{1}(Z(t) = z)\mathrm{d}G(z_{[0,t]}, y_{[0,t]})\right] \\ &= \sum_{z' \in \mathcal{Z}} \Lambda(z', z, t)\tilde{p}_f(z', t)\mathrm{d}t + \mathsf{E}\left[\mathbb{1}(Z(t) = z)f^\top(y(t), z(t))D^{-1}(z(t))\right]\mathrm{d}y(t) \\ &= \sum_{z' \in \mathcal{Z}} \Lambda(z', z, t)\tilde{p}_f(z', t)\mathrm{d}t + \tilde{p}_f(z,t)f^\top(y(t), z(t))D^{-1}(z(t))\mathrm{d}y(t).\end{aligned} \tag{108}$$

Equation (108) is called the Zakai equation in the literature. To derive the dynamics of the desired respective normalized quantity Eq. (101), further consider its denominator and notice that

$$\mathrm{d}\log \mathsf{E}[G] = \frac{\mathrm{d}\mathsf{E}[G]}{\mathsf{E}[G]} - \frac{1}{2}\frac{\mathrm{tr}\left\{\mathrm{d}\mathsf{E}[G]\,\mathrm{d}\mathsf{E}[G]^\top\right\}}{\mathsf{E}[G]^2}, \tag{109}$$

where we suppressed the arguments for conciseness. The quantity $\mathrm{d}\,\mathsf{E}[G]$ is precisely given by the Zakai equation (108) upon replacing the indicator by a constant, $\mathbb{1}(Z(t) = z) \to 1$,

$$\mathrm{d}\mathsf{E}[G] = \mathsf{E}\left[Gf^\top D^{-1}\right]\mathrm{d}y(t). \tag{110}$$

Inserting this into Eq. (109), as $\mathrm{d}y(t)\mathrm{d}y(t)^\top = D(z(t))\mathrm{d}t$, one finds

$$\mathrm{d}\log \mathsf{E}[G] = \frac{\mathsf{E}\left[Gf^\top D^{-1}\right]}{\mathsf{E}[G]}\mathrm{d}y(z) - \frac{1}{2}\frac{\mathsf{E}\left[Gf^\top D^{-1}\right]D^{-1}\mathsf{E}\left[D^{-1}f\right]}{\mathsf{E}[G]^2}\mathrm{d}t \tag{111}$$

where the terms on the right hand side are of the exact same form as Eq. (103). Consequently,

$$\mathsf{E}[G] = \exp\left\{\int_0^t \underbrace{\frac{\mathsf{E}\left[Gf^\top D^{-1}\right]}{\mathsf{E}[G]}}_{=:\varpi^\top}\mathrm{d}y(s) - \frac{1}{2}\int_0^t \underbrace{\frac{\mathsf{E}\left[Gf^\top D^{-1}\right]D^{-1}\mathsf{E}\left[D^{-1}fG\right]}{\mathsf{E}[G]^2}}_{\varpi^\top D^{-1}\varpi}\mathrm{d}s\right\}, \tag{112}$$

where, for clarity, we restate with arguments:

$$\varpi(t) = \frac{\mathsf{E}\left[G(z_{[0,t]}, y_{[0,t]})f(y(t), z(t))^\top D(z(t))^{-1}\right]}{\mathsf{E}\left[G(z_{[0,t]}, y_{[0,t]})\right]}. \tag{113}$$

Inserting this into the original Eq. (103), we arrive at

$$p_f(z,t) = \mathsf{E}\left[\mathbb{1}(Z(t)=z)\exp\left\{\int_0^t \left(f^\top(z(s),y(s))D(z(s))^{-1} - \varpi\right)(\mathrm{d}y(s)-\varpi(s)) \right.\right. \\ \left.\left. -\frac{1}{2}\int_0^t \left(f^\top(z(s),y(s)) - \varpi(s)\right) D(z(s))^{-1}\left(f(z(s),y(s)) - \varpi(s)\right)\mathrm{d}s\right\}\right]. \quad (114)$$

Repeating with this quantity the same derivation steps as for the Zakai equation (108) straightforwardly yields the Kushner-Stratonovich equation

$$\mathrm{d}p_f(z,t) = \sum_{z'\in\mathcal{Z}} \Lambda(z',z,t)p_f(z,t)\mathrm{d}t + p_f(z,t)\left(f(y(t),z(t)-\bar{f}(y(t))\right)^\top D(z(t))^{-1}\left(\mathrm{d}y(t)-\bar{f}(y(t))\right), \quad (115)$$

where $\bar{f}(y(t)) := \sum_{z\in\mathcal{Z}} f(y(t),z(t))p_f(z,t)$.

C. Bayesian Parameter estimation

We go through the derivations in the same order as presented in the main paper.

Initial Conditions The Dirichlet prior in the initial MJP state distribution π_{z_0},

$$p(\pi_{z_0}) = \mathrm{Dir}(\pi_{z_0} \mid \alpha_{z_0}), \quad (116)$$

with $\alpha_{z_0} \in \mathbb{R}_{>0}^{|\mathcal{Z}|}$, results in the usual conjugate update,

$$\begin{aligned} p(\pi_{z_0} \mid z_{[0,T]}, -) &\propto \mathrm{Cat}(z(0) \mid \pi_{z_0})\,\mathrm{Dir}(\pi_{z_0} \mid \alpha_{z_0}) \\ &\propto \mathrm{Dir}(\pi_{z_0} \mid \alpha_{z_0} + \delta_{z(0)}). \end{aligned} \quad (117)$$

The SSDE initial distribution parameters μ_0, Σ_0 are specified via a NIW prior,

$$\mathcal{NIW}(\mu_0,\Sigma_0 \mid \eta,\lambda,\Psi,\kappa) = \mathcal{N}\left(\mu_0 \,\middle|\, \eta, \frac{\Sigma_0}{\lambda}\right)\mathcal{IW}(\Sigma_0 \mid \Psi,\kappa). \quad (118)$$

Recall that the SSDE initial value $y(0) = y_0$ is Gaussian distributed,

$$p(y_0) = \mathcal{N}(y_0 \mid \mu_0, \Sigma_0). \quad (119)$$

Accordingly, the Bayesian posterior

$$\begin{aligned} p(\mu_0,\Sigma_0 \mid y_0, x_{[1,N]}, -) &\propto p(y_0 \mid x_{[1,N]},\mu_0,\Sigma_0)p(\mu_0,\Sigma_0) \\ &\propto p(y_0 \mid \mu_0,\Sigma_0)p(\mu_0,\Sigma_0) \\ &\propto \mathcal{NIW}\left(\mu_0,\Sigma_0 \,\middle|\, \tilde{\eta},\tilde{\lambda},\tilde{\Psi},\tilde{\kappa}\right) \end{aligned} \quad (120)$$

with the updates as in the main paper.

MJP rate parameters Utilizing the transition counts and the cumulative sojourn times as defined in the main paper,

$$\begin{aligned} N_{zz'} &= \sum_{k=0}^{K-1} \mathbb{1}(z_k = z \wedge z_{k+1} = z'), \\ T_z &= \sum_{k=0}^{K} \mathbb{1}(z_k = z)\tau_k, \end{aligned} \quad (121)$$

we can explicate the path likelihood $p(z_{[0,T]} \mid \Lambda_{zz'})$ and compute

$$\begin{aligned} p(\Lambda_{zz'} \mid z_{[0,T]}) &\propto p(z_{[0,T]}, - \mid \Lambda_{zz'}) p(\Lambda_{zz'}) \\ &\propto \prod_{k=0}^{K-1} \left\{\Lambda_{z_k} e^{-\Lambda_{z_k}\tau_k}\right\}^{\mathbb{1}(z_k=z)} \left\{\frac{\Lambda_{z_k z_{k+1}}}{\Lambda_{z_k}}\right\}^{\mathbb{1}(z_k=z \wedge z_{k+1}=z')} \left\{e^{-\Lambda_{z_K}\tau_K}\right\}^{\mathbb{1}(z_K=z)} p(\Lambda_{zz'}) \\ &\propto \left\{\Lambda_{zz'}\right\}^{\sum_{k=0}^{K-1}\mathbb{1}(z_k=z\wedge z_{k+1}=z')} e^{-\Lambda_z \sum_{k=0}^{K}\mathbb{1}(z_k=z)\tau_k} p(\Lambda_{zz'}) \\ &\propto \left\{\Lambda_{zz'}\right\}^{\sum_{k=0}^{K-1}\mathbb{1}(z_k=z\wedge z_{k+1}=z')} e^{-\Lambda_{zz'} \sum_{k=0}^{K}\mathbb{1}(z_k=z)\tau_k} p(\Lambda_{zz'}) \\ &\propto \mathrm{Gam}(\Lambda_{zz'} \mid s + N_{zz'}, r + T_z). \end{aligned} \tag{122}$$

SDE Drift Parameters We impose a Matrix-Normal (MN) prior over the drift parameters $\Gamma_m := [A_m, b_m]$,

$$\begin{aligned} p(\Gamma_m) &= \mathcal{MN}(\Gamma_m \mid M_m, D_m, K_m) \\ &= (2\pi)^{\frac{nm}{2}} |D_m|^{-\frac{n}{2}} |K_m|^{-\frac{m}{2}} \exp\left\{-\frac{1}{2}\,\mathrm{tr}\,\left((\Gamma_m - M_m)^\top D_m^{-1}(\Gamma_m - M_m)K_m^{-1}\right)\right\}. \end{aligned} \tag{123}$$

Note that we atypically use the subscript m for 'mode' here to avoid visual confusion with the letter z. To update the drift parameters Γ_m for all modes $m \in \mathcal{Z}$, we need to compute

$$p(\Gamma_m \mid z_{[0,T]}, y_{[0,T]}, x_{[1,N]}, -) \propto G(z_{[0,T]}, y_{[0,T]} \mid \Gamma_m) p(\Gamma_m). \tag{124}$$

The 'likelihood term' can be evaluated approximately by inserting the simulated paths. Now note that for the mode m, only the subintervals of $z_{[0,T]}$ contribute in which $z(t) = m$. More concretely, an MJP realization $z_{[0,T]}$ can be specified via the *jump times* $\{j_k\}_{k=1,\dots,K}$,

$$j_{k+1} = \inf_t \left(t \geq j_k \mid z(t) \neq z(j_k)\right) \tag{125}$$

and the *state sequence* $\{z_k\}_{k=1,\dots,K}$, $z_k \in \mathcal{Z}\ \forall k$:

$$z(t) = z_i,\ i = \sup_k (j_k < t). \tag{126}$$

This allows us to write

$$p(\Gamma_m \mid z_{[0,T]}, y_{[0,T]}, x_{[1,N]}, -) \propto G(z_{[0,T]}, y_{[0,T]} \mid \Gamma_m) p(\Gamma_m) \tag{127}$$

$$\begin{aligned} &= \exp\left\{\sum_{i\,:\,z(j_i)=m} \int_{j_i}^{j_{i+1}} f^\top(z(s), y(s)) D^{-1}(z(s)) \mathrm{d}y(s) \right. \\ &\qquad \left. -\frac{1}{2}\int_{j_i}^{j_{i+1}} f^\top(z(s), y(s)) D^{-1}(z(s)) f(z(s), y(s)) \mathrm{d}s\right\} p(\Gamma_m). \end{aligned} \tag{128}$$

Putting this sum over intervals aside for a moment for better readability, we find upon inserting a simulated SDE-path $y_{[0,T]}$

$$\exp\left\{\int_{t_0}^{t_1} f^\top(m, y(s)) D^{-1}(m) \mathrm{d}y(s) - \frac{1}{2}\int_{t_0}^{t_1} f^\top(m, y(s)) D^{-1}(m) f(m, y(s)) \mathrm{d}s\right\} \tag{129}$$

$$\approx \exp\left\{\sum_{l=1}^{L} f^\top(m, y_l) D^{-1}(m) \Delta y_l - \frac{1}{2}\sum_{l=1}^{L} f^\top(m, y_l) D^{-1}(m) f(m, y_l) h\right\}, \tag{130}$$

where h is time simulation time-step, $s_l = s_{l-1} + h$, the interval boundaries $s_1 = t_0$, $s_L = t_1$, and $\Delta y_l := y(s_l + h) - y(s_l)$ the difference of two successive points of the trajectory.

Inserting the drift $f(m, y_l) = \Gamma_m \bar{y}_l$, where $\bar{y}_l = \left[y_l^\top, 1_n^\top\right]^\top$, and writing for conciseness $D^{-1}(m) = D_m^{-1}$ we find

$$\exp\left\{\sum_{l=1}^{L} f^\top(m, y_l) D_m^{-1} \Delta y_l - \frac{1}{2}\sum_{l=1}^{L} f^\top(m, y_l) D_m^{-1} f(m, y_l) h\right\} \tag{131}$$

$$= \exp\left\{\sum_{l=1}^{L} \bar{y}_l^\top \Gamma_m^\top D_m^{-1} \Delta y_l - \frac{1}{2}\sum_{l=1}^{L} \bar{y}_l^\top \Gamma_m^\top D_m^{-1} \Gamma_m \bar{y}_l h\right\} \tag{132}$$

$$= \exp\left\{\sum_{l=1}^{L} \sqrt{h}\bar{y}_l^\top \Gamma_m^\top D_m^{-1} \frac{\Delta y_l}{\sqrt{h}} - \frac{1}{2}\sum_{l=1}^{L} \sqrt{h}\bar{y}_l^\top \Gamma_m^\top D_m^{-1} \Gamma_m \bar{y}_l \sqrt{h}\right\} \tag{133}$$

$$= \exp\left\{-\frac{1}{2}\sum_{l=1}^{L} \left(\frac{\Delta y_l}{\sqrt{h}} - \Gamma_m \bar{y}_l \sqrt{h}\right)^\top D_m^{-1} \left(\frac{\Delta y_l}{\sqrt{h}} - \Gamma_m \bar{y}_l \sqrt{h}\right) + \frac{1}{2}\sum_{l=1}^{L} \Delta y_l^\top D_m^{-1} \Delta y_l \frac{1}{h}\right\}. \tag{134}$$

We may now omit the last term on the right hand side, as it is independent of Γ_m. Note, however, that together with the measure of the Brownian motion in Eq. (25), we have (Del Moral & Penev, 2017)

$$\mathsf{P}\big(W_{[0,T]} \in \mathrm{d}w_{[0,T]}\big) \exp\left\{\frac{1}{2}\sum_{l=1}^{L} \Delta y_l^\top D_m^{-1} \Delta y_l \frac{1}{h}\right\} = |2\pi h D_m|^{-\frac{1}{2}} \prod_{l=1}^{L} \Delta y_l. \tag{135}$$

Hence, the above is equivalent to approximating the Radon-Nikodym derivative $G(z_{[0,T]}, y_{[0,T]})$ via the product of L Gaussian transition distributions. Making use of the trace function and defining the joint observation vectors

$$\Delta Y := \left[\frac{\Delta y_{s_1}}{\sqrt{h}}, \cdots, \frac{\Delta y_{s_L}}{\sqrt{h}}\right] \in \mathbb{R}^{n \times L}, \tag{136}$$

$$\bar{Y} := \left[\bar{y}_{s_1}\sqrt{h}, \cdots, \bar{y}_{s_1}\sqrt{h}\right] \in \mathbb{R}^{n+1 \times L}, \tag{137}$$

we arrive at

$$\exp\left\{-\frac{1}{2}\sum_{l=1}^{L} \left(\frac{\Delta y_l}{\sqrt{h}} - \Gamma_m \bar{y}_l \sqrt{h}\right)^\top D_m^{-1} \left(\frac{\Delta y_l}{\sqrt{h}} - \Gamma_m \bar{y}_l \sqrt{h}\right)\right\} \tag{138}$$

$$= \exp\left\{-\frac{1}{2}\,\mathrm{tr}\,\left(\left(\Delta Y - \Gamma_m \bar{Y}\right)^\top D_m^{-1} \left(\Delta Y - \Gamma_m \bar{Y}\right) \mathbb{1}_{L \times L}\right)\right\} \tag{139}$$

and notice that this expression corresponds to an (un-normalized) MN distribution Accordingly, still only considering a single jump interval $[j_{m_i}, j_{m_i+1}]$, we can write

$$p(\Gamma_m \mid z_{[0,T]}, y_{[0,T]}, x_{[1,N]}, -) \propto \mathcal{MN}\left(\Delta Y \mid \Gamma_m \bar{Y}, D_m, \mathbb{1}_{L \times L}\right) p(\Gamma_m) \tag{140}$$

$$= \mathcal{MN}\left(\Delta Y \mid \Gamma_m \bar{Y}, D_m, \mathbb{1}_{L \times L}\right) \mathcal{MN}\left(\Gamma_m \mid M_m, D_m, K_m\right) \tag{141}$$

It is known (Fox et al., 2008) that the Matrix-Normal distribution is a conjugate prior to the Matrix-Normal likelihood in the above form; consequently, the sought-after posterior is itself Matrix-Normal

$$p(\Gamma_m, z_{[0,T]}, y_{[0,T]}, x_{[1,N]}, -) = \mathcal{MN}\left(\Gamma_m \mid \tilde{M}_m, D_m, \tilde{K}_m\right) \tag{142}$$

with posterior hyperparameters

$$\begin{aligned} \tilde{K}_m &= \bar{Y}\bar{Y}^\top + K_m, \\ \tilde{M}_m &= (\Delta Y \bar{Y}^\top + M_m K_m)\tilde{K}_m^{-1}. \end{aligned} \tag{143}$$

Summation over all intervals with $z(j_{m_i}) = m$ is straightforward. Importantly, the above derivation also holds for adaptive step-sizes, $s_l = s_{l-1} + h_{l-1}$.

Observation Covariance We define the prior

$$p(\Sigma_x \mid x_{[1,N]}) = \mathcal{IW}(\Sigma_x \mid \Psi_x, \lambda_x). \tag{144}$$

With the Gaussian observations

$$X_i \sim \mathcal{N}(x_i \mid y_i, \Sigma_{\text{obs}}),$$

the standard result is

$$\begin{aligned} p(\Sigma_{\text{obs}} \mid x_{[1,N]}) &\propto p(x_{[1,N]} \mid \Sigma_{\text{obs}}) p(\Sigma_{\text{obs}}) && (145)\\ &\propto \mathcal{IW}(\Sigma_{\text{obs}} \mid \tilde{\Psi}_{\text{obs}}, \tilde{\lambda}_{\text{obs}}) && (146)\end{aligned}$$

with the updated hyperparameters as in the main paper.

D. Experimental Details

D.1. Hyperparameter Settings

We initialize all distribution hyperparameters, cf. Appendix C, empirically. To this end, we run k-means with the number of modes $|\mathcal{Z}|$ on the data and obtain empirical cluster means μ_z and covariances Σ_z. In the following, we denote by $z(t_i)$, we mean the k-means cluster assignment of observation i at time point t_i. We go through the settings in order of appearance in the main paper.

Initial Conditions The MJP initial Dirichlet hyperparameters are

$$\alpha_z = 1 + \delta(z(t_1)).$$

The SDE initial NIW hyperparameters

$$\eta = \frac{\sum_z \mu_z}{|Z|}, \quad \lambda = 1, \quad \Psi = \frac{\sum_z \Sigma_z}{|Z|} * 0.1, \quad \kappa = n + 2.$$

Note that, as done, e.g., in (Fox et al., 2008), we use a heuristic downscaling of the empirical covariances as they contain contributions by the measurement noise, the process covariance as well as the drift. Also, $\kappa = n + 2$ is the smallest scaling parameter that makes the IW distribution well defined.

MJP rates We compute the number of total observed transitions in the k-means trajectory, N_{trans} and set

$$s = N_{\text{trans}}, \quad r = 1.$$

SDE Drift Parameters We compute

$$\begin{aligned} \hat{A}_z &= \sum_{i=1}^{N} \mathbb{1}(z(t_i) = z) \frac{x_{i+1} - x_i}{x_{i+1} - t_i}, \\ \hat{b}_z &= -\hat{A}_z \mu_z, \end{aligned} \tag{147}$$

where the latter is because the set point for a linear system is found via

$$f(y) = Ay + b = A(y + A^{-1}b), \tag{148}$$

hence $f(y) = 0$ if $y = -A^{-1}b$, and we demand

$$\mu_z = -A_z^{-1} b \Rightarrow b_z = -A_z \mu_z. \tag{149}$$

With this, the MN hyperparameters

$$\begin{aligned} M_z &= [A_z, b_z] \\ K_z &= \mathbb{1}_{n+1}. \end{aligned} \tag{150}$$

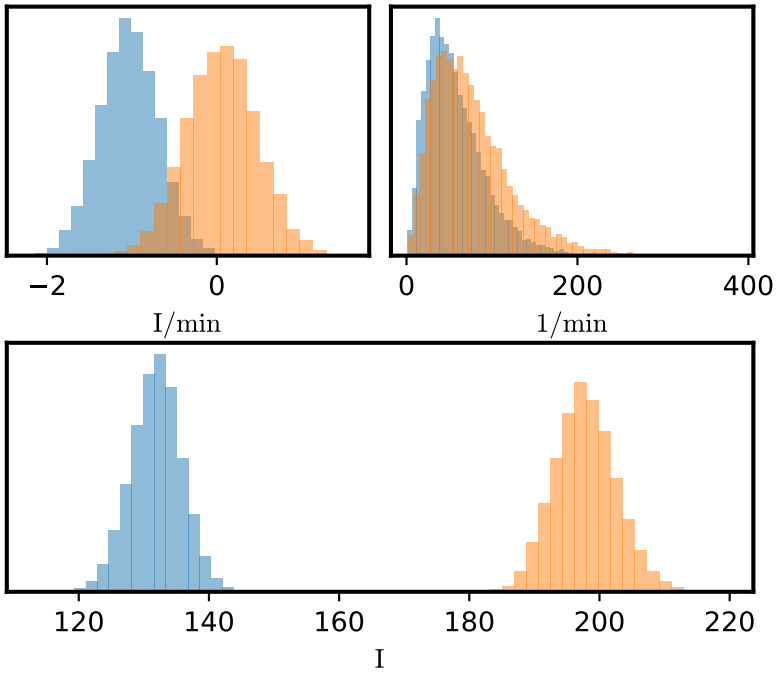

Figure 5. Bayesian parameter estimates. Top left: drift parameters A_z. Top right: rates $\Lambda_{zz'}$. Bottom: drift parameters b_z. I: fluorescence intensity (a.u.).

SDE Dispersion We set

$$\Psi_{D_z} = \Sigma_z * 0.1 \quad \lambda_{D_z} = n + 2, \tag{151}$$

with a heuristic downscaling as above.

Observation Covariance Lastly,

$$\Psi_x = \Sigma_z * 0.5 \quad \lambda_{D_z} = n + 2. \tag{152}$$

D.2. Inference of Gene-Switching Dynamics

We provide the posterior parameter distributions in Fig. 5. We initialized the variational method up for comparison in the same way, but set the parameters directly instead of distribution hyperparameters, as this method does not take a fully Bayesian approach, but rather works with point estimates.