Moment-Based Variational Inference for Markov Jump Processes

Christian Wildner[1]   **Heinz Koeppl**[1]

Abstract

We propose moment-based variational inference as a flexible framework for approximate smoothing of latent Markov jump processes. The main ingredient of our approach is to partition the set of all transitions of the latent process into classes. This allows to express the Kullback-Leibler divergence between the approximate and the exact posterior process in terms of a set of moment functions that arise naturally from the chosen partition. To illustrate possible choices of the partition, we consider special classes of jump processes that frequently occur in applications. We then extend the results to parameter inference and demonstrate the method on several examples.

1. Introduction

Markov jump processes (MJPs) are popular modelling tools in a number of domains including physics, biology, mathematical finance and chemistry (Anderson, 1991). In many cases, these models are used for planning and prediction. To achieve this, the model parameters have to be calibrated from sparse and noisy measurements. Mathematically, this leads to the problem of inference for stochastic processes. Unfortunately, inference in such scenarios is notoriously difficult because parameter inference requires estimation of the latent Markov chain given the observations as an intermediate step. This is especially true in scenarios with unbounded state-spaces such as population models. In recent years, population models have become popular in systems and synthetic biology to describe the intrinsic stochasticity of bio-chemical reactions in a cellular environment (Anderson & Kurtz, 2015).

Algorithms for latent state estimation are often based on the sequential Monte Carlo framework (Doucet & Johansen, 2011) or specialized MCMC approaches (Rao & Teh, 2013). Alternatively, if the state space is not too large, one can also apply a continuous version of the classical forward-backward algorithm for hidden Markov models. In the context of parameter learning, many approaches are likelihood-based and apply a form of the EM-algorithm (Bladt & Sørensen, 2005; Metzner et al., 2007; Liu et al., 2015). Bayesian approaches to joint estimation of latent states and parameters usually focus on the particle Markov chain Monte Carlo framework which has inspired many applications (Andrieu et al., 2010; Golightly & Wilkinson, 2011; Frigola et al., 2013).

While variational inference is ubiquituos in machine learning, most applications focus on probabilistic graphical models (Blei et al., 2017). An important contribution in the domain of continuous time dynamic models is the work of Archambeau et al. (2007) who developed a Gaussian approximation for stochastic differential equations. Opper & Sanguinetti (2008) proposed the first variational method for MJPs on the process level. The core idea of their approach is to approximate a coupled multi-component process by a product of independent component processes. This idea was adapted by Cohn et al. (2010) to continuous time Bayesian networks. More recently, Zhang et al. (2017) combined uniformization with the classical variational method for graphical models.

In this work, we propose a general procedure to derive variational inference algorithms for arbitrary MJPs on a countable state space. While based on the same divergence functional as the approach in (Opper & Sanguinetti, 2008), our approximation is conceptually different. The key step in our approach is to partition the state of all transitions into a number of predefined classes. We then express the divergence function in terms of natural moment functions that are induced by the partition. This allows to approximate the inference problem by an optimal control problem with a complexity controlled by the chosen partition for the transitions. Another difference to the mean-field approach is that we construct our variational family as a modification of the prior process, thus overcoming issues with absolute continuity that can arise from a product approximation.

The remaining part of the paper is organized as follows. Section 2 summarizes background material on MJPs. In Section 3 we construct our variational smoothing algorithm and discuss a few special process classes. Section 4 discusses an extension to latent parameter inference. Finally, in Section 5 we demonstrate the methods on several examples.

---

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

2. Background

2.1. Markov Jump Processes

We consider a Markov jump process $X(t)$ on a countable state space $\mathcal{X}$ over a finite interval $[0, T]$. As is well-known, such a process is fully specified by an initial measure $p_0 : \mathcal{X} \to [0, 1]$ and a transition function $Q : \mathcal{X} \times \mathcal{X} \times [0, T] \to [0, \infty)$. The transition function defines the infinitesimal transition probabilities in the sense that for $y \neq x$ we have

$$P(X(t + \Delta t) = y \mid X(t) = x) = Q(x, y, t)\Delta t + o(\Delta t)$$

For any $t > 0$ the marginal probability distribution $p(x, t) = P(X(t) = x)$ satisfies the master equation

$$\frac{d}{dt}p(x, t) = \sum_{y \neq x} Q(y, x, t)p(y, t) - \sum_{y \neq x} Q(x, y, t)p(x, t) \, .$$

Let $g : \mathcal{X} \to \mathbb{R}$ be a function with finite expectation $\mathsf{E}[g(X(t))] < \infty$ for $t \in [0, T]$. Then the expectation obeys a differential equation of the form

$$\frac{d}{dt}\mathsf{E}[g(X(t))] = \sum_{y} \mathsf{E}\left[(g(y) - g(X(t)) Q(X(t), y, t)\right] \, , \tag{1}$$

which can usually not be expressed in terms of $\mathsf{E}[g(X(t))]$ alone (Ethier & Kurtz, 2005).

2.2. Path Divergence

For two probability measures μ and ν on a common probability space, the Kullback-Leibler (KL) divergence from μ to ν is defined as

$$D[\mu \,||\, \nu] = \begin{cases} \int \log\left(\frac{d\mu}{d\nu}\right) d\mu & \mu \ll \nu, \\ \infty & \text{otherwise} \, , \end{cases}$$

where $\frac{d\mu}{d\nu}$ is the Radon-Nikodym derivative of measures and $\mu \ll \nu$ refers to μ being absolutely continuous with respect to ν.

Consider two MJPs X and Z on $\mathcal{X}$ over $[0, T]$ with possibly time-dependent transition functions Q^X, Q^Z that share the same initial distribution. Let P^X, P^Z denote the measures over the sample path space induced by X and Z. Then the divergence on the level of sample paths is given by

$$D\left[P^X \,||\, P^Z\right] = \int_0^T \sum_{x} p^X(x, t) \sum_{y \neq x} \left[Q^Z(x, y, t) \right.$$
$$\left. - Q^X(x, y, t) - Q^X(x, y, t) \log\left(\frac{Q^X(x, y, t)}{Q^Z(x, y, t)}\right) \right] dt \, , \tag{2}$$

where and $p^X(\cdot, t)$ refers to the marginal distribution of $X(t)$. For this expression to be finite it is required that

$Q^X(x, y, t)$ is zero whenever $Q^Z(x, y, t)$ is zero. A rigorous proof of (2) can be obtained using Girsanov's theorem for counting processes (Kipnis & Landim, 1998). Alternative derivations are given in (Cohn et al., 2010) based on explicit integration over waiting times or in (Opper & Sanguinetti, 2008) using a limit of discrete-time Markov chains for vanishing time steps.

2.3. Conditional Processes

From here on, we focus on a process X with time-independent transition function Q. Often the true state $X(t)$ of the system is only accessible through a p-dimensional observation process Y. We focus on the common scenario where observations occur at fixed discrete times $0 \leq t_1, \ldots, t_n \leq T$ that are not necessarily equidistant. Such an observation process may be written as a single matrix-valued random variable $Y = (Y(t_1), \ldots, Y(t_n))$. We assume that the observation at time t_k only depends on the latent state at time t_k and denote the conditional density of the observations as $p(y_k \mid X(t_k))$.

From a Bayesian perspective, we are interested in the conditional probability distribution $P(X(t) = x \mid y)$ for a particular realization $y \in \mathbb{R}^{p \times n}$. For this type of observation, it can be shown that the smoothed process $\tilde{X}$ associated with $P(X(t) = x \mid y)$ is a non-homogeneous MJP (Huang et al., 2016). The corresponding transition function $\tilde{Q}$ is now time dependent and given by

$$\tilde{Q}(x, y, t) = \frac{\sigma(y, t)}{\sigma(x, t)}Q(x, y) \, . \tag{3}$$

Here Q is the transition function of the prior process X and $\sigma(x, t) = p(y_k, \ldots, y_n \mid X(t) = x)$ with $k = \min_{k \in \mathbb{N}}\{t_k > t\}$. It can be shown that σ obeys the backward equation

$$\frac{d}{dt}\sigma(x, t) = -\sum_{y \neq x} Q(x, y)\left(\sigma(y, t) - \sigma(x, t)\right) \tag{4}$$

that has to be solved for the terminal condition $\sigma(x, T) = 1$ for all $x \in \mathcal{X}$ and jump conditions

$$\lim_{t \nearrow t_k} \sigma(x, t) = \sigma(x, t_k)p(y_k \mid x)$$

for all $x \in \mathcal{X}$ at the observation times t_k.

2.4. Variational Smoothing

Since the smoothed process $\tilde{X}$ as defined in Subsection 2.3 is typically intractable, we aim to find the best process level approximation Z within a simpler variational class $\mathcal{U}$. Following the usual variational inference framework (Jordan et al., 1999; Blei et al., 2017), we formalize this problem by choosing $Z^* \in \mathcal{U}$ such that

$$Z^* = \arg\min_{Z \in \mathcal{U}} D\left[P^Z \,||\, P^{\tilde{X}}\right] \tag{5}$$

minimizes the path divergence to the true posterior process. The functional optimization problem (5) still depends on the true posterior process $\tilde{X}$. As shown in the supplement, we can rewrite the objective function in (5) as

$$D\left[P^Z \| P^{\tilde{X}}\right] = D\left[P^Z \| P^X\right]$$
$$- \sum_{k=1}^{n} \mathsf{E}[\log p(y_k \mid Z(t_k)] + \text{const} \tag{6}$$

which is independent of the exact functional form of the posterior intensities.

3. Moment-Based Variational Smoothing

3.1. Variational Process Family

In our scenario, the smoothed process $\tilde{X}$ is an MJP with the modified intensities (3). Intuitively, the modified transition rate accounts for deviations of the posterior from the prior process. If the observations suggest that transition $x \to y$ has occurred more often than expected up to a certain time point, the posterior process will scale up the intensity to match the observation. The core idea of our approximation is to mimic this behavior with a process that is simpler than the full posterior. Thus, we define the variational transition function Q^Z as

$$Q^Z(x, y, t) = \lambda(x, y, t)Q(x, y), \tag{7}$$

where we have introduced the variational scaling factor λ. Note that (7) corresponds to an overparametrization of the variational problem. Solving (5) with the variational family induced by (7) recovers the true posterior transition function (3). In order to achieve a reduction in complexity, we group transitions together such that all transition within the group share a single scaling factor.

To make this more precise, let $\Psi = \{(x, y) \in \mathcal{X} \times X : x \neq y, Q(x, y) > 0\}$ be the set of all possible transitions of the prior process. Consider a partition $\Pi = \{\Pi_i \subset \Psi : i = 1, \dots, r\}$ such that $\Pi_i \cap \Pi_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^{r} \Pi_i = \Psi$. The partition Π induces a variational family by restricting the variational scaling factor λ to be of the form

$$\lambda(x, y) = \lambda_i(t) \quad \text{for} \quad (x, y) \in \Pi_i, \tag{8}$$

where the $\lambda_i : [0, T] \to [0, \infty)$ are from a suitably regular class of functions. Systematic ways to choose such a transition space partition will be discussed in Section 3.3 and Section 4.

3.2. Moment-Based Divergence

Define the natural moment functions associated with the partition Π as

$$\varphi_i(t) = \sum_{(x,y) \in \Pi_i} Q(x, y)p^Z(x, t), \tag{9}$$

where p^Z refers to the marginal distribution of the variational process $Z(t)$. It is also convenient to introduce the function $\tilde{Q}_i$ as

$$\tilde{Q}_i(x) = \sum_y \mathbf{1}_{\Pi_i}(x, y)Q(x, y)$$

with the indicator function $\mathbf{1}_{\Pi_i}(x, y) = 1$ iff the transition (x, y) is contained in Π_i. Intuitively, $\tilde{Q}_i(x)$ corresponds to the total exit rate from x to states within transition class i. We can now express the natural moment functions as

$$\varphi_i(t) = \mathsf{E}[\tilde{Q}_i(Z(t))].$$

Consider the divergence from the variational process to the prior process $D[P^Z \| P^X]$ in (6). Exploiting (7), (8) and (9) we obtain

$$D[P^Z \| P^X] = L[\varphi, \lambda]$$

where φ and λ are vector-valued functions with components λ_i and φ_i, respectively. The functional L is defined as

$$L[\varphi, \lambda] = \sum_{i=1}^{r} \int_0^T \varphi_i(t) \left(1 - \lambda_i(t) + \lambda_i(t) \log \lambda_i(t)\right) dt.$$

Since the variational process Z is a non-homogeneous MJP, the moment functions φ and the scaling factors λ are not independent, but obey a differential equation of the form (1)

$$\frac{d}{dt}\varphi_i(t) = \sum_{j=1}^{r} \lambda_j(t)\mathsf{E}\left[\sum_y \left(\tilde{Q}_i(y)\right.\right.$$
$$\left.\left. -\tilde{Q}_i(Z(t))\right) \mathbf{1}_{\Pi_j}(Z(t), y)Q(Z(t), y)\right]. \tag{10}$$

Our goal is to express the r.h.s. of (10) in terms of $\varphi(t)$, $\lambda(t)$ to obtain a closed-form description of the variational problem in terms these quantities. In general, this is not possible and we obtain an equation of the form

$$\frac{d}{dt}\varphi(t) = A(\lambda(t))\varphi(t) + B(\lambda(t))\tilde{\varphi}(t) \tag{11}$$

where A and B are suitable matrix valued functions and $\tilde{\varphi}$ is a collection of higher order moment functions that cannot be reduced to φ. In this case, we can apply an additional approximation in the form of moment closure as often used in the analysis of non-linear stochastic dynamical systems (Kuehn, 2016). Generally speaking, a moment closure scheme is a function ψ such that we may write $\tilde{\varphi}(t) = \psi(\varphi(t))$. Thus, we arrive at an equation of the form

$$\frac{d}{dt}\varphi(t) = f(\lambda(t), \varphi(t)), \tag{12}$$

where f depends on the process and the applied closure scheme. We can now recast the variational inference problem (5) into a non-linear optimal control problem of the form

$$
\begin{aligned}
\text{minimize} \quad & L[\lambda, \varphi] - F[\varphi] \\
\text{subject to} \quad & \frac{d}{dt}\varphi(t) = f(\lambda(t), \varphi(t)).
\end{aligned}
\tag{13}
$$

Here, F refers to the observation contribution in (6), i.e.

$$
F[\varphi] = \sum_{k=1}^{n} \mathsf{E}[\log p(y_k \mid Z(t_k)]
$$

which has to be understood as a functional of the natural moments φ. Before turning to special classes of MJPs, we end this section with a remark on moment closure. When the moment equations (11) are naturally closed (i.e. $B = 0$), than the control problem (13) corresponds exactly to the process level formulation (5) and the minimum of L corresponds to the usual evidence lower bound. When we use moment closure, there is no global process equivalent to (13) and the lower bound property is lost. This effect has also been observed for other variational approximations that go beyond product mean-field, for example cluster variational methods (Yedidia et al., 2000). In our case, we can recover a valid process level result using the solution λ^* in (7). Generating sample paths from the so defined approximate posterior gives a means to analyze the accuracy of the moment closure approximation.

3.3. Special Classes

In this section, we consider special classes of structured MJPs that give rise to a natural partition of the transition space.

State Space Lumping. Consider an MJP X with countable state space $\mathcal{X}$ and let $\{\Pi_1, \ldots, \Pi_s\}$ be a partition of the state space (in contrast to the partition of the transition space). The lumped process can now be defined as

$$
S(t) = j \quad , \quad \text{for } X(t) = x, \; x \in \Pi_j \,.
$$

A partition of the state space naturally induces a partition of the transition space given by

$$
\Pi_{ij} = \{(x, y) \in \{1, \ldots, s\}^2 : x \in \Pi_i, y \in \Pi_j\}\,.
$$

This partition subsumes all transitions of the lumped process. Accordingly, the natural moment functions become

$$
\varphi_{ij}(t) = \sum_{x,y} \mathbf{1}_{\Pi_i}(x)\mathbf{1}_{\Pi_j}(y)Q(x,y)q(x,t)\,.
$$

It is also interesting to observe that

$$
\tilde{Q}_{ij}(x) = \mathbf{1}_{\Pi_i}(x) \sum_y \mathbf{1}_{\Pi_j}(y)Q(x,y)
$$

which is the total transition rate from x in Π_i to Π_j. In the special case that the MJP is lumpable with respect to the partition Π (Rubino & Sericola, 1993), we obtain

$$
\tilde{Q}_{ij}(x) = \begin{cases} \tilde{Q}_{ij} & x \in \Pi_i \\ 0 & \text{otherwise} \end{cases},
$$

where $\tilde{Q}_{ij} = \sum_{y \in \Pi_j} Q(x,y)$ which is identical for all $x \in \Pi_i$. Inserting this into the dynamic equation yields

$$
\frac{d}{dt}\varphi_{ij} = \tilde{Q}_{ij}\left[\sum_{l \neq i} \lambda_{li}(t)\varphi_{il}(t) - \sum_{l \neq i} \lambda_{il}(t)\varphi_{il}(t)\right]
\tag{14}
$$

Dividing both sides by $\tilde{Q}_{ij}$, we observe that (14) is equivalent to the master equation of the lumped process S. Accordingly, our variational approach will recover the exact posterior of the lumped process. While most examples of practical interest are not lumpable, (14) provides a way to exploit known approximate lumping schemes for inference purposes by understanding the approximate lumping as a moment closure scheme.

Data Driven Partitioning. In many cases, the observations Y do not depend on the full latent state $X(t)$ but rather on a summary statistic $S(t) = T(X(t))$ for $T : \mathcal{X} \rightarrow \mathcal{S}$. In general, $S(t)$ will be a non-Markovian jump process on $\mathcal{S}$. However, it induces a natural partition of the form

$$
\Pi_{z,z'} = \{(x, x') \in \Psi : T(x) = z, T(x') = z'\}
$$

meaning that we group all transitions of X that lead to the same state change of S in $\mathcal{S}$.

Agent-Based MJPs. An agent-based MJP X on $\mathcal{X} = \{1, \ldots, m\}^d$ describes the behavior of d coupled agents where each agent may be in one of m states. In addition, the probability of two agents changing the state at the same time is assumed to be zero. In terms of the transition function, we obtain

$$
Q(x, y) = \begin{cases} Q_{jk}^i(x) & x_i = j, y_i = k, x_l = y_l \text{ for } l \neq i \\ 0 & \text{otherwise} \end{cases}
$$

where Q_{jk}^i is the transition rate of agent i from state j to state k as a function of the state of the other agents. This is the class of processes targeted by the existing mean-field approaches for jump processes (Opper & Sanguinetti, 2008; Cohn et al., 2010). Since our variational family is constructed by modifying the prior process, we cannot recover the product-type approximation used in the classical approaches. We can, however, mimic its behavior by subsuming all transitions that produce the same change of the same agent, i.e.

$$
\Pi_{jk}^i = \{(x, y) \in \Psi : x_i = j, y_i = k, x_l = y_l\}\,.
\tag{15}
$$

Using (15), we can evaluate the corresponding moment functions as

$$\varphi_{jk}^i(t) = \mathsf{E}[\mathbf{1}_j(Z_i(t))Q_{jk}^i(Z(t))],$$

which is quite intuitive, since it is the transitions rate of agent i to go from j to k averaged over all configurations of the remaining agents.

Population-Based MJPs. A population-based MJP X on $\mathcal{X} = \mathbb{N}_0^d$ describes the stochastic evolution of the abundances of d species over time. While they seem similar to agent models from the previous section, population models usually have an unbounded state space. In addition, one event may affect several species at the same time. This fact is formalized by a set of change vectors $\{v_i \in \mathbb{Z}_0^d : i = 1, \ldots, r\}$ and a transition function of the form

$$Q(x, y) = \begin{cases} h_i(x) & \text{if } y = x + v_i \\ 0 & \text{otherwise} \end{cases}.$$

If there are change vectors that affect several species, the product mean-field approach is ill-suited for this type of system, since a product of independent processes will not be absolutely continuous with respect to the prior process. In our framework, a natural choice of partition for this type of model is to combine all transitions that correspond to the same change vector

$$\Pi_i = \{(x, y) \in \Psi : y = x + v_i\}.$$

The corresponding natural moment functions become

$$\varphi_i(t) = \mathsf{E}[h_i(Z(t))].$$

The time-evolution of φ is given by

$$\frac{d}{dt}\varphi_i(t) = \sum_j \lambda_j(t)\mathsf{E}\left[(h_i(Z(t) + v_j)\right.$$
$$\left. - h_i(Z(t)))\, h_j(Z(t))\right].$$

For systems where h_i are linear functions of the state, this equation is closed in φ_i. An example of this class will be discussed later.

3.4. Minimizing the Divergence.

We now introduce an algorithm to find approximately optimal variational scaling functions λ^*. We do this following the indirect approach from optimal control (Bonnard & Caillau, 2006). The idea here is to transform the constrained problem (13) into an unconstraint problem by introducing the Lagrangian

$$J[\lambda, \varphi, \eta] = L[\lambda, \varphi] - F[\varphi]$$
$$- \int_0^T \eta(t)^T \left[f(\lambda(t), \varphi(t)) - \dot{\varphi}(t)\right],$$

where η is a vector-valued Lagrange multiplier function. By taking the functional derivative with respect to φ_i we obtain

$$\frac{d}{dt}\eta_i(t) = \sum_{j=1}^r (1 - \lambda_j(t) + \lambda_j(t)\log\lambda_j(t))$$
$$- \sum_{j=1}^r \frac{df_j}{d\varphi_i}\eta_j(t) \tag{16}$$

valid in between the observations. At the point of the observations, the functional F will induce jump conditions for η given by

$$\lim_{t \nearrow t_k} \eta_i(t) = \eta_i(t_k) + \frac{d}{d\varphi_i(t_k)}\mathsf{E}[p(y_k \mid Z(t_k))]. \tag{17}$$

Similarly, differentiation with respect to the scaling factor λ_i yields an algebraic constraint of the form

$$0 = \varphi_i(t)\log\lambda_i(t) - \sum_{j=1}^r \eta_j(t)\frac{df_j}{d\lambda_i}. \tag{18}$$

In the classical mean-field algorithm, the scaling functions λ can be expressed in terms of η by solving (18). The control problem is then solved using a form of the forward-backward sweep method (Mcasey et al., 2012). In our framework, it is generally not possible to eliminate the scaling factors. A variation of the forward-backward sweep explicitly including the scaling factors turned out to be unstable in our experiments.

We therefore propose a gradient-based algorithm based on the following argument. The moment functions φ are fully determined by the scaling factors λ and an initial condition. Hence, we may understand L as a functional of φ only. In this case, the r.h.s. of (18) corresponds to the gradient of L with respect to φ. Since the raw gradient updates obtained this way do not work too well, we take advantage of the probabilistic nature of the objective function to derive an analogue to the natural gradient (Amari, 1998). We do this by performing an expansion of the path divergence up to second order in λ. From this, we obtain updates of the form

$$\lambda^{(n+1)} = \lambda^{(n)} - h\tilde{\nabla}L[\lambda^{(n)}] \tag{19}$$

with step size h and the and the natural gradient

$$\tilde{\nabla}L[\lambda] = \lambda_i(t)\log\lambda_i(t) - \frac{\lambda_i(t)}{\varphi_i(t)}\sum_{j=1}^r \eta_j(t)\frac{df_j}{d\lambda_i}. \tag{20}$$

The updates (19), (20) ensure that we take small steps on the manifold defined by the variational family and improve convergence significantly. The resulting algorithm is summarized in Algorithm 1. A more detailed discussion including a derivation of the natural gradient can be found in the supplement Section 3.

Algorithm 1 Natural Gradient Descent for MB-VI

1: **Input:** Initial guess for the scaling factors $\lambda^{(0)}$, initial condition $\psi(0)$.
2: **repeat**
3:   Given $\lambda^{(n)}$, $\varphi(0)$, compute $\varphi^{(n)}$ using (12).
4:   Given $\lambda^{(n)}$, $\varphi^{(n)}$, compute $\eta^{(n)}$ using (16), (17).
5:   Given $\lambda^{(n)}$, $\varphi^{(n)}(t)$ and $\eta^{(n)}$, evaluate the natural gradient $\tilde{\nabla}L$ according to (20).
6:   Set $\lambda^{(n+1)}$ according to (19).
7: **until** $|L[\lambda^{(n)}] - L[\lambda^{(n-1)}]| <$ tolerance
8: **Output:** Optimized variational scaling factor λ^* .

4. Parameter Inference

Suppose now that the transitions function Q of the prior process depends on a collection of parameters θ. Using the usual framework of the variational EM algorithm, we can find an approximate maximum likelihood estimate by understanding the objective function $L[\lambda, \theta]$ as a function of the scaling factors and the parameters and then minimizing iteratively with respect to both arguments. In general, not much can be said about the structure of the resulting optimization problem. In the following, we focus on a special case that allows for closed form parameter updates.

Let Π be a partition of the transition space such that the prior process has parametrized transition function of the form

$$Q_\theta(x, y) = c_i(\theta)h(x, y) \quad \text{for} \quad (x, y) \in \Pi_i,$$

where c_i and h are known functions with c_i differentiable with respect to θ. We proceed by defining our variational family on Π by setting

$$\lambda(x, y, t) = \lambda_i(t)h(x, y) \quad \text{for} \quad (x, y) \in \Pi_i. \quad (21)$$

Note that (21) slightly differs from the ansatz introduced in Section 3.1 in the fact that it only contains a part of the prior intensity function. As a consequence, the variational family remains independent of the parameters θ. Setting

$$\varphi_i(t) = \sum_{(x,y) \in \Pi_i} h(x, y)q(x, t),$$

we can reuse the formalism introduced in section 3.2 to obtain stationarity conditions for λ. In addition, for fixed λ, φ we obtain optimality conditions for the parameters of the form

$$\sum_j (G_j c_j(\theta) - H_j) \frac{dc_j(\theta)}{d\theta_i} = 0,$$

where we have introduced the summary statistics

$$G_i = \int_0^T \varphi_i(t)dt,$$

$$H_i = \int_0^T \varphi_i(t)\lambda_i(t)dt.$$

Intuitively, H_i corresponds to the expected number of transitions in class i during $[0, T]$. In the case where we have a single parameter per transition class, i.e. $c_i(\theta) = \theta_i$, we get

$$\theta_i = \frac{H_j}{G_j}.$$

As an alternative to the EM approach, we may also follow a Bayesian framework, i.e. by assuming a prior $p(\theta)$. Choosing a variational ansatz that factorizes over parameters and latent trajectories, we get an exponential family variational parameter distribution of the form

$$q(\theta) \propto q(\theta) \prod_j c_j(\theta)^{H_j} \exp\left(-\sum_j c_j(\theta)G_j\right). \quad (22)$$

If the prior is a gamma distribution and the we have a single parameter per transitions class, (22) will also have the form of a gamma distribution.

5. Examples

In this section, we apply our method to three examples. We focus on models of the population type for which inference is notoriously difficult due to the unbounded state space. First we consider a linear birth death process. This simple example allows analytic treatment and provides some intuition. Next, we numerically study a stochastic gene expression model of the type that is frequently used in systems biology. As a third example, we consider a stochastic predator prey model to demonstrate our method in combination with moment closure.

5.1. Birth Death Process

Let $\mathcal{X} = \mathbb{N}_0$, $X_0 \in \mathcal{X}$ and $c_1, c_2 > 0$. For the transition vectors $v_1 = 1$, $v_2 = -1$ together with the intensity

$$Q(x, y) = \begin{cases} c_1 & y = x + 1 \\ c_2 x & y = x - 1 \end{cases}$$

the resulting process $\{X(t), t \geq 0\}$ is known as the linear birth death process and can be seen as the simplest case of a population-type model. We therefore following the approach of choosing the partition $\Pi = \{\Pi_1, \Pi_2\}$ with

$$\Pi_1 = \{(x, y) : y = x + 1\}, \ \Pi_1 = \{(x, y) : y = x - 1\}$$

which induce the moment functions

$$\varphi_1(t) = c_1, \ \varphi_2(t) = c_2\mathsf{E}[Z(t)].$$

Since, φ_1 is constant, the moment dynamics are reduced to a single closed equation of the form

$$\frac{d}{dt}\varphi_2(t) = c_2\lambda_1(t)\varphi_1 - c_2\lambda_2(t)\varphi_2(t). \quad (23)$$

Now suppose we consider the conditional process with end-point condition $X(T) = 0$. For simplicity, let us also assume that $X_0 = 0$. In this case, the backward equation can be integrated analytically. This allows to compute the variational scaling factors as

$$\lambda_1(t) = (1 - \exp(T - t)),$$
$$\lambda_2(t) = (1 - \exp(T - t))^{-1}.$$

Inserting this into (23) we obtain

$$\mathsf{E}[Z(t)] = \frac{c_1}{c_2}(1 - \exp[-c_2(T - t)])(1 - \exp(-c_2 t)).$$

For this example, the true posterior intensities can be calculated analytically (Huang et al., 2016) leading to the same result. Hence, the variational approximation coincides with the exact smoothed process in this simple special case. A graph of the first moment and the corresponding standard deviation can be found in Fig. 1.

Now consider the same example where instead of a terminal condition we have obtained a single observation $Y_T = y$ with zero mean additive Gaussian noise of variance σ^2. In this case, the objective function becomes more complicated, since the second order moment has to be included in order to describe the observations. Consequently, there seems to be no simple closed form expressions to characterize the variational posterior. We ran our algorithm for different values of σ^2 and verified empirically that the solution converges to the analytic expression for the exact terminal constraint case with $\sigma \to 0$. Fig. 1 shows the mean of the prior process and the noise-free conditional process compared two two observations with different noise levels. As expected, the variational method seems to interpolate between the observation and the prior process.

5.2. Gene Expression System

Consider the following simple gene expression model consisting of a single gene that can switch between an inactive state G^0 and an active state G_1 with rates c_1 and c_2. While the gene is active, mRNA is produced at rate c_3 (transcription). The mRNA is then processed by ribosomes to produce the target protein at rate c_5 (translation). Both mRNA and protein molecules undergo degradation reactions with rates c_4 and c_6 respectively.

The stochastic behavior of this system is modeled by a population MJP X with state space $\mathcal{X} \subset \mathbb{N}_0^3$ where the components correspond to the numbers of active genes X_1, mRNA molecules X_2 and protein molecules X_3. The nonzero elements of the transition function Q are given by

$$
\begin{aligned}
Q(x, x + e_1) &= c_1(x_1 - 1), &\quad Q(x, x - e_1) &= c_2 x_1, \\
Q(x, x + e_2) &= c_3 x_1, &\quad Q(x, x - e_2) &= c_4 x_2, \\
Q(x, x + e_3) &= c_5 x_2, &\quad Q(x, x - e_3) &= c_6 x_3.
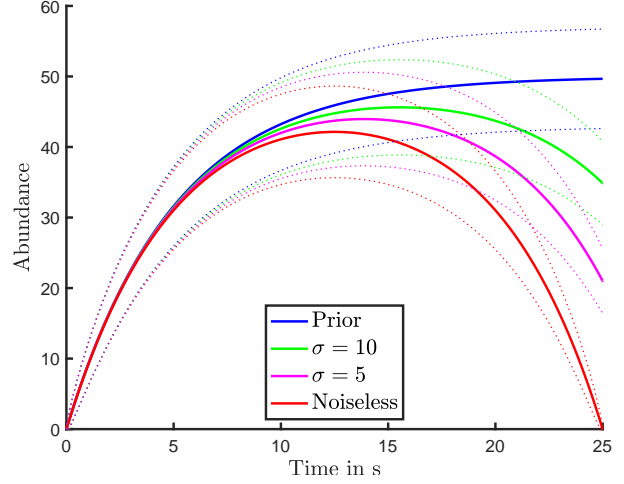\end{aligned}
$$

Figure 1. Mean (thick lines) and standard deviation around the mean (dotted lines) for different endpoint constraints. All results used the parameter configuration $c_1 = 5\mathrm{s}^{-1}$, $c_2 = 0.1\mathrm{s}^{-1}$.

In typical applications, the produced protein molecule is a fluorescent reporter. Measurements of the reporter concentration can be obtained from live-cell microscopy by integrating fluorescence intensity over the cross-section of the cell. Thus, these measurements are a scaled and noisy observation of the process component corresponding to the protein abundance. For simplicity, we assume that the noisy observations $Y_1, \ldots, Y_n$ are conditionally Gaussian given the protein abundance. To construct our variational approximation, we choose the population-type partitioning and obtain closed form moment equations that can be found in Section 5 of the supplement.

The results for variational posterior mean and variance are shown in Fig. 2. The example demonstrates that the posterior mean concentrates on the true realization of the latent process. Thus, the method allows to infer latent state dynamics, in particular, the unobserved activation patterns of the gene are recovered quite well. To investigate this further, we used the posterior mean of the gene state to devise a detector (threshold: 0.5) and evaluated the true positive fraction $\alpha = 0.94$ and false positive fraction $\beta = 0.15$ from $N = 100$ pooled trajectories. We note that false detections are mainly caused by short events. For example, if the gene is briefly inactive during a longer period of activity, this may not be visible in the protein activity and even less so in the observations.

Fig 2 also reveals a drawback of our method. From an exact smoothing, we would expect the variance of the posterior to be smaller at times of observations than in between the observations. We do not observe this behavior in our approach. The most likely reason for this is that a single
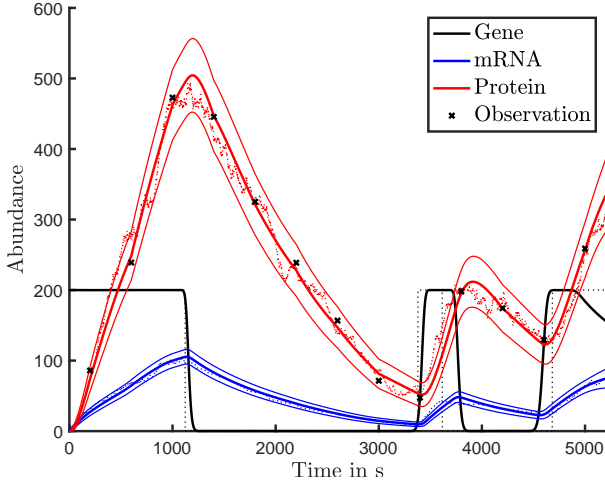
Figure 2. Approximate smoothing result for simulated observations from a gene expression model. Thick bold lines denote the mean of the variational posterior process, while the thin lines indicate an interval of one standard deviation around the mean. The true latent trajectories are shown as dotted lines. Note that we only considered a single gene but scaled the curve in the plot for better visibility.

Figure 3. Approximate smoothing result for simulated observations from a stochastic predator prey model. Thick bold lines denote the mean of the variational posterior process, while the thin lines indicate an interval of one standard deviation around the mean. The true latent trajectories are shown as dotted lines.

time-dependent factor per reaction channel does not provide enough degrees of freedom to capture this effect. In the end, our approach reduces the full smoothing problem to a set of nine ordinary differential equations.

5.3. Predator-Prey Dynamics

To demonstrate our approach on a non-linear model, we consider a stochastic form of the classical predator prey interaction model. Here, a prey species X_1 and a predator species X_2 interact with each other as defined by the transition function Q with non-zero elements

$$Q(x, x + e_1) = c_1 x_1, \qquad Q(x, x - e_1) = c_2 x_1 x_2,$$
$$Q(x, x + e_1) = c_3 x_1 x_2, \quad Q(x, x - e_1) = c_4 x_2.$$

Note that this system is unstable in the sense that, in the long run, we observe either extinction of both species or extinction of the predator and explosion of the prey population. Using our population-type partitioning, we obtain a set of moment equations that is not closed. In Supplement Sec. 5, we demonstrate how moment closure can be employed to obtain a closed system.

An application of the variational smoother to synthetic data of the predator prey model is given in Fig. 3. Due to the low abundances, it is possible to perform a truncation of the state space and solve the smoothing problem by integrating (4) backward in time. Comparing Fig. 3 to exact smoothing results reveals that the variance of the approximate posterior
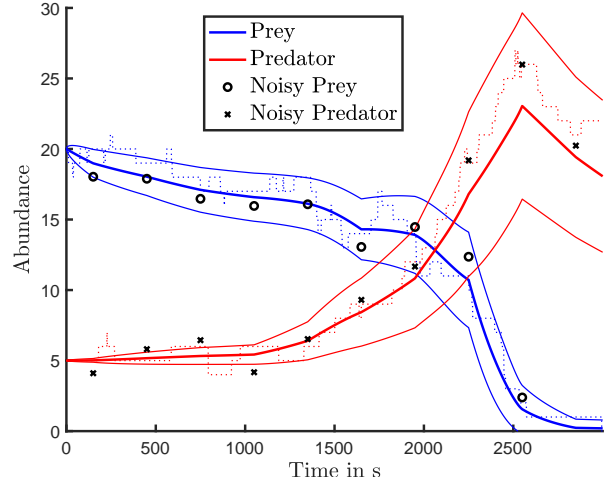
is higher than the variance of the exact smoothed process.

6. Discussion

On the basis of a partitioning of the transition space, we proposed a flexible framework to construct variational inference procedures for Markov jump processes. By defining the variational family as a modification of the prior process, our approach circumvents problems with absolute continuity that have plagued earlier product-type mean-field approaches. The main advantage of our method is, however, that the complexity of the approximation depends on the chosen partition of the transition space and can thus be adapted to the problem at hand. To get some insight into the choice of partition, we discussed classes of structured MJPs that give rise to a natural partitioning. As other variational procedures, moment based variational inference can be used for approximate parameter inference. An interesting observation here is that a suitable choice of partitioning may naturally lead to a gamma variational distribution of the parameters. Finally, we demonstrated the method on synthetic examples of the population type.

Our work opens several lines of future research. First of all, we envision applications to real data, e.g. in the context of systems and synthetic biology. On the theoretical side, it may be interesting to assess the effect of different partition and moment closure choices on the approximation quality. Finally, a more systematic comparison with related methods is required.

Acknowledgements

References

Amari, S.-I. Natural Gradient Works Efficiently in Learning. *Neural Comput.*, 10(2):251–276, 1998.

Anderson, D. F. and Kurtz, T. G. *Stochastic Analysis of Biochemical Systems*. Stochastics in Biological Systems. Springer, New York, 2015.

Anderson, W. J. *Continuous-time Markov chains : an applications oriented approach*, volume 7 of *Springer series in statistics. Probability and its applications*. New York [u.a.], 1991.

Andrieu, C., Doucet, A., and Holenstein, R. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3): 269–342, 2010.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-Taylor, J. Variational Inference for Diffusion Processes. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 17–24. Curran Associates Inc., 2007.

Bladt, M. and Sørensen, M. Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):395–410, 2005.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Bonnard, B. and Caillau, J.-B. Introduction to Nonlinear Optimal Control. In Loría, A., Lamnabhi-Lagarrigue, F., and Panteley, E. (eds.), *Advanced Topics in Control Systems Theory: Lecture Notes from FAP 2005*, pp. 1–60. Springer London, 2006.

Cohn, I., El-Hay, T., Friedman, N., and Kupferman, R. Mean Field Variational Approximation for Continuous-Time Bayesian Networks. *J. Mach. Learn. Res.*, 11:2745–2783, 2010.

Doucet, A. and Johansen, A. M. *A tutorial on particle filtering and smoothing: fifteen years later*. 2011.

Ethier, S. N. and Kurtz, T. G. *Markov Processes : Characterization and Convergence*. Wiley series in probability and statistics. 2005.

Frigola, R., Lindsten, F., Schön, T. B., and Rasmussen, C. E. Bayesian Inference and Learning in Gaussian Process State-Space Models with Particle MCMC. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 3156–3164. Curran Associates, Inc., 2013.

Golightly, A. and Wilkinson, D. J. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1 (6):807–820, 2011.

Huang, L., Pauleve, L., Zechner, C., Unger, M., Hansen, A. S., and Koeppl, H. Reconstructing dynamic molecular states from single-cell time series. *Journal of The Royal Society Interface*, 13(122), 2016.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999.

Kipnis, C. and Landim, C. *Scaling Limits of Interacting Particle Systems*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1998.

Kuehn, C. Moment Closure—A Brief Review. In Schöll, E., Klapp, S. H. L., and Hövel, P. (eds.), *Control of Self-Organizing Nonlinear Systems*, pp. 253–271. 2016.

Liu, Y.-Y., Li, S., Li, F., Song, L., and Rehg, J. M. Efficient Learning of Continuous-Time Hidden Markov Models for Disease Progression. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 3600–3608. Curran Associates, Inc., 2015.

Mcasey, M., Mou, L., and Han, W. Convergence of the forward-backward sweep method in optimal control. *Computational Optimization and Applications*, 53, 09 2012.

Metzner, P., Horenko, I., and Schütte, C. Generator estimation of Markov jump processes based on incomplete observations nonequidistant in time. *Phys. Rev. E*, 76(6): 066702, 2007.

Opper, M. and Sanguinetti, G. Variational inference for Markov jump processes. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 1105–1112. Curran Associates, Inc., 2008.

Rao, V. and Teh, Y. W. Fast MCMC sampling for markov jump processes and extensions. *J. Mach. Learn. Res.*, 14 (1):3295–3320, January 2013.

Rubino, G. and Sericola, B. A finite characterization of weak lumpable Markov processes. Part II: The continuous time case. *Stochastic Processes and their Applications*, 45(1): 115–125, 1993.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. Generalized Belief Propagation. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pp. 668–674. MIT Press, 2000.

Zhang, B., Pan, J., and Rao, V. A. Collapsed variational Bayes for Markov jump processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3749–3757. Curran Associates, Inc., 2017.