Exact inference for a class of hidden Markov models on general state spaces

Guillaume Kon Kam King

Université Paris-Saclay, INRAE, MaIAGE,
78350, Jouy-en-Josas, France
e-mail: guillaume.kon-kam-king@inrae.fr

Omiros Papaspiliopoulos

Department of Decision Sciences, Bocconi University,
Via Röntgen 1, Milan, Italy
e-mail: omiros@unibocconi.it

Matteo Ruggiero

University of Torino and Collegio Carlo Alberto,
Corso Unione Sovietica 218/bis, 10134, Torino, Italy
e-mail: matteo.ruggiero@unito.it

Abstract: Exact inference for hidden Markov models requires the evaluation of all distributions of interest – filtering, prediction, smoothing and likelihood – with a finite computational effort. This article provides sufficient conditions for exact inference for a class of hidden Markov models on general state spaces given a set of discretely collected indirect observations linked non linearly to the signal, and a set of practical algorithms for inference. The conditions we obtain are concerned with the existence of a certain type of dual process, which is an auxiliary process embedded in the time reversal of the signal, that in turn allows to represent the distributions and functions of interest as finite mixtures of elementary densities or products thereof. We describe explicitly how to update recursively the parameters involved, yielding qualitatively similar results to those obtained with Baum–Welch filters on finite state spaces. We then provide practical algorithms for implementing the recursions, as well as approximations thereof via an informed pruning of the mixtures, and we show superior performance to particle filters both in accuracy and computational efficiency. The code for optimal filtering, smoothing and parameter inference is made available in the Julia package DualOptimalFiltering.

Keywords and phrases: Hidden Markov models, optimal filtering, smoothing, diffusion process, Cox–Ingersoll–Ross, Wright–Fisher.

Contents

1. Introduction

Inference for Hidden Markov Models (HMMs), is a fundamental statistical problem concerned with learning the parameters and the trajectory of an unobserved Markov process, called *signal*, given noisy or indirect observations, typically collected at discrete times. Sometimes also referred to as state-space models, HMMs have found widespread application in a variety of frameworks that include genomics (Brown et al., 1993; Yau et al., 2011; Guha et al., 2008; Titsias et al., 2016), proteomics (Bae et al., 2005), time series analysis (Sarkar and Dunson, 2019), temporal clustering (Crane, 2017), signal processing (Fox et al., 2011), econometrics (Hamilton, 1990; Chib, 1996), brain imaging, target tracking and animal movement (Quick et al., 2017; Langrock et al., 2015), to mention a few

examples. General treatments of HMMs can be found for example in Chopin and Papaspiliopoulos (2020); Cappé et al. (2005); Särkkä (2013).

Let $\{X_t, t \geq 0\}$ be a Markov process on $\mathcal{X} \subseteq \mathbb{R}^K$, with initial distribution ν_0 and transition density $P_t^{\psi}(x' \,|\, x)$, parametrised by a finite-dimensional vector ψ. The process $\{X_t, t \geq 0\}$, referred to as the *hidden signal*, is assumed to be unobserved and to evolve in continuous time. Observations $Y_t \in \mathcal{Y} \subseteq \mathbb{R}^D$ are collected at discrete times $0 \leq t_0 < t_1 < \ldots$ and taken to be conditionally independent of everything else given the current value of the signal, to which they relate through the *emission density* $f_x^{\psi}(\cdot)$, parametrised by both ψ and x, whereby $Y_t \,|\, (X_t = x) \overset{\text{ind}}{\sim} f_x^{\psi}(\cdot)$.

In this framework, the quantities of statistical interest are typically given by the density of the signal given past observations $p^{\psi}(x_{t_{i+k}} \,|\, y_{t_0}, \ldots, y_{t_i})$, i.e., the *predictive distribution*; or given observations up to present time $p^{\psi}(x_{t_i} \,|\, y_{t_0}, \ldots, y_{t_i})$, called *filtering distribution*; or given past, present and future observations $p^{\psi}(x_{t_{i-k}} \,|\, y_{t_0}, \ldots, y_{t_i})$, called *smoothing distribution*. In addition, the likelihood of the observations $p^{\psi}(y_{t_0}, \ldots, y_{t_i})$ is another primary object of interest, as it allows performing inference on the model parameters ψ. Here the likelihood is obtained by integrating out the hidden trajectory of the signal from the joint density $p^{\psi}(x_{t_0}, \ldots, x_{t_i}, y_{t_0}, \ldots, y_{t_i})$. In the following, we will drop the superscript ψ for notational simplicity, and return to the problem of drawing inference on ψ in Section 5.3.

Performing exact inference for HMMs entails being able to compute the above distributions with a finite computational effort. Two important classes of models have long been known to allow such computation. The first requires the signal state space $\mathcal{X}$ to be a finite set, whereby the quantities of interest are obtained through the Baum–Welch filter (see Cappé et al., 2005) and elaborations thereof. The second is given by linear Gaussian systems (e.g. Ornstein–Uhlenbeck signals and linearly linked Gaussian emissions), in which case all quantities of interest are obtained by updating parameters of Gaussian distributions through the celebrated Kalman–Bucy filter and elaborations thereof. The common feature of the two above cases is the existence of a finite-dimensional process which completely characterises the distributions of interest, so that these can be obtained by appropriately updating this process, called *finite-dimensional filter*. The *computational complexity* of the filter is given by the number of operations needed for performing such updates, which for finite-dimensional filters grows linearly in the number of observations. Outside the above mentioned classes, finite-dimensional filters are typically rare and difficult to obtain. See Ferrante (1992); Ferrante and Runggaldier (1990); Ferrante and Vidoni (1998); Guenther (1981); Runggaldier and Spizzichino (2001).

A major breakthrough in the study of HMMs was achieved in Chaleyat-Maurel and Genon-Catalot (2006), who introduced the notion of *computable inference* for HMMs. See also Genon-Catalot (2003); Genon-Catalot and Kessler (2004); Chaleyat-Maurel and Genon-Catalot (2009); Comte et al. (2011). This is based on a general assumption made on the signal transition semigroup, and applies to models for which the distributions of interest can be characterised

by a finite-dimensional process whose size, however, can increase as the number of observations increases. In such cases, the recursive updates can be shown to have polynomial computational complexity in the number of observations. Chaleyat-Maurel and Genon-Catalot (2009); Comte et al. (2011), for example, found explicit solutions for specific one-dimensional HMMs along these lines. In this framework, Papaspiliopoulos and Ruggiero (2014) identified a structural property of the transition density of the signal that implies the conditions in Chaleyat-Maurel and Genon-Catalot (2006) and thus leads to computable filtering. Specifically, they showed that if the signal has a dual process, i.e. a certain auxiliary process embedded in its time reversal, given by a pure-death process on a multidimensional grid subordinated to an ODE, then all filtering distributions are finite mixtures of parametric densities which are conjugate to the emission density of the observations. These conditions are formulated for generic K-dimensional HMMs, and the implied computations do not rely on specific properties of the model at hand (Chaleyat-Maurel and Genon-Catalot (2009) for example exploit the specific eigenstructure of the transition semigroup). Later, Papaspiliopoulos et al. (2016) showed computable filtering is possible for two signals taking values on the space of atomic measures, namely Fleming–Viot and Dawson–Watanabe measure-valued diffusions, and more recently Ascolani et al. (2020) applied their results to Bayesian predictive inference in a nonparametric framework.

The present article obtains theory and practical algorithms for the whole inferential agenda for HMMs. While Papaspiliopoulos and Ruggiero (2014); Papaspiliopoulos et al. (2016) focused only on the theoretical aspects of filtering, here we develop and implement results also on smoothing and inference on the model parameters. Under a set of sufficient conditions essentially analogous to those in Papaspiliopoulos and Ruggiero (2014), we show that all distributions of interest for the signal can be expressed as finite mixtures of elementary densities, and the likelihood of the observations takes the form of a finite product of mixtures. We provide explicit recursive formulae that describe the parameter updates, and also show how to obtain samples from the joint smoothing distribution of the signal. Moreover, although the number of components in the finite mixtures can grow rapidly, the number of components carrying most of the probability mass is seemingly stationary (see Figure 1). It thus seems natural to approximate the mixtures by, e.g., pruning all components with negligible mass, resulting in a roughly constant number of components. In Section 4 we discuss several pruning strategies and detail practical algorithms for computable inference with this class of HMMs. In particular we propose an automatic mixture pruning scheme that results in linear computational costs, at the expense of some approximation error.

By devising appropriate metrics for evaluating the quality of the mixture approximations, we compare the performance of our schemes against suitable particle filters and find superior performance. Figure 2 provides a glimpse into these results, showing that for the two classes of models used in Figure 1, computing the likelihood with one of the proposed approximation strategies outperforms particle filters both in accuracy and in computational efficiency.
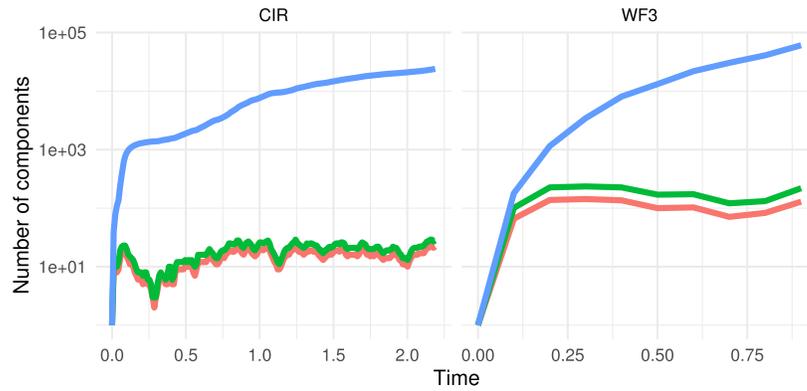
FIG 1. *Number of components (in log scale) in the filtering densities as a function of time for two models illustrated in detail later (a Cox–Ingersoll–Ross process and a 3-component Wright–Fisher process), needed to account for 95% (red), 99% (green) and 100% (blue) of the mass.*
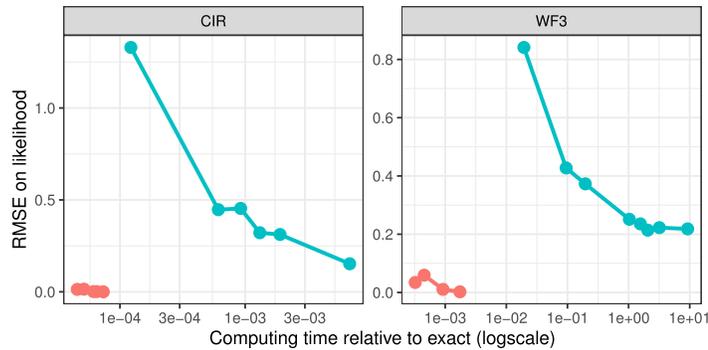
FIG 2. *Root mean squared error for computing the likelihood, as a function of computing time, for one approximation strategy proposed later (red dots, for various levels of approximation) compared with using a particle filter (blue dots, for various number of particles), for the models in Figure 1. The computing time is relative to the time needed to compute the likelihood exactly (feasible for these model using the results presented later).*

The code for performing optimal filtering, smoothing and parameter inference using the result presented here is made available as the Julia package `DualOptimalFiltering`[1]. We also implemented a novel particle filtering algorithm using the recent exact sampling method of Jenkins and Spanò (2017), with the algorithm for exact sampling made available as the Julia package `ExactWrightFisher`[2].

The rest of the paper is organised as follows. Section 2 presents our theoretical results in terms of sufficient conditions for computable inference, details the

---

[1]Available at https://github.com/konkam/DualOptimalFiltering.jl
[2]Available at https://github.com/konkam/ExactWrightFisher.jl

associated general recursive formulae for updating the distributions of interest, and illustrates how this is concretely done in the case of a Cox–Ingersoll–Ross (CIR) and a 3-component Wright–Fisher (WF) signal, by describing explicitly the parameter updates for these models. Section 3 shows how to sample from the joint smoothing distribution in order to generate trajectories of the signal conditional on the data, and illustrates the strategy for the two above models. While for the first this comes out quite naturally from the computation, for the WF it requires a non trivial argument, which leverages on recent results by Jenkins and Spanò (2017). Section 4 discusses the algorithmic implementation and proposes some acceleration strategies via informed pruning of the mixtures, whereas Section 5 presents numerical experiments concerning signal recovery, parameter estimation via maximum likelihood and full Bayesian inference via Monte Carlo Markov Chain (MCMC) algorithms. We compare the performance of our strategies against particle filtering. Concluding remarks are provided in a brief discussion. The most relevant proofs are included in Appendix A, while a few additional results are deferred to Appendix B.

2. Exact recursive inference for a class of HMMs

2.1. Setting and inferential goals

Let the hidden signal X be a continuous-time homogeneous Markov process on $\mathcal{X} \subseteq \mathbb{R}^K$ with initial distribution ν_0 and transition density P_t. We consider the classical HMM setting where given X, observations Y_t are independent and the conditional distribution of Y_t given $X_t = x$, called *emission density*, only depends on x and not t. We further assume the emission densities have a common dominating measure on $\mathcal{Y}$ for all $x_t \in \mathcal{X}$.

Denote by $x_{0:T} := (x_0, \ldots, x_T)$ the unknown signal values at the discrete times $0 = t_0 < \ldots < t_T$ and by $y_{0:T} := (y_0, \ldots, y_T)$ the observations collected at the same time points. For notational simplicity, and without loss of generality, we assume that only one data point is observed at each time and that the observation times are equally spaced by an interval $\Delta = t_i - t_{i-1}$. We therefore denote by P_Δ the transition density of the signal on such intervals, write $x_i := x_{t_i}$ and $y_i := y_{t_i}$, and (with some abuse) refer to t_i as time i.

Inference in this setting is concerned with evaluating various conditional distributions of the signal given the observations. These are typically classified into: *predictive densities*

$$\nu_{i|0:i-1}(x_i) := p(x_i \,|\, y_{0:i-1}) = \int_{\mathcal{X}} p(x_{i-1} \,|\, y_{0:i-1}) P_\Delta(x_i \,|\, x_{i-1}) \mathrm{d}x_{i-1} \qquad (1)$$

which evaluate the law of the signal given past data; *filtering densities*

$$\nu_{i|0:i}(x_i) := p(x_i|y_{0:i}) \propto \int_{\mathcal{X}^i} p(x_{0:i}, y_{0:i}) \mathrm{d}x_{0:i-1}, \qquad (2)$$

given past and present data, where $p(x_{0:i}, y_{0:i})$ is the joint density of signal and observations; and *smoothing densities*

$$\nu_{i|0:T}(x_i) := p(x_i \,|\, y_{0:T}) \propto \int_{\mathcal{X}^T} p(x_{0:T}, y_{0:T}) \mathrm{d}x_{0:i-1} \mathrm{d}x_{i+1:T} \tag{3}$$

which aim at improving previous estimates of the signal once the whole dataset is available. A further quantity of primary interest is the likelihood of the observations

$$p(y_{0:T}) = \int_{\mathcal{X}^{T+1}} p(x_{0:T}, y_{0:T}) \mathrm{d}x_{0:T}, \tag{4}$$

obtained by marginalising out the trajectory of the signal from the joint density in the integral.

Computable inference, as introduced in Chaleyat-Maurel and Genon-Catalot (2006) (cf. Introduction), requires to characterise the above distributions, for the model at hand, as finite mixtures of densities, possibly with a different number of components at different collection times. The following section identifies a set of sufficient conditions that guarantee such characterizations are available.

2.2. Sufficient conditions for computable inference

We identify a set of sufficient conditions under which (1), (2) and (3) can be written as finite mixtures of densities belonging to a given parametric family, and (4) as a finite product of mixtures. Assumptions 1 to 3 below are the same as in Papaspiliopoulos and Ruggiero (2014), who studied computable filtering, while Assumption 4 is needed for computable smoothing.

Assumption 1 (Reversibility). *The signal X_t is reversible with respect to π, i.e., the detailed balance condition $\pi(x)P_\Delta(x' \,|\, x) = \pi(x')P_\Delta(x \,|\, x')$ holds.*

Note that here π need not be a normalised density, hence formally we are not assuming stationarity of the signal. E.g., one-dimensional diffusions can have a sigma-finite reversible measure which is not the stationary measure. In particular, our results carry over to these signals, where the reversible measure can act as an improper prior, as long as the distributions in Assumption 3 below can be normalised, i.e., as long as the first conditioning on a data point yields a proper distribution.

Define now, for $K \geq 1$, the space of positive integer vectors and associated norm

$$\mathcal{M} := \mathbb{Z}_+^K = \{\, \mathbf{m} = (m_1, \ldots, m_K) : m_j \in \mathbb{Z}_+ \}, \quad |\mathbf{m}| = \sum_{i=1}^{K} m_i,$$

where $\mathbf{0}$ denotes the vector of zeros and $\mathbf{e}_i = (\delta_{ij})_{j \geq 1}$ is the canonical unit vector in the ith direction. We write $\mathbf{m} \leq \mathbf{n}$ if and only if $m_j \leq n_j$ for all j, and $\mathbf{m} - \mathbf{n}$ for the vector with jth element $m_j - n_j$. We will also need the grid of integer vectors lying below points in $\mathbf{M} \subset \mathcal{M}$, denoted by

$$\mathcal{B}(\mathbf{M}) = \{\mathbf{n} : \mathbf{n} \leq \mathbf{m}, \mathbf{m} \in \mathbf{M}\}, \tag{5}$$

and with a little abuse of notation we also let $\mathcal{B}(\mathbf{m}) := \mathcal{B}(\{\mathbf{m}\}) = \{\mathbf{n} : \mathbf{n} \leq \mathbf{m}\}$.

The second assumption provides key requirements on the existence of a certain *dual process* for the signal (a definition is given below). To this end, let, for $\Theta \subseteq \mathbb{R}^l$, $l \in \mathbb{Z}_+$, the function $r : \Theta \to \Theta$ be such that the differential equation

$$\mathrm{d}\Theta_t/\mathrm{d}t = r(\Theta_t), \quad \Theta_0 = \theta_0, \tag{6}$$

has a unique solution, denoted $\Theta_t(\theta_0)$, for all θ_0. Let $q : \mathbb{Z}_+ \to \mathbb{R}_+$ be an increasing function, $\rho : \Theta \to \mathbb{R}_+$ be a continuous function, and let (M_t, Θ_t) be a two-component process on $\mathcal{M} \times \Theta$, where Θ_t evolves autonomously according to (6), and when $(M_t, \Theta_t) = (\mathbf{m}, \theta)$, the process jumps to $(\mathbf{m} - \mathbf{e}_j, \theta)$ at rate $\rho(\theta)q(|\mathbf{m}|)m_j$. I.e., M_t is a non-homogeneous pure-death process on $\mathcal{M}$, and Θ_t a deterministic process that modulates the jump rates of M_t. The transition probabilities of M_t, denoted

$$p_{\mathbf{m},\mathbf{n}}(t;\theta) := \mathbb{P}(M_t = \mathbf{n} \,|\, M_0 = \mathbf{m}, \Theta_0 = \theta), \quad \mathbf{n} \le \mathbf{m}, \tag{7}$$

and zero otherwise, are fully described in Lemma B.1.

Assumption 2 (Duality). *There exists a family of functions $h : \mathcal{X} \times \mathcal{M} \times \Theta \to \mathbb{R}_+$ with $\sup_x h(x, \mathbf{m}, \theta) < \infty$ for all $(\mathbf{m}, \theta)$, such that*

$$\mathbb{E}[h(X_t, \mathbf{m}, \theta) \,|\, X_0 = x] = \mathbb{E}[h(x, M_t, \Theta_t) \,|\, (M_0, \Theta_0) = (\mathbf{m}, \theta)]. \tag{8}$$

When (8) holds, (M_t, Θ_t) is said to be *dual* to X_t with respect to functions h, called *duality functions*. See Jansen and Kurt (2014). The conditional expectations in (8) are taken with respect to the law of X_t on the left hand side, and with respect to that of (M_t, Θ_t) on the right hand side. When $K = 0$ or $l = 0$, the dual process is just Θ_t or M_t respectively, and we adopt the convention that $\rho(\theta) \equiv 1$ whenever $l = 0$.

The third assumption essentially amounts to what in Bayesian statistics is known as *conjugacy*.

Assumption 3 (Conjugacy). *For π as in Assumption 1 and h as in Assumption 2, the emission density $f_{x_t}(\cdot)$ is conjugate to densities in the parametric family*

$$\mathcal{F} = \{g(x, \mathbf{m}, \theta) = h(x, \mathbf{m}, \theta)\pi(x), \mathbf{m} \in \mathcal{M}, \theta \in \Theta\}, \tag{9}$$

i.e., there exist an increasing function $t : \mathcal{Y} \times \mathcal{M} \to \mathcal{M}$ and a function $T : \mathcal{Y} \times \Theta \to \Theta$ such that if $X \sim g(x, \mathbf{m}, \theta)$ and $Y \,|\, (X = x) \sim f_x(y)$, then $X \,|\, (Y = y) \sim g(x, t(y, \mathbf{m}), T(y, \theta))$.

Here $g(x, \mathbf{m}, \theta) = h(x, \mathbf{m}, \theta)\pi(x)$ is identified with the prior distribution on a hidden state X_t, and $g(x, t(y, \mathbf{m}), T(y, \theta))$ is the posterior distribution of X_t given $y \sim f_{x_t}(\cdot)$.

The fourth assumption is needed for obtaining the smoothing densities in computable form.

Assumption 4 (*h*-stability). *For h as in Assumption 2, there exist functions $d : \mathcal{M}^2 \to \mathcal{M}$ and $e : \Theta^2 \to \Theta$ such that for all $x \in \mathcal{X}$, $\mathbf{m}, \mathbf{m}' \in \mathcal{M}$, $\theta, \theta' \in \Theta$*

$$h(x, \mathbf{m}, \theta)h(x, \mathbf{m}', \theta') = C_{\mathbf{m},\mathbf{m}',\theta,\theta'} h(x, d(\mathbf{m}, \mathbf{m}'), e(\theta, \theta')), \tag{10}$$

where $C_{\mathbf{m},\mathbf{m}',\theta,\theta'}$ is constant in x.

At close inspection, Assumption 4 may appear to follow from the conjugacy in Assumption 3. In fact, when two observations Y, Y' independently give posteriors $g(x, \mathbf{m}, \theta)$ and $g(x, \mathbf{m}', \theta')$, then $g(x, d(\mathbf{m}, \mathbf{m}'), e(\theta, \theta'))$ is simply the posterior obtained by using data (Y, Y') jointly, and $C_{\mathbf{m}, \mathbf{m}', \theta, \theta'}$ simply provides a reparameterization. However, it appears not immediate that this argument is valid for any value of θ, so we state it as an assumption.

2.3. Recursive formulae

In this section we derive our theoretical results, and present general recursive formulae for updating the computable representations of (1)–(2)–(3)–(4).

2.3.1. Filtering and prediction

Denote by

$$\mu_{\mathbf{m}, \theta}(y) := \int_{\mathcal{X}} f_x(y) g(x, \mathbf{m}, \theta) \mathrm{d}x \tag{11}$$

the marginal density of $Y \sim f_x(\cdot)$ when $X \sim g(x, \mathbf{m}, \theta)$, for g in (9). For Θ_Δ as in (6), $t(\cdot, \cdot), T(\cdot, \cdot)$ as in Assumption 3 and $i = 0, \ldots, T$, define also the quantities

$$\begin{aligned}
\vartheta_{i|0:i} &:= T(y_i, \vartheta_{i|0:i-1}), & \vartheta_{i|0:i-1} &:= \Theta_\Delta(\vartheta_{i-1|0:i-1}), & \vartheta_{0|0:-1} &:= \theta_0 \\
\mathbf{M}_{i|0:i} &:= t(y_i, \mathbf{M}_{i|0:i-1}), & \mathbf{M}_{i|0:i-1} &:= \mathcal{B}(\mathbf{M}_{i-1|0:i-1}), & \mathbf{M}_{0|0:-1} &:= \{\mathbf{0}\}.
\end{aligned} \tag{12}$$

Here, $\vartheta_{i|0:i-1}$ denotes the state of the deterministic component of the dual process at time i, after the propagation from time $i - 1$ and before updating with the datum collected at time i, and $\vartheta_{i|0:i}$ the state after such update. Similarly, $\mathbf{M}_{i|0:i-1}$ denotes the set, at time i before the update, of what are called here *active indices*, i.e., the points in $\mathcal{M}$ identifying mixture components with strictly positive weight, and $\mathbf{M}_{i|0:i}$ those active after the update. Note that the quantities in (12) are deterministic and can be computed on the basis of the dataset $y_{0:T}$, by means of the update functions t, T in Assumption 3, of the solution of (6) and of (5).

The following Theorem provides the recursive formulae for prediction and filtering in (1) and (2).

Theorem 1. *Let Assumption 1 to 3 hold, and let*

$$\sum_{\mathbf{m} \in \mathbf{M}_{i-1|0:i-1}} w_{\mathbf{m}}^{(i-1)} g(x, \mathbf{m}, \vartheta_{i-1|0:i-1})$$

be the density of x_{i-1} conditional on $y_{0:i-1}$. Then (1) and (2) are the finite mixtures of densities

$$\nu_{i|0:i-1}(x) = \sum_{\mathbf{m} \in \mathbf{M}_{i|0:i-1}} w_{\mathbf{m}}^{(i-1)'} g(x, \mathbf{m}, \vartheta_{i|0:i-1}), \tag{13}$$

$$\nu_{i|0:i}(x) = \sum_{\mathbf{m} \in \mathbf{M}_{i|0:i}} w_{\mathbf{m}}^{(i)} g(x, \mathbf{m}, \vartheta_{i|0:i}) \tag{14}$$

where, for $p_{\mathbf{n},\mathbf{m}}(\Delta; \vartheta_{i|0:i})$ and $\mu_{\mathbf{m},\theta}$ as in (7) and (11), the mixture weights $w_{\mathbf{m}}^{(i-1)'}, w_{\mathbf{m}}^{(i)}$ are given by

$$
\begin{aligned}
w_{\mathbf{m}}^{(i-1)'} &= \sum_{\mathbf{n} \in \mathbf{M}_{i-1|0:i-1}: \ \mathbf{n} \geq \mathbf{m}} w_{\mathbf{n}}^{(i-1)} p_{\mathbf{n},\mathbf{m}}(\Delta; \vartheta_{i-1|0:i-1}), \quad \mathbf{m} \in \mathbf{M}_{i|0:i-1}, \\
w_{\mathbf{m}}^{(i)} &\propto \mu_{\mathbf{n}, \vartheta_{i|0:i-1}}(y_i) w_{\mathbf{n}}^{(i-1)'}, \quad \mathbf{m} = t(y_i, \mathbf{n}), \mathbf{n} \in \mathbf{M}_{i|0:i-1},
\end{aligned} \tag{15}
$$

and 0 elsewhere.

This result is due to Papaspiliopoulos and Ruggiero (2014) and included here in the present notation for ease of reference. Note that a predictive distribution for the signal at time $T + t$ for arbitrary $t > 0$ can be easily obtained from (13) by letting $\Delta = t$ in the transition probabilities $p_{\mathbf{n},\mathbf{m}}(\Delta; \vartheta_{i|0:i})$ (cf. Appendix B).

2.3.2. Likelihood

From (4), we can write the likelihood as

$$p(y_{0:T}) = \int_{\mathcal{X}^{T+1}} \nu(x_0) \prod_{i=0}^{T} f_{x_i}(y_i) \prod_{i=1}^{T} P_{\Delta_i}(x_i \mid x_{i-1}) \mathrm{d}x_{0:T}$$

by integrating the signal trajectory out of $p(y_{0:T} \mid x_{0:T})$ and using the conditional independence of y_i given x_i. Alternatively, writing

$$p(y_{0:T}) = p(y_0) \prod_{i=1}^{T} p(y_i \mid y_{0:i-1}) = \int_{\mathcal{X}} f_{x_0}(y_0) \nu(x_0) \prod_{i=1}^{T} \int_{\mathcal{X}} f_{x_i}(y_i) p(x_i \mid y_{0:i-1}),$$

highlights the dependence on the predictive densities (1).

The following theorem shows that in the present setting, the above expressions are finite products of finite mixtures of distributions in (9), marginalised over the hidden state as in (11).

Theorem 2. *Under the assumptions of Theorem 1, setting $\nu_0 = \pi$, we have*

$$p(y_{0:T}) = \mu_{\mathbf{0},\theta_0}(y_0) \prod_{i=1}^{T} \sum_{\mathbf{n} \in \mathbf{M}_{i|0:i-1}} w_{\mathbf{n}}^{(i-1)'} \mu_{\mathbf{n}, \vartheta_{i|0:i-1}}(y_i) \tag{16}$$

with $\mu_{\mathbf{n}, \vartheta_{i|0:i-1}}$ as in (11), $\vartheta_{i|0:i-1}, \mathbf{M}_{i|0:i-1}$ as in (12) and $w_{\mathbf{n}}^{(i-1)'}$ as in (15).

The proof of Theorem 2 is provided in Appendix A.

2.3.3. Smoothing

Let $0 \leq i \leq T - 1$. Bayes' Theorem and conditional independence allow to write (3) as

$$p(x_i \mid y_{0:T}) \propto p(y_{i+1:T} \mid x_i) p(x_i \mid y_{0:i}) \tag{17}$$

where the right hand side involves the filtering distribution $p(x_i \,|\, y_{0:i})$, available from Theorem 1, and the *cost-to-go function* $p(y_{i+1:T} \,|\, x_i)$, sometimes called information filter, which is the likelihood of future observations given the signal.

Denote by $\overleftarrow{\vartheta}_i, \overleftarrow{\vartheta}'_i, \overleftarrow{\mathbf{M}}_i, \overleftarrow{\mathbf{M}}'_i$ the quantities defined in (12) computed backwards. Equivalently, these are computed as in (12) with data in reverse order, i.e. using $y_{T:0}$ in place of $y_{0:T}$, namely

$$
\begin{aligned}
\overleftarrow{\vartheta}_{i|i+1:T} &= \Theta_\Delta(\overleftarrow{\vartheta}_{i+1|i+1:T}), & \overleftarrow{\vartheta}_{i|i:T} &= T(y_i, \overleftarrow{\vartheta}_{i|i+1:T}), & \overleftarrow{\vartheta}_{T|T} &= T(y_T, \theta_0) \\
\overleftarrow{\mathbf{M}}_{i|i+1:T} &= \mathcal{B}(\overleftarrow{\mathbf{M}}_{i+1|i+1:T}), & \overleftarrow{\mathbf{M}}_{i|i:T} &= t(y_i, \overleftarrow{\mathbf{M}}_{i|i+1:T}), & \overleftarrow{\mathbf{M}}_{T|T} &= \{t(y_T, \mathbf{0})\}.
\end{aligned}
\tag{18}
$$

The following proposition identifies an explicit expression for the cost-to-go function.

Proposition 3. *Let Assumptions 1–2 above hold. For all $0 \le i \le T - 1$, we have*

$$
p(y_{i+1:T} \,|\, x_i) = \sum_{\mathbf{m} \in \overleftarrow{\mathbf{M}}_{i|i+1:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+1)} h(x_i, \mathbf{m}, \overleftarrow{\vartheta}_{i|i+1:T})
$$

with

$$
\overleftarrow{w}_{\mathbf{m}}^{(i+1)} = \sum_{\substack{\mathbf{n} \in \overleftarrow{\mathbf{M}}_{i+1|i+2:T}: \\ t(y_{i+1}, \mathbf{n}) \ge \mathbf{m}}} \overleftarrow{w}_{\mathbf{n}}^{(i+2)} \mu_{\mathbf{n}, \overleftarrow{\vartheta}_{i+1|i+2:T}}(y_{i+1}) p_{t(y_{i+1}, \mathbf{n}), \mathbf{m}}(\Delta; \overleftarrow{\vartheta}_{i+1|i+1:T}) \tag{19}
$$

and $\overleftarrow{\vartheta}_{i+1}, \overleftarrow{\vartheta}_{i|i+1:T}, \overleftarrow{\mathbf{M}}_{i+1|i+1:T}, \overleftarrow{\mathbf{M}}_{i|i+1:T}$ as in (18).

The proof of Proposition 3 is provided in Appendix A.

An intuition on the above result can be obtained by considering that for $i = T - 1$, $p(y_T \,|\, x_{T-1})$ simply averages over the possible values of the signal at time T, and computes the likelihood of Y_T given such values. This would in general yield an infinite expansion based on the transition kernel of the signal. Here the duality relation in Assumption 2 allows to express this quantity as a finite linear combination of duality functions, which are in turn ratios of likelihoods over marginal likelihoods.

The following Theorem shows that under the stated assumptions, the smoothing density (17) as well takes the form of a finite mixture of densities in (9).

Theorem 4. *Let Assumption 1 to 4 above hold and let $\nu_0 = \pi$. Then, for $0 \le i \le T - 1$,*

$$
p(x_i \,|\, y_{0:T}) = \sum_{\mathbf{m} \in \overleftarrow{\mathbf{M}}_{i|i+1:T}, \ \mathbf{n} \in \mathbf{M}_{i|0:i}} w_{\mathbf{m}, \mathbf{n}}^{(i)} g(x_i, d(\mathbf{m}, \mathbf{n}), e(\overleftarrow{\vartheta}_{i|i+1:T}, \vartheta_{i|0:i})),
$$

with

$$
w_{\mathbf{m}, \mathbf{n}}^{(i)} \propto \overleftarrow{w}_{\mathbf{m}}^{(i+1)} w_{\mathbf{n}}^{(i)} C_{\mathbf{m}, \mathbf{n}, \overleftarrow{\vartheta}_{i|i+1:T}, \vartheta_{i|0:i}},
$$

$w_{\mathbf{n}}^{(i)}$ as in (15), $\overleftarrow{w}_{\mathbf{m}}^{(i+1)}$ as in (19) and $C_{\mathbf{m}, \mathbf{n}, \overleftarrow{\vartheta}_{i|i+1:T}, \vartheta_{i|0:i}}$ as in (10).

The proof of Theorem 4 is provided in Appendix A.

Here the intuition is instead that the distribution of the signal at time i is evaluated by appropriately "interpolating" its distribution at adjacent times $i-1, i+1$, given data $y_{0:i-1}$ and $y_{i+1:T}$ respectively. This interpolation is performed in principle using again the transition kernel of the signal, but it is again the duality of Assumption 2 that allows to reduce the resulting expression from a doubly infinite series to a double finite sum. Here both $\mathbf{M}_{i|0:i}$ and $\overleftarrow{\mathbf{M}}_{i|i+1:T}$, which are sets of active indices at time i, have finite cardinality, hence the smoothing densities can be computed recursively with a finite number of operations.

2.4. Illustration

We illustrate the above results for two HMMs of interest that fall in our setting: a one dimensional signal driven by a Cox–Ingersoll–Ross (CIR) diffusion (which is also a continuous-state branching process), with Poisson distributed observations; and a signal driven by a Wright–Fisher (WF) diffusion on the $(K-1)$-dimensional simplex with categorical observations. Note that our results include also finite state space models and linear Gaussian systems (cf. Papaspiliopoulos and Ruggiero, 2014), whose details for filtering and smoothing are well known and omitted here.

2.4.1. *Cox–Ingersoll–Ross signals*

The CIR process is the solution of the stochastic differential equation

$$\mathrm{d}X_t = (\delta\sigma^2 - 2\gamma X_t)\mathrm{d}t + 2\sigma\sqrt{X_t}\mathrm{d}B_t, \qquad \delta, \gamma, \sigma > 0, \tag{20}$$

whose stationary distribution is the gamma density $\mathrm{Ga}(\delta/2, \gamma/\sigma^2)$ with shape $\delta/2$ and rate γ/σ^2. A conjugate emission density $f_x(y)$ is the Poisson distribution $\mathrm{Po}(\lambda x)$ with mean λx. Choosing as prior $\pi(x) = g(x, 0, \theta_0) = \mathrm{Ga}(x; \delta/2, \gamma/\sigma^2)$, with $\theta_0 = \gamma/\sigma^2$, and letting $f_x(y)$ as above, the update of $g(x, 0, \theta_0)$ given one observation y yields $g(x, m, \theta) = g(x, t(y, 0), T(y, \theta_0)) = \mathrm{Ga}(\delta/2 + y, \gamma/\sigma^2 + \lambda)$. Thus in this case

$$h(x, m, \theta) = \frac{\Gamma(\delta/2)}{\Gamma(\delta/2 + m)}\left(\frac{\gamma}{\sigma^2}\right)^{-\delta/2}\theta^{\delta/2+m}x^m\exp\left\{-\left(\theta - \gamma/\sigma^2\right)x\right\} \tag{21}$$

together with $t(y, m) = m + y$, $T(y, \theta) = \theta + \lambda$. It can be easily verified that the marginals are Negative-Binomial densities, e.g. for y as above

$$\mu_{m,\theta}(y) = \mathrm{NB}\left(y; \frac{\delta}{2} + m, \frac{\theta}{\theta + \lambda}\right).$$

Furthermore, (20) has a dual given by a non-homogeneous death process on $\mathbb{Z}_+$ with transition probabilities

$$p_{m,m-i}(\Delta; \theta) = \mathrm{Bin}(m - i; m, \Theta_\Delta(\theta)), \qquad \Theta_\Delta(\theta) = \frac{(\gamma/\sigma^2)\theta e^{2\gamma\Delta}}{\theta e^{2\gamma\Delta} + \gamma/\sigma^2 - \theta}, \tag{22}$$

where θ in (22) is the last available value of the gamma rate parameter.

In this model, all filtering and predictive densities are finite mixtures of gamma densities. Specifically, let w_n be the weights of the predictive density at time i, given by the expression

$$\nu_{i|0:i-1}(x_i) = \sum_{0 \le n \le N_{i-1}} w_n \mathrm{Ga}(x; \delta/2 + n, \vartheta_{i|0:i-1}), \qquad N_{i-1} = \sum_{j=0}^{i-1} y_j. \quad (23)$$

Here $\vartheta_{i|0:i-1}$ is obtained as in (12) by recursively computing $\vartheta_{0|0} = T(y_0, \gamma/\sigma^2)$, $\vartheta_{1|0} = \Theta_\Delta(\vartheta_{0|0})$, $\vartheta_{1|0:1} = T(y_1, \vartheta_{1|0})$, and so on. The marginal density of a single observation y_i is obtained by integrating the emission density $\mathrm{Po}(\lambda x_i)$ with respect to (23), yielding

$$\mu_{\nu_{i|0:i-1}}(y_i) = \int_\mathcal{X} \mathrm{Po}(y_i; \lambda x_i)\nu_{i|0:i-1}(x_i) = \sum_{0 \le n \le N_{i-1}} w_n \mathrm{NB}\left(y_i; \frac{\delta}{2} + n, \frac{\vartheta_{i|0:i-1}}{\vartheta_{i|0:i}}\right), \quad (24)$$

with N_{i-1} and w_n as in (23). The density of all observations (16) is thus a product of mixtures as in (24), with weights computed recursively as described in Theorem 2. Upon observing y_i, the filtering density at time i reads

$$\begin{aligned}
\nu_{i|0:i}(x_i) &= \sum_{y_i \le m \le N_{i-1}+y_i} w_m^{(i)} \mathrm{Ga}(x_i; \delta/2 + m, \vartheta_{i|0:i}) \\
w_m^{(i)} &\propto w_n \mathrm{NB}\left(y_i; \frac{\delta}{2} + n, \frac{\vartheta_{i|0:i-1}}{\vartheta_{i|0:i}}\right), \quad m = n + y_i, \quad 0 \le n \le N_{i-1},
\end{aligned} \quad (25)$$

with w_n given in (23), and the predictive density for x_{i+1} given $y_{0:i}$ is

$$\begin{aligned}
\nu_{i+1|0:i}(x_{i+1}) &= \sum_{0 \le n \le N_i} w_n^{(i)'} \mathrm{Ga}(x_{i+1}; \delta/2 + n, \vartheta_{i+1|0:i}), \qquad N_i = N_{i-1} + y_i \\
w_n^{(i)'} &= \sum_{n \le \ell \le N_i} w_\ell^{(i)} p_{\ell,n}(\Delta; \vartheta_{i|0:i}),
\end{aligned} \quad (26)$$

with $p_{\ell,m}$ and $\Theta_\Delta(\theta)$ as in (22) and $w_\ell^{(i)}$ as in (25).

Additionally, Assumption 4 is satisfied with $d(m_1, m_2) = m_1 + m_2, e(\theta_1, \theta_2) = \theta_1 + \theta_2 - \gamma/\sigma^2$ and

$$C_{m_1,m_2,\theta_1,\theta_2} = \Gamma(\delta/2)\left(\frac{\gamma}{\sigma^2}\right)^{-\delta/2} \frac{\Gamma(\frac{\delta}{2} + m_1 + m_2)}{\Gamma(\frac{\delta}{2} + m_1)\Gamma(\frac{\delta}{2} + m_2)} \frac{(\theta_1)^{\delta/2+m_1}(\theta_2)^{\delta/2+m_2}}{(\theta_1 + \theta_2 - \gamma/\sigma^2)^{\delta/2+m_1+m_2}}.$$

It is easy to show that, with $\theta_0 = \gamma/\sigma^2$, then $\theta_1, \theta_2, e(\theta_1, \theta_2) \ge \gamma/\sigma^2$. Hence, the marginal smoothing density of x_i given $y_{0:T}$ can be obtained by combining the filtering density $\nu_{i|0:i}(x_i)$ and the cost-to-go function $p(y_{i+1:T}|x_i)$. The latter is

$$p(y_{i+1:T} \mid x_i) = \sum_{0 \le m \le M_{i+1}} \overleftarrow{w}_m^{(i+1)} h(x_i, m, \vartheta'), \qquad M_{i+1} = \sum_{j=i+1}^T y_i, \qquad (27)$$

with h as in (21) and $\vartheta' = \overleftarrow{\vartheta}_{i|i+1:T}$ as in (18). The weights in (27) are obtained from those of $p(y_{i+2:T}|x_{i+1})$, denoted $\overleftarrow{w}_n^{(i+2)}$, as

$$\overleftarrow{w}_m^{(i+1)} = \sum_{n:\,n+y_{i+1}\geq m} \overleftarrow{w}_n^{(i+2)}\mathrm{NB}\left(y_{i+1};\frac{\delta}{2}+n,\frac{\vartheta''}{\vartheta''+\lambda}\right)\mathrm{Bin}(m;n+y_{i+1},\vartheta')$$

with $\vartheta'' = \overleftarrow{\vartheta}_{i+1|i+2:T}$. The smoothing density is therefore the finite mixture of Gamma densities

$$p(x_i\,|\,y_{0:T}) = \sum_{y_i\leq n\leq N_i}\sum_{0\leq m\leq M_{i+1}} w_{m,n}^{(i)}\mathrm{Ga}\left(x_i;\frac{\delta}{2}+m+n,\vartheta'+\vartheta-\gamma/\sigma^2\right)$$

with

$$w_{m,n}^{(i)} \propto \overleftarrow{w}_m^{(i+1)}w_n^{(i)}C_{m,\ell,\vartheta',\vartheta}$$

where $\overleftarrow{w}_m^{(i+1)}$ is as in (27) and $w_n^{(i)}$ as in (25).

2.4.2. *Wright–Fisher signals*

The K-component WF model is a diffusion process taking values in the simplex of nonnegative vectors $\mathbf{x} = (x_1,\ldots,x_K)$ whose coordinates sum up to one (later simply called the simplex). It is characterised by its infinitesimal operator

$$\mathcal{A} = \frac{1}{2}\sum_{i=1}^{K}(\alpha_i - |\boldsymbol{\alpha}|x_j)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{K} x_i(\delta_{ij}-x_j)\frac{\partial^2}{\partial x_i\partial x_j}, \quad |\boldsymbol{\alpha}| = \sum_{j=1}^{K}\alpha_j,$$

where $\alpha_j > 0$ for all j, whose domain is the class of twice differentiable functions on the simplex. This diffusion is stationary and reversible with respect to the Dirichlet distribution with parameter $\boldsymbol{\alpha}$ whose density on the simplex with respect to the Lebesgue measure is proportional to $\prod_{j=1}^{K} x_j^{\alpha_j}$. See Ethier and Kurtz (1986).

A distribution conjugate to the Dirichlet, seen as prior density for a simplex-valued variable, is the Multinomial distribution, denoted here $\mathrm{MN}(\mathbf{y};|\mathbf{y}|,\mathbf{x})$, where x_j is the probability of drawing category j in a sample of size $|\mathbf{y}|$ and $\mathbf{y} = (y_1,\ldots,y_K)$ are the multiplicities being drawn. Upon observing $\mathbf{y}$, the Dirichlet density is updated by replacing $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}+\mathbf{y}$, i.e., α_j with α_j+y_j for $j = 1,\ldots,K$. Moreover, the WF signal is known to be dual to a death process D_t on $\mathbb{Z}_+^K$ that jumps from $\mathbf{m}$ to $\mathbf{m}-\mathbf{e}_j$ at rate $m_j(|\boldsymbol{\alpha}|+|\mathbf{m}|-1)/2$, with respect to functions

$$h(\mathbf{x},\mathbf{m}) = \frac{\Gamma(|\boldsymbol{\alpha}|+|\mathbf{m}|)}{\Gamma(|\boldsymbol{\alpha}|)}\prod_{j=1}^{K}\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j+m_j)}x_j^{m_j}.$$

This dual process has no deterministic component, and its transition probabilities $p_{\mathbf{m},\mathbf{n}}(\Delta)$ are obtained by specialising Lemma B.1 in Appendix B to the

case $\rho \equiv 1$. It can be easily verified that the marginal distribution of $\mathbf{y}$ is a Dirichlet-Multinomial density, e.g., given the parameter $\boldsymbol{\alpha} + \mathbf{m}$,

$$\mu_{\mathbf{m}}(\mathbf{y}) = \mathrm{DM}(\mathbf{y}; \boldsymbol{\alpha} + \mathbf{m}) := \binom{|\mathbf{y}|}{\mathbf{y}} \frac{\prod_{j=1}^{K}(\alpha_j + m_j)_{(y_j)}}{(|\boldsymbol{\alpha} + \mathbf{m}|)_{(|\mathbf{y}|)}},$$

where $a_{(n)} = a(a+1)\cdots(a+n-1)$ is the Pochhammer symbol. In this model, all filtering and predictive densities are finite mixtures of Dirichlet densities. Specifically, let the density of the signal at time i, before conditioning on $\mathbf{y}_i$, be

$$\nu_{i|0:i-1}(\mathbf{x}_i) = \sum_{\mathbf{0} \leq \mathbf{n} \leq \mathbf{N}_{i-1}} w_{\mathbf{n}}\mathrm{Dir}(\mathbf{x}_i; \boldsymbol{\alpha} + \mathbf{n}), \qquad \mathbf{N}_{i-1} = \sum_{j=0}^{i-1} \mathbf{y}_j.$$

The marginal density of $\mathbf{y}_i$ is obtained by integrating the Multinomial emission density with respect to (23), yielding

$$\mu_{\nu_{i|0:i-1}}(\mathbf{y}_i) = \int_{\mathcal{X}} \mathrm{MN}(\mathbf{y}_i; |\mathbf{y}_i|, \mathbf{x})\nu_{i|0:i-1}(\mathbf{x}_i) = \sum_{\mathbf{0} \leq \mathbf{n} \leq \mathbf{N}_{i-1}} w_{\mathbf{n}}\mathrm{DM}(\mathbf{y}; \boldsymbol{\alpha} + \mathbf{n}),$$

with $\mathbf{N}_{i-1}$ and $w_{\mathbf{n}}$ as above, and the density of all observations (16) is thus a product of such mixtures, with weights computed recursively. Upon observing $\mathbf{y}_i$, the filtering density at time i reads

$$\begin{aligned}
\nu_{i|0:i}(\mathbf{x}_i) &= \sum_{\mathbf{y}_i \leq \mathbf{m} \leq \mathbf{N}_{i-1}+\mathbf{y}_i} w_{\mathbf{m}}^{(i)}\mathrm{Dir}(\mathbf{x}_i; \boldsymbol{\alpha} + \mathbf{m}) \\
w_{\mathbf{m}}^{(i)} &\propto w_{\mathbf{n}}\mathrm{DM}(\mathbf{y}; \boldsymbol{\alpha} + \mathbf{n}), \quad \mathbf{m} = \mathbf{n} + \mathbf{y}_i, \quad \mathbf{0} \leq \mathbf{n} \leq \mathbf{N}_{i-1},
\end{aligned} \tag{28}$$

and the predictive density for $\mathbf{x}_{i+1}$ given $\mathbf{y}_{0:i}$ is

$$\begin{aligned}
\nu_{i+1|0:i}(\mathbf{x}_{i+1}) &= \sum_{\mathbf{0} \leq \mathbf{n} \leq \mathbf{N}_i} w_{\mathbf{n}}^{(i)'}\mathrm{Dir}(\mathbf{x}_{i+1}; \boldsymbol{\alpha} + \mathbf{n}), \qquad \mathbf{N}_i = \mathbf{N}_{i-1} + \mathbf{y}_i, \\
w_{\mathbf{n}}^{(i)'} &= \sum_{\mathbf{n} \leq \mathbf{l} \leq \mathbf{N}_i} w_{\mathbf{l}}^{(i)} p_{\mathbf{l},\mathbf{n}}(\Delta),
\end{aligned}$$

with $w_{\mathbf{l}}^{(i)}$ as in (28). Additionally, Assumption 4 for $h(\mathbf{x}, \mathbf{m})$ is satisfied with $d(\mathbf{m}_1, \mathbf{m}_2) = \mathbf{m}_1 + \mathbf{m}_2$ and

$$C_{\mathbf{m}_1,\mathbf{m}_2} = \frac{\Gamma(|\boldsymbol{\alpha} + \mathbf{m}_1|)\Gamma(|\boldsymbol{\alpha} + \mathbf{m}_2|)}{\Gamma(|\boldsymbol{\alpha}|)\Gamma(|\boldsymbol{\alpha} + \mathbf{m}_1 + \mathbf{m}_2|)} \prod_{j=1}^{K} \frac{\Gamma(\alpha_j)\Gamma(\alpha_j + m_{j1} + m_{j2})}{\Gamma(\alpha_j + m_{j1})\Gamma(\alpha_j + m_{j2})},$$

and the cost-to-go function $p(\mathbf{y}_{i+1:T}|\mathbf{x}_i)$ is

$$\begin{aligned}
p(\mathbf{y}_{i+1:T}\,|\,\mathbf{x}_i) &= \sum_{\mathbf{0} \leq \mathbf{m} \leq \mathbf{M}_{i+1}} \overleftarrow{w}_{\mathbf{m}}^{(i+1)} h(\mathbf{x}_i, \mathbf{m}), \qquad \mathbf{M}_{i+1} = \sum_{j=i+1}^{T} \mathbf{y}_i, \\
\overleftarrow{w}_{\mathbf{m}}^{(i+1)} &= \sum_{\mathbf{n}:\, \mathbf{n}+\mathbf{y}_i \geq \mathbf{m}} \overleftarrow{w}_{\mathbf{n}}^{(i+2)} \mathbf{DM}(\mathbf{y}_{i+1}; \boldsymbol{\alpha} + \mathbf{n})p_{\mathbf{n}+\mathbf{y}_i,\mathbf{m}}(\Delta).
\end{aligned} \tag{29}$$

Therefore the smoothing density is the finite mixture of Dirichlet densities

$$p(\mathbf{x}_i \,|\, \mathbf{y}_{0:T}) = \sum_{\mathbf{y}_i \leq \mathbf{n} \leq \mathbf{N}_{i-1}} \sum_{\mathbf{0} \leq \mathbf{m} \leq \mathbf{M}_{i+1}} w_{\mathbf{m},\mathbf{n}}^{(i)} \mathrm{Dir}(\mathbf{x}_i; \boldsymbol{\alpha} + \mathbf{m} + \mathbf{n}),$$

with

$$w_{\mathbf{m},\mathbf{n}}^{(i)} \propto \overleftarrow{w}_{\mathbf{m}}^{(i+1)} w_{\mathbf{n}}^{(i)} C_{\mathbf{m},\mathbf{n}}$$

where $\overleftarrow{w}_m^{(i+1)}$ is as in (29) and $w_n^{(i)}$ as in (28).

3. Simulation from the joint smoothing distribution

The results presented in Section 2 can be elaborated to provide methods to simulate trajectories of the signal conditional on the entire dataset. These can be interpreted as samples from the posterior distribution of the signal, evaluated at the skeleton of observation times. These draws in turn can be used to estimate the model parameters through an MCMC strategy. In particular, the availability of a method for sampling conditional trajectories opens the way to performing full inference on the model parameters, using a Metropolis-within-Gibbs algorithm, with a Gibbs step for the trajectory and a Metropolis–Hastings step for the parameters. We implement such strategy in Section 5.3.

In this framework, two decompositions are in principle available for the joint conditional density of the signal. The *backward* decomposition reads

$$p(x_{0:T}|y_{0:T}) = p(x_T|y_{0:T}) \prod_{i=0}^{T-1} p(x_i|x_{i+1}, y_{0:T}) = p(x_T|y_{0:T}) \prod_{i=0}^{T-1} p(x_i|x_{i+1}, y_{0:i})$$

where the backward kernel, in virtue of Bayes' Theorem, can be written

$$p(x_i|x_{i+1}, y_{0:i}) = \frac{p(x_{i+1}|x_i)p(x_i|y_{0:i})}{p(x_{i+1}|y_{0:i})} = \frac{P_\Delta(x_{i+1}|x_i)\nu_{i|0:i}(x_i)}{\nu_{i+1|0:i}(x_{i+1})} \qquad (30)$$

and $\nu_{i|0:i}$ and $\nu_{i+1|0:i}$ are the filtering and prediction densities as in (13)–(14). The typical approach to smoothing using this decomposition is generally known as the Forward Filtering Backward Sampling (FFBS) algorithm (cf. Chopin and Papaspiliopoulos, 2020; Cappé et al., 2005), which consists in first implementing a forward pass by filtering the whole signal, and then generating a draw of the conditional trajectory by means of a backward pass which starts from a sample from the last filtering distribution and uses (30) repeatedly.

An alternative approach uses the *forward* decomposition

$$p(x_{0:T}|y_{0:T}) = p(x_0|y_{0:T}) \prod_{i=i}^{T} p(x_i|x_{i-1}, y_{0:T}) = p(x_0|y_{0:T}) \prod_{i=0}^{T} p(x_i|x_{i-1}, y_{i:T}),$$

where the forward kernel, again by Bayes' Theorem, can be written

$$p(x_i|x_{i-1}, y_{i:T}) = \frac{p(x_i|x_{i-1})p(y_{i:T}|x_i)}{p(y_{i:T}|x_{i-1})} = \frac{P_\Delta(x_i|x_{i-1})p(y_i|x_i)p(y_{i+1:T}|x_i)}{p(y_{i:T}|x_{i-1})}.$$
$$(31)$$

The cost-to-go functions appearing on the right hand side are available through Proposition 3. Here the approach to obtaining a sample from the joint smoothing distribution is to compute the cost-to-go functions backward and to perform forward sampling.

In both cases, the finite mixture representations obtained in Section 2.3 (filtering density for the backward decomposition, cost-to-go functions for the forward decomposition) need to be combined with the transition density of the signal, so further analysis on how to simulate in pratice from $p(x_i|x_{i+1}, y_{0:i})$ or $p(x_i|x_{i-1}, y_{i:T})$ depends on the specific model. We show how this can be done for the CIR and WF model.

3.1. Cox–Ingersoll–Ross signals

Under the specifications of Section 2.4.1, the unconditional transition density of the signal can be written as the infinite mixture of gamma densities

$$P_\Delta(x_{i+1}|x_i) = \sum_{k \geq 0} \mathrm{Po}(k; \Theta'_\Delta x_i) \mathrm{Ga}(x_{i+1}; \delta/2 + k, e^{2\gamma\Delta}\Theta'_\Delta), \qquad \Theta'_\Delta = \frac{\gamma/\sigma^2}{e^{2\gamma\Delta} - 1},$$

which implies, using (30), that

$$p(x_i|x_{i+1}, y_{0:i}) = \frac{\sum_{k \geq 0} \mathrm{Ga}(x_{i+1}; \delta/2 + k, e^{2\gamma\Delta}\Theta'_\Delta)\mathrm{Po}(k; \Theta'_\Delta x_i)\nu_{i|0:i}(x_i)}{\nu_{i+1|0:i}(x_{i+1})}.$$

When $\nu_{i|0:i}(x_i)$ is as in (25), the terms in x_i yield

$$\mathrm{Po}(k; \Theta'_\Delta x_i)\nu_{i|0:i}(x_i) = \sum_{y_i \leq m \leq N_i} w_m^{(i)} \mu_{m, \vartheta_{i|0:i}}(k) \mathrm{Ga}(x_i; \delta/2 + m + k, \vartheta_{i|0:i} + \Theta'_\Delta)$$

where we have used Bayes' Theorem to exploit the conjugacy of the Gamma–Poisson model (cf. Section 2.4.1) and where

$$\mu_{m, \vartheta_{i|0:i}}(k) = \mathrm{NB}\left(k; \frac{\delta}{2} + m, \frac{\vartheta_{i|0:i}}{\vartheta_{i|0:i} + \Theta'_\Delta}\right)$$

is the marginal distribution of k when $k|x_i \sim \mathrm{Po}(k; \Theta'_\Delta x_i)$ and $x_i \sim \mathrm{Ga}(x_i; \delta/2 + m, \vartheta_{i|0:i})$. It follows that

$$p(x_i|x_{i+1}, y_{0:i}) = \sum_{k \geq 0} \tilde{w}_k(x_{i+1}, y_{0:i}) \sum_{y_i \leq m \leq N_i} \hat{w}_{m,k}^{(i)} \mathrm{Ga}(x_i; \delta/2 + m + k, \vartheta_{i|0:i} + \Theta'_\Delta) \tag{32}$$

where

$$\tilde{w}_k(x_{i+1}, y_{0:i}) = \frac{\mathrm{Ga}(x_{i+1}; \delta/2 + k, e^{2\gamma\Delta}\Theta'_\Delta)}{\nu_{i+1|0:i}(x_{i+1})}, \qquad \hat{w}_{m,k}^{(i)} = w_m^{(i)} \mu_{m, \vartheta_{i|0:i}}(k),$$

and $\nu_{i+1|0:i}(x_{i+1})$ is as in (26).

Equation (32) is again an infinite mixture of gamma densities, implying that a simple algorithm for simulating from $p(x_i|x_{i+1}, y_{0:i})$ is to sample the double index (k, m) from the discrete mixture with probability masses equal to the weights in (32), and then sample x_i from a gamma distribution whose parameters are determined by the drawn (k, m). A schematic version of this strategy is as follows:

- draw $U \sim \mathrm{U}(0, 1)$
- set κ, M to be the minimum values j, p that yield

$$U \leq \sum_{k \leq j-1} \tilde{w}_k(x_{i+1}, y_{0:i}) \sum_{y_i \leq m \leq N_i} \hat{w}_{m,k}^{(i)} + \tilde{w}_j(x_{i+1}, y_{0:i}) \sum_{y_i \leq m \leq p} \hat{w}_{m,j}^{(i)}$$

- draw $X_i \sim \mathrm{Ga}(x_i; \delta/2 + \kappa + M, \theta + \lambda + \Theta'_\Delta)$.

3.2. Wright–Fisher signals

A similar strategy could in principle be applied to the WF model as well, under the specifications of Section 2.4.2. The unconditional transition density of the WF diffusion can be written (cf., e.g., Ethier and Griffiths, 1993)

$$P_\Delta(\mathbf{x}_i|\mathbf{x}_{i-1}) = \sum_{m=0}^\infty q_m \sum_{\mathbf{l} \in \mathbb{N}^K: \; l_1+\ldots+l_K=m} \mathrm{MN}\,(\mathbf{l}|m, \mathbf{x}_{i-1})\,\mathrm{Dir}\,(\mathbf{x}_i|\mathbf{l}+\mathbf{x}_{i-1}). \tag{33}$$

However, the weights q_m, determined in Griffiths (1980); Tavaré (1984), have an infinite expansion, resulting in an intractable doubly infinite series for the transition density. Recently, Jenkins and Spanò (2017) devised a strategy for sampling exactly from (33), by means of an algorithm that deals with alternated weights that decrease only after a certain index. Leveraging on their results, we are able to obtain an algorithm for sampling from the conditional kernel. First we provide the following representation.

Theorem 5. *Let $\mathbf{x}_{0:T}$ be as in Section 2.4.2, and let $\mathbf{y}_{0:T}$ come from a multinomial emission density. Then (31) can be written as the mixture of Dirichlet densities*

$$p(\mathbf{x}_i|\mathbf{x}_{i-1}, \mathbf{y}_{i:T}) = \sum_{m=0}^\infty \tilde{q}_m \sum_{\substack{\mathbf{l} \in \mathbb{N}^K: \\ l_1+\ldots+l_K=m}} \tilde{w}_{\mathbf{l}}^m \sum_{\mathbf{k} \in \overleftarrow{\mathbf{M}}_{i|i+1:T}} \tilde{w}_{\mathbf{l},\mathbf{k}} \mathrm{Dir}(\mathbf{x}_i|\mathbf{l}+\mathbf{x}_{i-1}+\mathbf{y}_i+\mathbf{k}).$$
$$\tag{34}$$

Explicit expressions for the weights $\tilde{q}_m, \tilde{w}_{\mathbf{l}}^m$ and $\tilde{w}_{\mathbf{l},\mathbf{k}}$ in (34), together with a proof of the statement, are provided in Appendix A. The two inner sums in the above representation are finite mixtures from which sampling is straightforward. The final step is to sample from the infinite mixture with weights $\tilde{q}_m$. The strategy proposed in Jenkins and Spanò (2017) can be adapted to this aim, upon observing that we can bracket the discrete density of m by two convergent series and use the strategy for dealing with alternated series from Devroye

([1986](#)). A detailed explanation and an explicit sampling algorithm are provided in Appendix A. Note that this approach also underlies the foundational work in Beskos et al. ([2006](#)) for simulation of diffusion processes.

4. Implementation and acceleration

4.1. Computational challenges

The expressions we have obtained in Section 2.3 are similar in spirit to the well-known Baum–Welch algorithms (cf. Introduction), with the difference that in our case the finite state space is a subset of $\mathbb{Z}_+^K$ of whose dimension varies instead of being fixed. In practice, each component in the involved mixture distributions is associated to a node of the grid in $\mathbb{Z}_+^K$ given by the current set of active indices. As already noted by Chaleyat-Maurel and Genon-Catalot ([2006](#)), each update operation with new data increases the nominal number of components in the filtering mixture distribution by shifting the mixture to a higher set of nodes, where the upward shift represents the accumulation of further information, associating null weights to all nodes below the shifted set of active indices. All these nodes then become active indices with the prediction operation, thus increasing the effective number of mixture components (those with non null weight). A similar intuition applies to the evaluation of the cost-to-go functions, which are combined with the filtering distributions to form the marginal smoothing distributions.

More formally, for the two models illustrated in Section 2.4, the number of mixture components in the filtering distributions evolves as $|\mathbf{M}_n| = \prod_{k=0}^{K}(1 + \sum_{i=1}^{n} y_{i,k})$, where $y_{i,k}$ is the k-th coordinate of $\mathbf{y}_i$ observed at time i and K is the dimension of the signal space, whence the polynomial complexity mentioned in the Introduction. On the implementation side, an important aspect to note is that the prediction step is more expensive than the update step, as at each iteration it involves computing the probability that the dual process reaches any point $\mathbf{m}$ in $\mathcal{B}(\mathbf{M}_{i|0:i})$ from any point in $\mathbf{M}_{i|0:i}$ which is not lower than $\mathbf{m}$ (cf. (5), (12) and Theorem 1). When new data arrive at a constant frequency, it is possible to limit the cost of the prediction operation by storing the transition terms $p_{\mathbf{m},\mathbf{n}}(\Delta)$, which can be used multiple times during the successive iterations. Nonetheless, the rapid growth in their number renders this strategy daunting in terms of memory storage. A close inspection reveals that the $p_{\mathbf{m},\mathbf{n}}(\Delta)$ are themselves a product of a smaller number of terms (cf. Lemma B.1 in Appendix B). This number grows only quadratically with the sum of all observations and can be saved, thus improving the computational efficiency.

A major point that allows to substantially improve the computation is however given by the fact that each transition of the death process will assign a non negligible probability weight only to a small number of nodes, the intuition being that as the time interval becomes large, the probability mass progressively concentrates on the grid origin $(0, \ldots, 0) \in \mathbb{Z}_+^K$. Cf. Figure 4 in Ascolani et al. ([2020](#)). As a result, the number of mixture components in the filtering distributions with non negligible weight may be roughly stationary or increase at a much

lower rate, depending on the ratio of the time lag between observations and the number of observations per collection time. Cf. Figure 1 here and Section 4 in Papaspiliopoulos and Ruggiero (2014). A similar behaviour is expected from the cost-to-go functions, which are a linear combination of functions with many negligible coefficients relative to the largest. This suggest that a great computational improvement can be obtained by informedly pruning the mixtures components with negligible weight, a strategy which we develop in Section 4.2.

Another hurdle in the computation, of a more technical flavour, is due to the alternating signs appearing in the death process transition probabilities (cf. Lemma B.1 in Appendix B), susceptible to both over and underflow. Here we solve this challenge to numerical stability using the `Nemo.jl` library for arbitrary precision computation (Fieker et al., 2017).

4.2. Pruning

Based on the intuitions and experimental observations illustrated in the previous section, we consider three pruning strategies:

1. *Fixed number*: retain only a fixed number of components after ranking them by weight. This aims at controlling the computational budget at each iteration.
2. *Fixed mass*: retain the minimum number of components needed to reach a fixed fraction of the total mass. This aims at controlling the total approximation error incurred by the pruning, and in fact is closely linked to the total variation distance between the exact and the pruned mixture.
3. *Fixed threshold*: retain only the components whose weight is above a fixed threshold. This approach has the advantage of not requiring to rank the mixture components before pruning. However, it not obvious how to choose a threshold for the cost-to-go functions, whose coefficients do not sum up to one.

The pruning operation should be performed at each time step, followed by a renormalisation of the remaining weights (in the case of the filtering distributions),. Although the overall approximation error is the result of several previous approximations, the error accumulation is counterbalanced by the incorporation of fresh information on the data generating distribution at each collection time. In fact, the pruning should be performed after the update with the incoming data, which concentrates the mass of the mixture closer to the true data generating mechanism, thus likely on a fewer number of components. This has the further advantage of reducing the computational cost of the prediction step, which is more expensive than the update. Hence pruning after the update entails the maximal computational gain as it retains the required mass through the minimal number of components and reduces the number of transitions to be computed. Note that since the likelihood is obtained by means of the filtering weights (cf. Section 2.3.2), the above strategy essentially applies to the computation of the likelihood as well.

Algorithm 1: Filtering

Pruning setting: ON (approximate filtering) / OFF (exact filtering)
Input: $Y_{0:n}$, $t_{0:n}$ and $\nu = h(x, \mathbf{0}, \theta_0) \in \mathcal{F}$ for some $\theta_0 \in \Theta$
Result: $\vartheta_{i|0:i}$, $\mathbf{M}_{i|0:i}$ and $W_{0:n}$ with $W_i = \{w_{\mathbf{m}}^i, \mathbf{m} \in \mathbf{M}_{i|0:i}\}$
Initialise
    Set $\vartheta_{0|0} = T(Y_0, \theta_0)$ with T as in (12)
    Set $\mathbf{M}_{0|0} = \{t(Y_0, \mathbf{0})\} = \{\mathbf{m}^*\}$ and $W_0 = \{1\}$ with t as in Assumption 3
    Compute $\vartheta_{1|0}$ from $\vartheta_{0|0}$ as in (12)
    Set $\mathbf{M}^* = \mathcal{B}(\mathbf{M}_{0|0})$ and $W^* = \{p_{\mathbf{m}^*, \mathbf{n}}(\Delta, \vartheta_{0|0}), \mathbf{n} \in \mathbf{M}^*\}$ with $\mathcal{B}$ as in (5) and $p_{\mathbf{m}, \mathbf{n}}$
    as in (B40)

for *i from* 1 *to* n **do**
    Update
        Set $\vartheta_{i|0:i} = T(Y_i, \vartheta_{i|0:i-1})$
        Set $W_i = \{\frac{w_{\mathbf{m}}^* \mu_{\mathbf{m}, \vartheta_{i|0:i-1}}(Y_i)}{\sum_{\mathbf{n} \in \mathbf{M}^*} w_{\mathbf{n}}^* \mu_{\mathbf{n}, \vartheta_i}(Y_i)}, \mathbf{m} \in \mathbf{M}^*\}$ with $\mu_{\mathbf{m}, \theta}$ defined as in (11)
        Set $\mathbf{M}_{i|0:i} = \{t(Y_i, \mathbf{m}), \mathbf{m} \in \mathbf{M}^*\}$ and update the labels in W_i
        Copy $\vartheta_{i|0:i}$, $\mathbf{M}_{i|0:i}$ and W_i to be reported as the output

    if *pruning* ON **then**
        Prune $\mathbf{M}_{i|0:i}$ and remove the corresponding weights in W_i
        Normalise the weights in W_i

    Predict
        Compute $\vartheta_{i+1|0:i}$ from $\vartheta_{i|0:i}$

        Set $\mathbf{M}^* = \mathcal{B}(\mathbf{M}_{i|0:i})$ and $W^* = \left\{ \displaystyle\sum_{\mathbf{m} \in \mathbf{M}_{i|0:i}, \mathbf{m} \geq \mathbf{n}} w_{\mathbf{m}}^i p_{\mathbf{m}, \mathbf{n}}(\Delta, \vartheta_{i|0:i}), \mathbf{n} \in \mathbf{M}^* \right\}$

end

We lay out two practical algorithms for implementing the recursions of Section 2.3, both with the option of pruning. If kept off, the algorithms thus provide an exact evaluation of the respective distributions. The pseudo code for computing the filtering densities is provided in Algorithm 1, while that for the smoothing densities in Algorithm 2. Algorithm B4 in Appendix B illustrates how to modify Algorithm 1 to compute the likelihood.

5. Numerical experiments

5.1. Retrieving the signal

We illustrate inference on the signal trajectory for the models illustrated in Section 2.4. We reparametrise the CIR model by considering

$$\mathrm{d}X_t = a(b - X_t)\,\mathrm{d}t + s\sqrt{X_t}\,\mathrm{d}B_t \tag{35}$$

in place of (20), where b is the mean, a the speed of adjustment towards the mean and s controls the volatility, corresponding to $a = 2\gamma, b = \delta\sigma^2/(2\gamma)$ and

---

Algorithm 2: Smoothing

Pruning setting: ON (approximate smoothing) / OFF (exact smoothing)

Input: $Y_{0:n}$, $t_{0:n}$ and $\nu = h(x, \mathbf{0}, \theta_0) \in \mathcal{F}$ for some $\theta_0 \in \Theta$

Result: $\vartheta_{i|0:n}$, $\mathbf{M}_{i|0:n}$ and $W_{0:n|0:n}$ with $W_{i|0:n} = \{w_{\mathbf{m}}^i, \mathbf{m} \in \mathbf{M}_{i|0:n}\}$

Initialise

    Obtain $\vartheta_{i|0:i}$, $\mathbf{M}_{i|0:i}$ and $W_{0:n}$ with $W_i = \{w_{\mathbf{m}}^i, \mathbf{m} \in \mathbf{M}_{i|0:i}\}$ from the filtering algorithm, with pruning if relevant.

    Set $\overleftarrow{\vartheta}_{n|n} = T(Y_n, \theta_0)$ with T as in (12)

    Set $\overleftarrow{\mathbf{M}}_{n|n} = \{t(Y_n, \mathbf{0})\} = \{\mathbf{m}^*\}$ and $\overleftarrow{W}_n = \{1\}$ with t as in Assumption 3

    Compute $\overleftarrow{\vartheta}_{n-1|n}$ from $\overleftarrow{\vartheta}_{n|n}$ as in (12)

    Set $\overleftarrow{\mathbf{M}}^* = \mathcal{B}(\overleftarrow{\mathbf{M}}_{n|n})$ and $\overleftarrow{W}^* = \{p_{\mathbf{m}^*, \mathbf{n}}(\Delta, \overleftarrow{\vartheta}_{n|n}), \mathbf{n} \in \overleftarrow{\mathbf{M}}^*\}$ with $\mathcal{B}$ as in (5) and $p_{\mathbf{m}, \mathbf{n}}$ as in (B40)

for *i from $n-1$ to 1* **do**

    Update

        Set $\overleftarrow{\vartheta}_{i|i:n} = T(Y_i, \vartheta_{i|i+1:n})$

        Set $\overleftarrow{W}_i = \{w_{\mathbf{m}}^* \mu_{\mathbf{m}, \overleftarrow{\vartheta}_{i|i+1:T}(Y_i)}, \mathbf{m} \in \overleftarrow{\mathbf{M}}^*\}$ with $\mu_{\mathbf{m}, \theta}$ defined as in (11)

        Set $\overleftarrow{\mathbf{M}}_{i|i:n} = \{t(Y_i, \mathbf{m}), \mathbf{m} \in \mathbf{M}^*\}$

    if *pruning* ON **then**

        Prune $\overleftarrow{\mathbf{M}}_{i|i:n}$ and remove the corresponding values in $\overleftarrow{W}_i$

    Predict

        Compute $\overleftarrow{\vartheta}_{i-1|i:n}$ from $\overleftarrow{\vartheta}_{i|i:n}$

        Set $\overleftarrow{\mathbf{M}}^* = \mathcal{B}(\overleftarrow{\mathbf{M}}_{i|i:n})$ and $\overleftarrow{W}^* = \left\{ \displaystyle\sum_{\mathbf{m} \in \overleftarrow{\mathbf{M}}_{i|i:n}, \mathbf{m} \geq \mathbf{n}} w_{\mathbf{m}}^i p_{\mathbf{m}, \mathbf{n}}(\Delta, \overleftarrow{\vartheta}_{i|i:n}), \mathbf{n} \in \overleftarrow{\mathbf{M}}^* \right\}$

end

for *i from 1 to $n-1$* **do**

    Set $\vartheta_{i|0:n} = e(\overleftarrow{\vartheta}_{i|i+1:n}, \vartheta_{i|0:i})$

    Set $\mathbf{M}_{i|0:n} = \{d(\mathbf{m}, \mathbf{n}); \mathbf{m} \in \overleftarrow{\mathbf{M}}_{i|i+1:n}, \mathbf{n} \in \mathbf{M}_{i|0:i}\}$ considering only the unique values.

    Set
    $W_{i|0:n} = \{w_{\mathbf{l}}^i \propto \displaystyle\sum_{\substack{\mathbf{m} \in \overleftarrow{\mathbf{M}}_{i|i+1:n}, \mathbf{n} \in \mathbf{M}_{i|0:i}, \\ d(\mathbf{m}, \mathbf{n}) = \mathbf{l}}} \overleftarrow{w}_{\mathbf{m}}^{(i+1)} w_{\mathbf{n}}^{(i)} C_{\mathbf{m}, \mathbf{n}, \overleftarrow{\vartheta}_{i|i+1:n}, \vartheta_{i|0:i}}, \mathbf{l} \in \mathbf{M}_{i|0:n}\}$

end

$s = 2\sigma$. We simulate the trajectory of a CIR process starting from $X_0 = 3$ with $a = 5$, $b = 9.6$ and $s = 8$, corresponding to a Gamma$(1.5, 0.15625)$ stationary distribution. We draw 10 observations at each of 200 collection times separated by 0.011 seconds.

For the WF, we implement[3] the exact simulation scheme of Jenkins and Spanò (2017) to simulate a signal with $K = 3$ components, initialising the process at random from a Dirichlet$(0.3, 0.3, 0.3)$ distribution, and draw 15 ob-

---

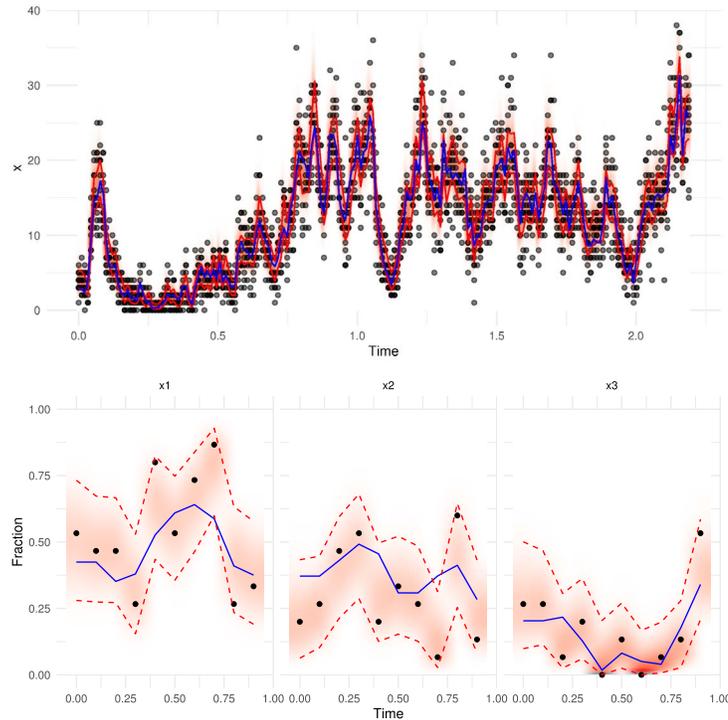[3] Our code is available at https://github.com/konkam/ExactWrightFisher.jl

FIG 3. *Hidden signal (blue solid), data points (bullets) and 95% pointwise credible intervals (red dashed) derived from the filtering distribution for the CIR (top) and WF (bottom) simulated datasets. A heat map also represents the filtering densities, darker red indicating higher density. For the WF, each panel plots the marginal for one coordinate, and the single bullet is the proportion of the observations from the corresponding type.*

servations at each of 10 observation times separated by 0.1 seconds.

Figure 3 shows the *exact* filtering for both models, obtained following the parameter updates detailed in Section 2.4 by means of Algorithm 1 (with pruning off). The non observed trajectories of the signals are correctly recovered by the filter.

Figure 4 shows a few trajectories sampled from the joint smoothing density of the CIR model, i.e., the conditional distribution of the signal given the data. This is performed by following the strategy outlined in Section 3 and exploits the correlation with neighbouring time points, thus improving the estimates obtained by means of the marginal smoothing distributions only (Section 2.3.3).

5.2. Performance

We assess the performance of the pruning strategies proposed in Section 4.2 applied to the two above models both from a visual and a quantitative perspective. Figure 5 illustrates the effect of approximating the conditional likelihood, based
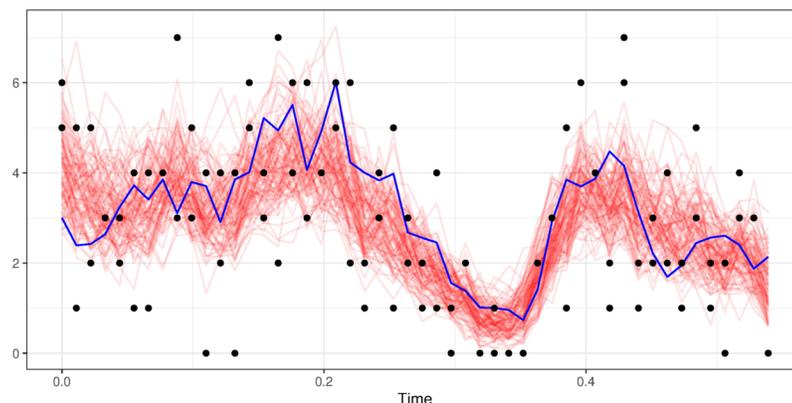
FIG 4. *True trajectory (blue) and a few trajectories sampled from the joint smoothing distribution (red) of the CIR model, together with the data (dots).*
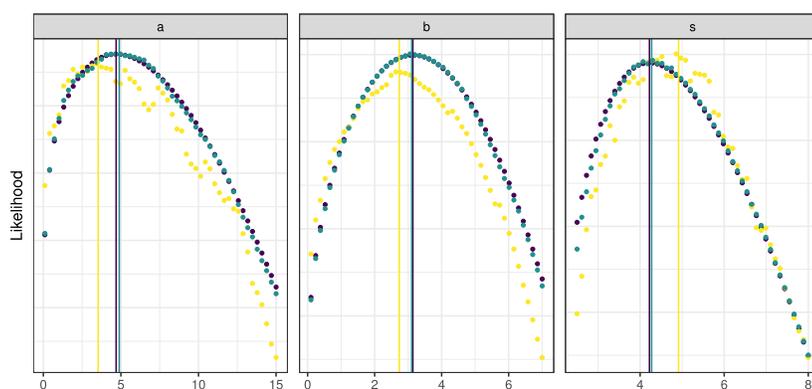

FIG 5. *Comparison between exact and approximate likelihood obtained via pruning, for parameters a, b, s in (35). The dots represent the conditional likelihood for one parameter, computed on a grid and fixing the other parameters at their true value. The vertical lines represent the maximum likelihood estimate obtained by maximising these conditional likelihoods (paired colours). The darkest colour represents the exact likelihood, the intermediate colour represent an approximate likelihood obtained by pruning all but the 10 largest components at each time, and the lightest colour represents the likelihood obtained by pruning all but the single largest component at each time. The simulated data is the same as used in Section 5.1.*

on the same simulated scenario as the previous section. We test the most drastic approximation which retains a single component. This degrades the smoothness of the likelihood and perturbs the concavity of the curves, which in turn renders the optimisation problem challenging. A less extreme approximation, which retains the 10 largest components in the filtering distribution, provide seemingly concave curves whose optima are very close to the exact maximum likelihood estimates.
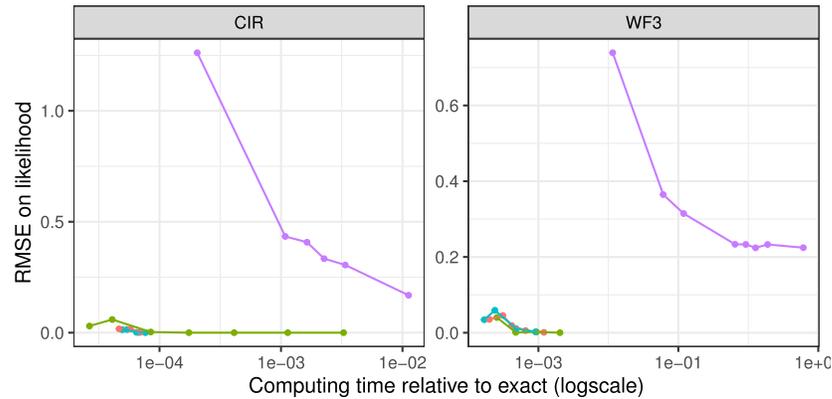
FIG 6. *Accuracy of the likelihood estimate (absolute error) against computing time (relative to computing the exact likelihood), for the CIR (left) and the WF (right). Blue:* fixed threshold *approximations with thresholds 0.01, 0.005, 0.001, 0.0005, 0.0001 for CIR and 0.01, 0.005, 0.001, 0.0001 for WF. Green:* fixed number *approximations with 5, 10, 25, 50, 100, 200, 400 components for CIR and 50, 100, 200, 400 for WF. Red:* fixed fraction *approximation with probability masses 0.95, 0.99, 0.999 for CIR and 0.8, 0.9, 0.95, 0.99, 0.999 for WF. Violet: particle filter approximations with 1000, 5000, 7500, 10 000, 15 000, 50 000 particles for CIR and 100, 500, 1000, 5000, 7500, 10 000, 15 000, 50 000 for WF.*

Next we evaluate more systematically the performance of the pruning strategies for filtering, smoothing and likelihood computation, by comparing them against a bootstrap particle filter with adaptive resampling (for the filtering distribution and the likelihood), and a FFBS particle smoother with adaptive resampling (for the smoothing distribution). The bootstrap particle filter requires sampling from the transition density of the processes, which is straightforward for the CIR process and feasible for the WF through the strategy in Jenkins and Spanò (2017). The FFBS particle smoother (or the two-filter particle smoother) requires the evaluation of the transition density of the process. This is tractable for the CIR model, using the representation of the transition density via Bessel functions (see (36) below) but not for the WF model, where no straightforward representation appears to be available. Thus, we do not implement the particle smoother for the WF model.

We first investigate the performance of the pruning approximations for computing the likelihood. As computing the filtering and the likelihood is very similar, the performance of the approximations is almost the same. To quantify the loss of precision due to the approximation, we compute the absolute error on the likelihood resulting from the pruning, and a root mean squared error for the bootstrap particle filter based on 50 replicates. To quantify the gain in efficiency with the approximation, we measure the time needed to compute the approximate likelihood of the whole dataset relative to the time needed to compute the exact likelihood. For the particle filter, we measure the time needed to obtain a single estimate. Figure 6 shows the absolute error on the likelihood plotted against the time needed to compute the approximate likelihood, and provides
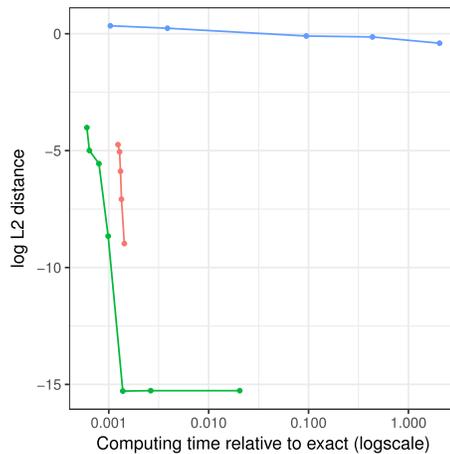
FIG 7. *Log of the maximum L_2 distance between the exact smoothing distributions and the approximated smoothing distributions for the CIR model. Green:* fixed number *approximation with 5, 10, 25, 50, 100, 200 components. Red:* fixed fraction *approximation with 0.8, 0.9, 0.95, 0.99, 0.999 percent of the total mass retained. Blue:* particle filter approximation with 50, 100, 500, 1000, 2500 particles.*

clear evidence that the pruning strategies entail a gain in terms of computing time by three orders of magnitude for the WF case and by four to five orders of magnitude for the CIR case. At the same time, the pruning strategies offer a drastically better precision than the bootstrap particle filters. For the CIR, the particle filter runtime necessary to obtain a precision similar to the pruning strategies seems to be three orders of magnitude larger, while it seems even larger for the WF.

To quantify the loss of precision incurred in estimating a filtering or a smoothing distribution, the simplest approach which extends to a multivariate setting is to compute the L_2 distance between the exact and the approximate distributions obtained via pruning. The L_2 distance between two mixtures of gamma or Dirichlet distributions has an explicit analytical expression under certain conditions (see Appendix B.2), which are satisfied in our setting. As there is one smoothing distribution per observation time, for our simulation scenario we consider the maximum over time of the above distances. In order to compare the pruning and the particle approximations, we exploit the fact that a sample from the particle filter or smoother comes from a mixture of gamma or Dirichlet distributions. We estimate the filtering and smoothing distributions associated with the particle algorithms by means of gamma and Dirichlet kernel density estimates (Chen, 2000; Boone et al., 2014). The particle filtering or smoothing distribution is then estimated by a mixture of kernel components with equal weights, and it is possible to compute the L_2 distance with the exact distribution as with the pruning approximations. Figure 7 shows that the pruning approximations to the smoothing distributions are even more efficient than for the likelihood and the filtering. This is because the particle smoothing comput-
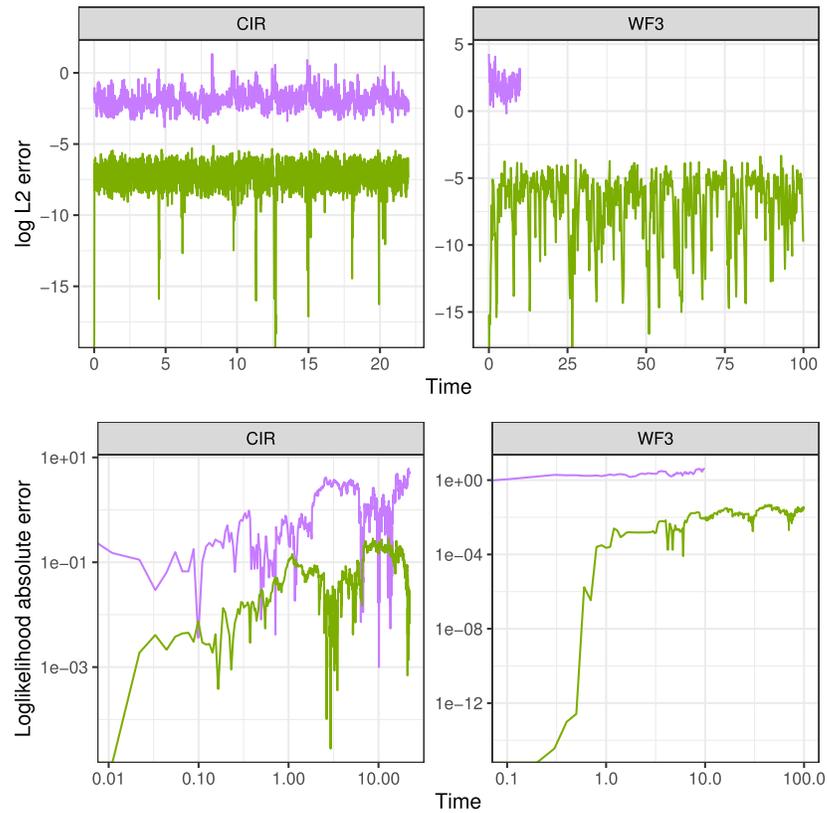
FIG 8. *For a comparable computational budget, evolution of the L2 error on the filtering distributions (top) and of the log-likelihood estimate accuracy (bottom) as a function of time (number of observations) for the CIR (left) and the WF (right). Green:* fixed number *approximation with 180 components for CIR and 200 for WF. Violet: on run of a particle filter approximation with 50 particles. There are 2000 observations for the CIR model, 1000 for the WF model with the* fixed number *approximation and 100 for the particle approximation (which yields a comparable computational budget).*

ing time seems dominated by the evaluation of the Bessel function in the CIR transition density, while our exact dual smoothing algorithm avoids evaluating the transition density altogether.

After studying a scenario with fixed data and varying computational effort, we fix the computational effort and vary the amount of data. Figure 8 shows the evolution of the L_2 error on the filtering distribution and of the absolute error on the loglikelihood as the number of observations increases. The particle filter approximation exhibits the behaviour, usually observed for a fixed number of particles, that the L_2 error on the filtering distribution appears bounded while the absolute error on the likelihood increases with the number of observations (cf. Chopin and Papaspiliopoulos (2020); Cérou et al. (2011)). The error of the pruning approximation is much lower for the same computational budget, and

it seems to grow at a similar rate. Overall, the figure suggests the advantage of the pruning strategy can persist for longer time horizons.

In summary, all pruning strategies are shown to provide very fast and accurate alternatives to the exact strategies, outperforming particle filtering by orders of magnitude.

5.3. Inference on model parameters

A fast approximation of the likelihood using the strategy outlined in Section 4.2 unlocks the door to performing inference on the model parameters (denoted ψ in the Introduction). Additionally, using results from Section 3, one can also address joint inference on signal trajectory and model parameters. In this section, we illustrate both inference problems on simulated datasets. We use the two models considered in previous sections with the specifications of Section 5.1.

For the CIR model, we set parameters $a = 5.0$, $b = 2.4$ and $s = 4.0$ with 2 observations at each of 200 times separated by 0.011 time units. We use weakly informative exponential priors with rate parameter equal to 0.01 for each parameter. For the WF model, we choose $\alpha = (1.1, 2.5, 2.1)$, with 15 categorical variables observed every 0.1 time units over 100 collection times. We use a weakly informative prior given by a half-normal distribution of location 5 and scale 4 for each parameter.

5.3.1. Marginal inference on the parameters

We build an algorithm which uses the filter to estimate the joint posterior on the parameters, via Metropolis–Hastings. Specifically, we implement a standard symmetric random walk Metropolis–Hastings algorithm with Gaussian jumps to sample from the posterior distribution on the parameters using pruning approximations to the likelihood. A good jump size was estimated on a pilot run, based on an estimation of the variance-covariance of the posterior near the mode and scaling the sizes to achieve a good acceptance rate, in the range [0.2,0.4]. We ran three chains until convergence, which was estimated from the potential scale reduction factor (implementation from `MCMCDiagnostics.jl`). Figure 9 presents the autocorrelation plots for both models. The approximation of the likelihood is obtained by pruning all but the 10 largest components of the filtering distributions at each time step. Figure 10 shows that we correctly recover the original parameters.

5.3.2. Joint inference on signal and parameters

Using results from Section 3, we are also able to perform joint inference on the trajectory and the model parameters. We illustrate this in the case of the CIR model, using a Gibbs sampler with a Metropolis–Hastings step. With respect to the previous algorithm for inference on the parameters, this can be viewed as a data augmentation, where states and parameters are sampled jointly.
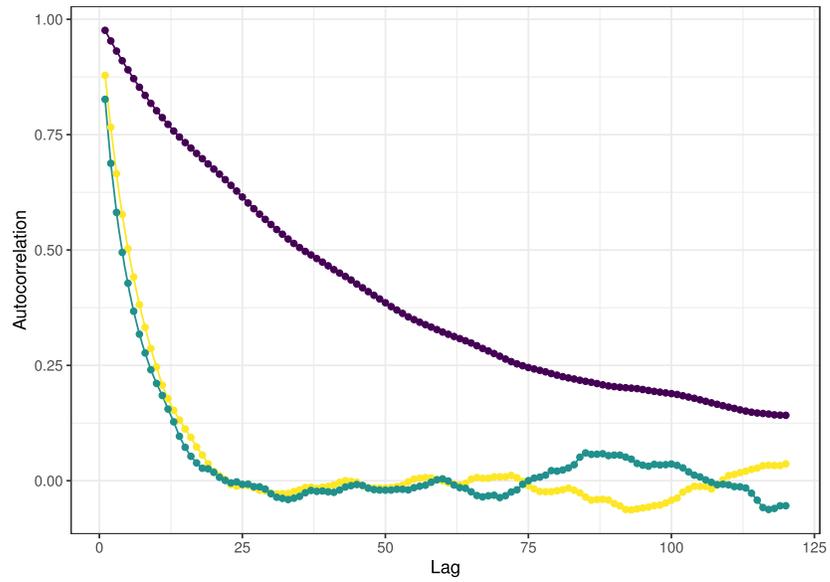
Fig 9. *Autocorrelation function of the MCMC chains for θ (CIR) and for α_1 (WF). The darkest line corresponds to the joint CIR inference, the intermediate line corresponds to the marginal CIR inference and the lightest line to the marginal WF inference.*
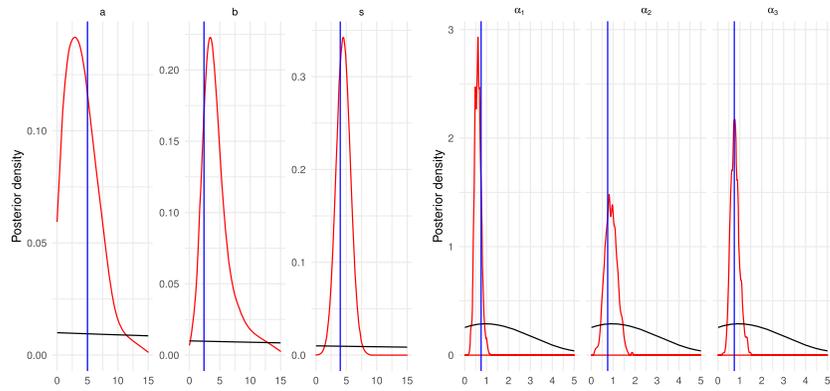
Fig 10. *Prior (black) and posterior distributions (red) for CIR (left) and WF (right) parameters, together with true parameters (blue vertical).*

An outline of the strategy is as follows:

- Gibbs step: update the signal by sampling a new trajectory from the joint smoothing density, conditionally on both parameters and data;
- Metropolis–Hastings step: update the parameters by computing the likelihood of the data and the trajectory given the parameters.
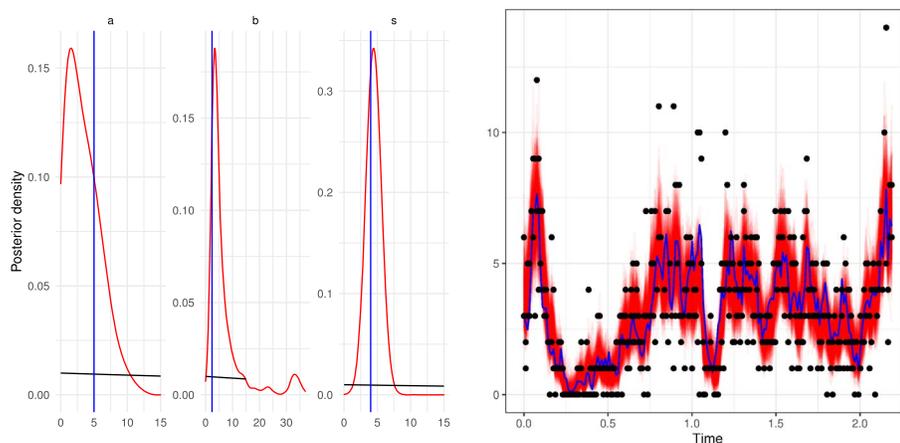
FIG 11. *Left: prior (black) and posterior distributions (red) for the CIR model parameters, along with the true values (blue verticals). Right: a few samples from the posterior distribution (red) for the trajectory and true trajectory (blue) with observed data (black dots).*

Note that for the Metropolis–Hastings step, we can exploit the following well-known representation for evaluating the transition density of the CIR process, written

$$P_\Delta(X_{i+1}|X_i) = c\,e^{-u-v}\left(\frac{v}{u}\right)^{q/2} I_q(2\sqrt{uv}), \tag{36}$$

where $c = [(1 - e^{-2\gamma\Delta})\sigma^2]^{-1}\gamma$, $q = \delta/2 - 1$, $u = cX_i e^{-2\gamma\Delta}$, $v = cX_{i+1}$, and $I_q(2\sqrt{uv})$ is a modified Bessel function of the first kind of order q. The absence of such an expression for the transition density of the WF process prevents from applying the same strategy directly.

The Metropolis–Hastings step is relatively cheap since it simply requires evaluating the likelihood of the parameters conditional on the data and a sampled trajectory, which can be done exactly. The Gibbs step, which consists in simulating a full trajectory conditional on the sampled parameters, can be computationally intensive as it involves summing the weights in the filtering mixtures several times, but can be accelerated using a pruning strategy for the filtering distributions. We used the same priors as for the inference on the model parameters, and ran the Gibbs algorithm for 10000 iterations. We tuned the proposal distributions for the Metropolis–Hastings step using the posterior from the marginal sampler for convenience, but this could also be done using a pilot run. Figure 11 shows that both the parameter values and the hidden trajectory of the CIR were correctly recovered. Figure 9 shows that for this simulated dataset, the chain used for joint inference had stronger autocorrelation than those for marginal inference. The autocorrelation plots for the other parameters are provided in Appendix B (Figure B12).

6. Discussion

Papaspiliopoulos and Ruggiero ([2014](#)) unveiled that the existence of a certain type of *dual process* for the hidden signal leads to exact expressions of the filtering distributions, expressed in *computable form* as mixtures whose number of components varies with the number of observations but is always finite, thus entailing a polynomial computational complexity. Here we have shown that, under essentially the same set of conditions, one can also obtain in computable form expressions for the likelihood and for the marginal smoothing distributions, which can thus be evaluated exactly. These expressions allow to design algorithms for addressing inference on the signal trajectory and on the model parameters, detailed in the paper, thus covering the whole agenda of inference for a hidden Markov model.

Concerning the identification of the dual process for a given class of HMMs, in absence of further insight into the model finer properties (like countable constructions of the forward signal available, e.g., in population genetics), one can work directly with the process infinitesimal generator. After identifying the correct family of distributions that satisfy Assumption [3](#), the output of the computation of the generator on such functions may suggest minor modifications that lead to the generator of the correct dual. This was done for example in Papaspiliopoulos and Ruggiero ([2014](#)) for the CIR model.

We have also outlined acceleration strategies based on informed pruning of the mixture representations of the distributions of interest, and showed that the resulting approach vastly outperforms both in accuracy and computing time a generic approximation scheme like particle filtering and smoothing, bringing the computational burden close to linear in the number of observations. We also observed empirically that the growth rate for the pruning approximation error scales similarly to the particle filter approximations. Our algorithms have been made publicly available via the dedicated Julia package `DualOptimalFiltering`, which implements a general interface applicable to any HMM satisfying the conditions outlined in this paper.

Acknowledgements

Appendix A

Proof of Theorem [2](#)

Using ([13](#)) we have

$$\int_{\mathcal{X}} f_{x_i}(y_i)\nu_{i|0:i-1}(x_i) = \int_{\mathcal{X}} f_{x_i}(y_i) \sum_{\mathbf{m}\in\mathbf{M}_{i|0:i-1}} w_{\mathbf{m}}^{(i)'} g(x,\mathbf{m},\vartheta_{i|0:i-1})$$

$$= \sum_{\mathbf{m} \in \mathbf{M}_{i|0:i-1}} w_{\mathbf{m}}^{(i)'} \int_{\mathcal{X}} f_{x_i}(y_i) g(x, \mathbf{m}, \vartheta_{i|0:i-1})$$

$$= \sum_{\mathbf{m} \in \mathbf{M}_{i|0:i-1}} w_{\mathbf{m}}^{(i)'} \mu_{\mathbf{m}, \vartheta_{i|0:i-1}}(y_i),$$

the last identity following from (11). Hence the marginal likelihood of y_i is

$$\mu_{\nu_{i|0:i-1}}(y_i) := \int_{\mathcal{X}} f_{x_i}(y_i) \nu_{i|0:i-1}(x_i) = \sum_{\mathbf{m} \in \mathbf{M}_{i|0:i-1}} w_{\mathbf{m}}^{(i)'} \mu_{\mathbf{m}, \vartheta_{i|0:i-1}}(y_i) \qquad (37)$$

with $w_{\mathbf{m}}^{(i)'}$ as in (15) and $\mu_{\mathbf{m}, \vartheta_{i|0:i-1}}$ as in (11). Dividing and multiplying by $\mu_{\mathbf{0}, \theta_0}(y_0) = \int_{\mathcal{X}} f_{x_0}(y_0) \pi(x_0)$ we can write

$$\int_{\mathcal{X}} P_{\Delta}(x_1 \,|\, x_0) f_{x_0}(y_0) \pi(\mathrm{d}x_0) = \mu_{\mathbf{0}, \theta_0}(y_0) \int_{\mathcal{X}} P_{\Delta}(x_1 \,|\, x_0) \phi_{y_0}(\pi(\mathrm{d}x_0)).$$

The right hand side equals $\mu_{\mathbf{0}, \theta_0}(y_0) \nu_{1|0:1}(x_1)$, so the equation for the likelihood becomes

$$p(y_{0:T}) = \mu_{\mathbf{0}, \theta_0}(y_0) \int_{\mathcal{X}^T} \prod_{i=2}^{T} f_{x_i}(y_i) P_{\Delta}(x_i \,|\, x_{i-1}) f_{x_1}(y_1) \nu_{1|0:1}(x_1)$$

Iterating with $\mu_{\nu_{i|0:i-1}}(y_i) = \int_{\mathcal{X}} f_{x_i}(y_i) \nu_{i|0:i-1}(x_i)$ and $\phi_{y_i}(\nu_{i|0:i-1})$ in place of $\mu_{\mathbf{0}, \theta_0}(y_0)$ and $\phi_{y_0}(\nu)$ leads to writing

$$p(y_{0:T}) = \mu_{\mathbf{0}, \theta_0}(y_0) \prod_{i=1}^{T} \mu_{\nu_{i|0:i-1}}(y_i)$$

which, through (37), yields the result. □

Proof of Proposition 3

We prove Proposition 3 by induction. For $i = T - 1$, we have

$$p(y_T \,|\, x_{T-1}) = \int_{\mathcal{X}} p(y_T \,|\, x_T) P_{\Delta}(x_T \,|\, x_{T-1}) \mathrm{d}x_T$$

$$= \int_{\mathcal{X}} f_{x_T}(y_T) h(x_t, \mathbf{0}, \theta_0) P_{\Delta}(x_T \,|\, x_{T-1}) \mathrm{d}x_T$$

$$= \mu_{\mathbf{0}, \theta_0}(y_T) \int_{\mathcal{X}} h(x_t, t(y_T, \mathbf{0}), T(y_T, \theta_0)) P_{\Delta}(x_T \,|\, x_{T-1}) \mathrm{d}x_T$$

$$= \mu_{\mathbf{0}, \theta_0}(y_T) \mathbb{E}^{x_{T-1}} [h(X_T, \mathbf{n}, \overleftarrow{\vartheta}_{T:T})]$$

where in the penultimate equality we used the fact that using Assumption 3 and (11) allow to write

$$f_x(y) h(x, \mathbf{m}, \theta) = \mu_{\mathbf{m}, \theta}(y) \, h(x, t(y, \mathbf{m}), T(y, \theta)), \qquad (38)$$

and in the last equality we let $\mathbf{n} = t(y_T, \mathbf{0})$ and $\overleftarrow{\vartheta}_{T:T}$ is as in (18). Lemma B.2 now implies

$$p(y_T \,|\, x_{T-1}) = \sum_{\substack{\mathbf{m} \in \overleftarrow{\mathbf{M}}_{T-1:T}, \mathbf{n} \in \overleftarrow{\mathbf{M}}_{T:T}, \\ \mathbf{n} \geq \mathbf{m}}} \mu_{\mathbf{0},\theta_0}(y_T) p_{\mathbf{n},\mathbf{m}}(\Delta; \overleftarrow{\vartheta}_{T:T}) h(x_{T-1}, \mathbf{m}, \overleftarrow{\vartheta}_{T-1:T})$$

with $\overleftarrow{\mathbf{M}}_{T-1:T} = \mathcal{B}(\overleftarrow{\mathbf{M}}_{T:T}) = \mathcal{B}(\{\mathbf{n}\})$. Hence the statement holds for $i = T - 1$ with

$$\overleftarrow{w}_{\mathbf{m}}^{(T)} = \mu_{\mathbf{0},\theta_0}(y_T) p_{\mathbf{n},\mathbf{m}}(\Delta; \overleftarrow{\vartheta}_{T:T}), \quad \mathbf{n} = t(y_T, \mathbf{0}),$$

since $\{\mathbf{i} \,:\, \mathbf{i} \in \overleftarrow{\mathbf{M}}_{T:T}, \mathbf{i} \geq \mathbf{m}\} = \{\mathbf{n}\}$ in (19). Assume now it holds for $p(y_{i+2:T} \,|\, x_{i+1})$, i.e.

$$p(y_{i+2:T} \,|\, x_{i+1}) = \sum_{\mathbf{m} \in \overleftarrow{\mathbf{M}}_{i+1|i+2:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+2)} h(x_{i+1}, \mathbf{m}, \overleftarrow{\vartheta}_{i+1|i+2:T}).$$

Then

$$p(y_{i+1:T} \,|\, x_i) = \int_{\mathcal{X}} p(y_{i+1} \,|\, x_{i+1}) p(y_{i+2:T} \,|\, x_{i+1}) P_\Delta(x_{i+1} \,|\, x_i) \mathrm{d}x_{i+1}$$

$$= \sum_{\mathbf{m} \in \mathbf{M}_{i+1|i+2:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+2)} \int_{\mathcal{X}} f_{x_{i+1}}(y_{i+1}) h(x_{i+1}, \mathbf{m}, \overleftarrow{\vartheta}_{i+1|i+2:T}) P_\Delta(x_{i+1} \,|\, x_i) \mathrm{d}x_{i+1}$$

$$= \sum_{\mathbf{m} \in \mathbf{M}_{i+1|i+2:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+2)} \mu_{\mathbf{m},\overleftarrow{\vartheta}_{i+1|i+2:T}}(y_{i+1})$$

$$\times \int_{\mathcal{X}} h(x_{i+1}, t(y_{i+1}, \mathbf{m}), T(y_{i+1}, \overleftarrow{\vartheta}_{i+1|i+2:T})) P_\Delta(x_{i+1} \,|\, x_i) \mathrm{d}x_{i+1}$$

$$= \sum_{\mathbf{m} \in \mathbf{M}_{i+1|i+2:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+2)} \mu_{\mathbf{m},\overleftarrow{\vartheta}_{i+1|i+2:T}}(y_{i+1}) \mathbb{E}_{x_i}[h(X_{i+1}, t(y_{i+1}, \mathbf{m}), \overleftarrow{\vartheta}_{i+1|i+1:T})]$$

where the third identity follows from (38) and in the fourth we have used $T(y_{i+1}, \overleftarrow{\vartheta}_{i+1|i+2:T}) = \overleftarrow{\vartheta}_{i+1|i+1:T}$. Since $\mathbf{M}_{i+1|i+1:T} = t(y_{i+1}, \mathbf{M}_{i+1|i+2:T})$, applying again Lemma B.2 yields

$$\sum_{\mathbf{m} \in \mathbf{M}_{i+1|i+2:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+2)} \mu_{\mathbf{m},\overleftarrow{\vartheta}_{i+1|i+2:T}}(y_{i+1}) \mathbb{E}_{x_i}[h(X_{i+1}, t(y_{i+1}, \mathbf{m}), \overleftarrow{\vartheta}_{i+1|i+1:T})]$$

$$= \sum_{\mathbf{m} \in \mathbf{M}_{i+1|i+2:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+2)} \mu_{\mathbf{m},\overleftarrow{\vartheta}_{i+1|i+2:T}}(y_{i+1})$$

$$\sum_{\mathbf{n} \leq t(y_{i+1}, \mathbf{m})} p_{t(y_{i+1}, \mathbf{m}), \mathbf{n}}(\Delta; \overleftarrow{\vartheta}_{i+1|i+1:T}) h(x_i, \mathbf{n}, \overleftarrow{\vartheta}_{i|i+1:T})$$

$$= \sum_{\mathbf{m} \in \mathbf{M}_{i|i+1:T}} \sum_{\mathbf{n} \in \mathbf{M}_{i+1|i+2:T}, \mathbf{n} \geq \mathbf{m}} \overleftarrow{w}_{\mathbf{n}}^{(i+2)} \mu_{\mathbf{n},\overleftarrow{\vartheta}_{i+1|i+2:T}}(y_{i+1})$$

$$p_{t(y_{i+1}, \mathbf{n}), \mathbf{m}}(\Delta; \overleftarrow{\vartheta}_{i+1|i+1:T}) h(x_i, \mathbf{m}, \overleftarrow{\vartheta}_{i|i+1:T})$$

where $\overleftarrow{\vartheta}_{i|i+1:T} = \Theta_\Delta(\overleftarrow{\vartheta}_{i+1|i+1:T})$. Finally, since $\mathbf{M}_{i|i+1:T} = \mathcal{B}(\mathbf{M}_{i+1|i+1:T})$, we obtain

$$\sum_{\mathbf{m}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+1)} h(x_i, \mathbf{m}, \overleftarrow{\vartheta}_{i|i+1:T})$$

with

$$\overleftarrow{w}_{\mathbf{m}}^{(i+1)} = \sum_{\substack{\mathbf{n}\in\overleftarrow{\mathbf{M}}_{i+1|i+2:T}, \\ t(y_{i+1},\mathbf{n})\geq\mathbf{m}}} \overleftarrow{w}_{\mathbf{n}}^{(i+2)} \mu_{\mathbf{n},\overleftarrow{\vartheta}_{i+1|i+2:T}}(y_{i+1}) p_{t(y_{i+1},\mathbf{n}),\mathbf{m}}(\Delta; \overleftarrow{\vartheta}_{i+1|i+1:T}). \qquad \square$$

Proof of Theorem 4

From Propositions 1 and 3, the numerator on the right hand side of (17) reads

$$\sum_{\mathbf{m}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+1)} h(x_i, \mathbf{m}, \overleftarrow{\vartheta}_{i|i+1:T}) \sum_{\mathbf{n}\in\mathbf{M}_{i|i:T}} w_{\mathbf{n}}^{(i)} h(x_i, \mathbf{n}, \vartheta_{i|0:i}) \pi(x_i).$$

By Assumption 4 the previous equals

$$\sum_{\mathbf{m}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}} \sum_{\mathbf{n}\in\mathbf{M}_{i|i:T}} \overleftarrow{w}_{\mathbf{m}}^{(i+1)} w_{\mathbf{n}}^{(i)} C_{\mathbf{m},\mathbf{n},\overleftarrow{\vartheta}_{i|i+1:T},\vartheta_{i|0:i}} h(x_i, d(\mathbf{m},\mathbf{n}), e(\overleftarrow{\vartheta}_{i|i+1:T}, \vartheta_{i|0:i})) \pi(x_i).$$

Since

$$\int_{\mathcal{X}} h(x_i, d(\mathbf{m},\mathbf{n}), e(\overleftarrow{\vartheta}_{i|i+1:T}, \vartheta_{i|0:i})) \pi(x_i) = 1$$

by Assumption 3 and the above sum is finite, the latter can be normalised, so (17) is the distribution

$$p(x_i \mid y_{0:T}) = \sum_{\mathbf{m}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}, \mathbf{n}\in\mathbf{M}_{i|i:T}} w_{\mathbf{m},\mathbf{n}}^{(i)} g(x_i, d(\mathbf{m},\mathbf{n}), e(\overleftarrow{\vartheta}_{i|i+1:T}, \vartheta_{i|0:i})),$$

with

$$w_{\mathbf{m},\mathbf{n}}^{(i)} = \frac{\overleftarrow{w}_{\mathbf{m}}^{(i+1)} w_{\mathbf{n}}^{(i)} C_{\mathbf{m},\mathbf{n},\overleftarrow{\vartheta}_{i|i+1:T},\vartheta_{i|0:i}}}{\sum_{\mathbf{i}\in\overleftarrow{\mathbf{M}}_{i|i+1:T},\mathbf{j}\in\mathbf{M}_{i|i:T}} \overleftarrow{w}_{\mathbf{i}}^{(i+1)} w_{\mathbf{j}}^{(i)} C_{\mathbf{i},\mathbf{j},\overleftarrow{\vartheta}_{i|i+1:T},\vartheta_{i|0:i}}}. \qquad \square$$

Proof of Theorem 5

The forward kernel (31) involves the cost-to-go functions, available from Proposition 3, the Multinomial emission density, and the transition density (33) of the WF process. Then we can write

$$p(\mathbf{x}_i|\mathbf{x}_{i-1}, \mathbf{y}_{i:T}) = p(\mathbf{y}_{i:T}|\mathbf{x}_{i-1})^{-1} p(\mathbf{x}_i|\mathbf{x}_{i-1}) p(\mathbf{y}_i|\mathbf{x}_i) p(\mathbf{y}_{i+1:T}|\mathbf{x}_i)$$

$$
= p(\mathbf{y}_{i:T}|\mathbf{x}_{i-1})^{-1} \sum_{m=0}^{\infty} q_m \sum_{\substack{\mathbf{l}\in\mathbb{N}^K; \\ l_1+\ldots+l_K=m}} \mathrm{MN}(\mathbf{l}|m,\mathbf{x}_{i-1})\mathrm{Dir}(\mathbf{x}_i\,|\,\mathbf{1}+\mathbf{x}_{i-1}+\mathbf{y}_i)
$$

$$
\frac{\Gamma(|\mathbf{y}_i|)\mathcal{B}(\mathbf{1}+\mathbf{x}_{i-1}+\mathbf{y}_i)}{\prod_j \Gamma(y_{ij})\mathcal{B}(\mathbf{1}+\mathbf{x}_{i-1})} \sum_{\mathbf{k}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}} \overleftarrow{w}_{\mathbf{k}}^{(i+1)} h(\mathbf{x}_i,\mathbf{k})
$$

$$
= p(\mathbf{y}_{i:T}|\mathbf{x}_{i-1})^{-1} \sum_{m=0}^{\infty} q_m \sum_{\substack{\mathbf{l}\in\mathbb{N}^K; \\ l_1+\ldots+l_K=m}} \sum_{\mathbf{k}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}} \mathrm{MN}(\mathbf{l}|m,\mathbf{x}_{i-1})
$$

$$
\times \frac{\Gamma(|\mathbf{y}_i|)\mathcal{B}(\mathbf{1}+\mathbf{x}_{i-1}+\mathbf{y}_i)}{\prod_j \Gamma(y_{ij})\mathcal{B}(\mathbf{1}+\mathbf{x}_{i-1})} \overleftarrow{w}_{\mathbf{k}}^{(i+1)} h(\mathbf{x}_i,\mathbf{k})\mathrm{Dir}(\mathbf{x}_i\,|\,\mathbf{1}+\mathbf{x}_{i-1}+\mathbf{y}_i)
$$

$$
= p(\mathbf{y}_{i:T}|\mathbf{x}_{i-1})^{-1} \sum_{m=0}^{\infty} q_m \sum_{\substack{\mathbf{l}\in\mathbb{N}^K; \\ l_1+\ldots+l_K=m}} \sum_{\mathbf{k}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}} \mathrm{MN}(\mathbf{l}|m,\mathbf{x}_{i-1})
$$

$$
\times \frac{\Gamma(|\mathbf{y}_i|)}{\prod_j \Gamma(y_{ij})}\frac{\Gamma(|\boldsymbol{\alpha}|+|\mathbf{k}|)\prod_j \Gamma(\alpha_j)}{\Gamma(|\boldsymbol{\alpha}|)\prod_j \Gamma(\alpha_j+k_j)}\frac{\mathcal{B}(\mathbf{1}+\mathbf{x}_{i-1}+\mathbf{y}_i+\mathbf{k})}{\mathcal{B}(\mathbf{1}+\mathbf{x}_{i-1})}\overleftarrow{w}_{\mathbf{k}}^{(i+1)}
$$

$$
\times \mathrm{Dir}(\mathbf{x}_i\,|\,\mathbf{1}+\mathbf{x}_{i-1}+\mathbf{y}_i+\mathbf{k}).
$$

Denote now

$$
w_{\mathbf{l},\mathbf{k}} = \mathcal{B}(\mathbf{1}+\mathbf{x}_{i-1}+\mathbf{y}_i+\mathbf{k})\frac{\Gamma(|\boldsymbol{\alpha}|+|\mathbf{k}|)}{\prod_j \Gamma(\alpha_j+k_j)}\overleftarrow{w}_{\mathbf{k}}^{(i+1)}
$$

$$
r_{\mathbf{l}} = \sum_{\mathbf{k}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}} w_{\mathbf{l},\mathbf{k}}
$$

$$
\tilde{w}_{\mathbf{l},\mathbf{k}} = r_{\mathbf{l}}^{-1} w_{\mathbf{l},\mathbf{k}}
$$

$$
r_m = \sum_{\mathbf{l}\in\mathbb{N}^K; l_1+\ldots+l_K=m} r_{\mathbf{l}}\frac{\mathrm{MN}(\mathbf{l}|m,\mathbf{x}_{i-1})}{\mathcal{B}(\mathbf{1}+\mathbf{x}_{i-1})}
$$

$$
\tilde{w}_{\mathbf{l}}^m = r_m^{-1} r_{\mathbf{l}}\frac{\mathrm{MN}(\mathbf{l}|m,\mathbf{x}_{i-1})}{\mathcal{B}(\mathbf{1}+\mathbf{x}_{i-1})}
$$

$$
\tilde{q}_m = p(\mathbf{y}_{i:T}|\mathbf{x}_{i-1})^{-1}\frac{\Gamma(|\mathbf{y}_i|)\prod_j \Gamma(\alpha_j)}{\prod_j \Gamma(y_{ij})\Gamma(|\boldsymbol{\alpha}|)}r_m q_m = \tilde{r}_m q_m.
$$

We can then write $p(\mathbf{x}_i|\mathbf{x}_{i-1},\mathbf{y}_{i:T})$ as the mixture

$$
p(x_i|x_{i-1},y_{i:T}) = \sum_{m=0}^{\infty} \tilde{q}_m \sum_{\substack{\mathbf{l}\in\mathbb{N}^K; \\ l_1+\ldots+l_K=m}} \tilde{w}_{\mathbf{l}}^m \sum_{\mathbf{k}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}} \tilde{w}_{\mathbf{l},\mathbf{k}}\mathrm{Dir}(\mathbf{x}_i|\mathbf{1}+\mathbf{x}_{i-1}+\mathbf{y}_i+\mathbf{k}),
$$

$$(39)$$

where, by construction,

$$
\forall m\in\mathbb{N}, \forall \mathbf{l}\in\mathbb{N}^K \text{ such that } l_1+\ldots+l_K=m : \sum_{\mathbf{k}\in\overleftarrow{\mathbf{M}}_{i|i+1:T}} \tilde{w}_{\mathbf{l},\mathbf{k}} = 1
$$

$$\forall m \in \mathbb{N} : \sum_{\mathbf{l} \in \mathbb{N}^K ; l_1 + \ldots + l_K = m} \tilde{w}_{\mathbf{l}}^m = 1.$$

As the right hand side of (39) is a sum of integrable positive functions, integrating both sides with respect to $\mathbf{x}_i$ implies

$$\sum_{m=0}^{\infty} \tilde{q}_m = 1 \iff \sum_{m=0}^{\infty} r_m q_m = p(y_{i:T} | x_{i-1}) \frac{\prod_j \Gamma(y_{ij}) \Gamma(|\alpha|)}{\Gamma(|\mathbf{y}_i|) \prod_j \Gamma(\alpha_j)}$$

implying the forward kernel is an infinite mixture of finite mixtures of Dirichlet distributions with respect to m, $\mathbf{l}$ and $\mathbf{k}$. $\qquad \square$

Sampling from WF the forward kernel

The two inner sums in (34) are finite mixtures, so the challenge to sample from the WF forward kernel lies in how to sample from the infinite mixture with weights $\tilde{q}_m$. To this end, let $b_i^{(t,\theta)}(m) = a_{im}^\theta e^{-i(i+\theta-1)t/2}$ where $a_{im}^\theta = \frac{(\theta+2i-1)(\theta+m)_{(i-1)}}{m!(i-m)!}$. Then the q_m in (33) can be written

$$q_m = \sum_{i=0}^{\infty} (-1)^i b_i(m).$$

Then the $\tilde{q}_m$ in (34) can be written as $\tilde{r}_m q_m$ where, for any given m, $\tilde{r}_m$ is a strictly positive real number (cf. Proof of Theorem 5). Similarly, we define $\tilde{b}_i^{(t,\theta)}(m) = \tilde{r}_m b_i^{(t,\theta)}(m)$.

Note now that all the properties allowing to sample from the distribution defined by the weights q_m in Proposition 1 of Jenkins and Spanò (2017) carry over to the distribution defined by the weights $\tilde{q}_m$. In fact, the following proposition holds:

Proposition 6. *Let*

$$\tilde{C}_m^{(t,\theta)} := \inf \left\{ j \geq 0 : \tilde{b}_{j+m+1}^{(t,\theta)}(m) < \tilde{b}_{j+m}^{(t,\theta)}(m) \right\}.$$

Then

1. $\tilde{C}_m^{(t,\theta)} < \infty$, *for all* m.
2. $\tilde{b}_i^{(t,\theta)}(m) \downarrow 0$ *as* $i \to \infty$ *for all* $i \geq m + \tilde{C}_m^{(t,\theta)}$.
3. $\tilde{C}_m^{(t,\theta)} = 0$ *for all* $m > D_0^{(t,\theta)}$, *where for* $\varepsilon \in [0,1)$,

$$D^{(t,\theta)} := \inf \left\{ i \geq \left(\frac{1}{t} - \frac{\theta+1}{2} \right) \vee 0 : (\theta + 2i + 1) e^{-\frac{(2i+\theta)t}{2}} < 1 \right\}.$$

Proof. Let $b_i^{(t,\theta)}(m) = a_{im}^\theta e^{-i(i+\theta-1)t/2}$ where $a_{im}^\theta = \frac{(\theta+2i-1)(\theta+m)_{(i-1)}}{m!(i-m)!}$, and let

$$C_m^{(t,\theta)} := \inf \left\{ i \geq 0 : b_{i+m+1}^{(t,\theta)}(m) < b_{i+m}^{(t,\theta)}(m) \right\}.$$

Then Proposition 1 in Jenkins and Spanò (2017) implies

1. $C_m^{(t,\theta)} < \infty$, for all m
2. $b_i^{(t,\theta)}(m) \downarrow 0$ as $i \to \infty$ for all $i \geq m + C_m^{(t,\theta)}$; and
3. $C_m^{(t,\theta)} = 0$ for all $m > D_0^{(t,\theta)}$, where for $\varepsilon \in [0,1)$,

$$D_0^{(t,\theta)} := \inf\left\{ i \geq \left(\frac{1}{t} - \frac{\theta+1}{2}\right) \vee 0 : (\theta + 2i + 1)e^{-\frac{(2i+\theta)t}{2}} < 1 \right\}.$$

Note now that $\tilde{C}_m^{(t,\theta)} = C_m^{(t,\theta)}$, since

$$
\begin{aligned}
\tilde{C}_m^{(t,\theta)} &:= \inf\left\{ j \geq 0 : \tilde{b}_{j+m+1}^{(t,\theta)}(m) < \tilde{b}_{j+m}^{(t,\theta)}(m) \right\} \\
&= \inf\left\{ j \geq 0 : \tilde{r}_m b_{j+m+1}^{(t,\theta)}(m) < \tilde{r}_m b_{j+m}^{(t,\theta)}(m) \right\} \\
&= \inf\left\{ j \geq 0 : b_{j+m+1}^{(t,\theta)}(m) < b_{j+m}^{(t,\theta)}(m) \right\} = C_m^{(t,\theta)}
\end{aligned}
$$

Next, observe that

1. $C_m^{(t,\theta)} < \infty$, for all m implies $\tilde{C}_m^{(t,\theta)} < \infty$, for all m.
2. $b_i^{(t,\theta)}(m) \downarrow 0$ as $i \to \infty$ for all $i \geq m + C_m^{(t,\theta)}$ implies that $\tilde{b}_i^{(t,\theta)}(m) \downarrow 0$ as $i \to \infty$ for all $i \geq m + \tilde{C}_m^{(t,\theta)}$, as $\tilde{b}_i^{(t,\theta)}(m) = \tilde{r}_m b_i^{(t,\theta)}(m)$ with $\tilde{r}_m$ a positive constant.
3. $C_m^{(t,\theta)} = 0$ for all $m > D_0^{(t,\theta)}$, where for $\varepsilon \in [0,1)$ implies $\tilde{C}_m^{(t,\theta)} = 0$ for all $m > D_0^{(t,\theta)}$, where for $\varepsilon \in [0,1)$.

This completes the proof. $\qquad\square$

The interpretation of Proposition 6 is that once the coefficients $\tilde{b}_i^{(t,\theta)}(m)$ start to decay, they keep decaying indefinitely so it becomes possible to bound them and use the alternated series trick from Devroye (1986).

To this end, let $\mathbf{i} = (i_0, i_1, \ldots, i_M)$ and define

$$\tilde{S}_\mathbf{i}^-(M) := \sum_{m=0}^{M} \sum_{j=0}^{2i_m+1} (-1)^j \tilde{b}_{m+j}^{(t,\theta)}(m), \quad \tilde{S}_\mathbf{i}^+(M) := \sum_{m=0}^{M} \sum_{j=0}^{2i_m} (-1)^j \tilde{b}_{m+j}^{(t,\theta)}(m).$$

$\tilde{S}_\mathbf{i}^-$ and $\tilde{S}_\mathbf{i}^+$ form the two convergent series bracketing the target distribution, which in turn allows to use the alternated series method. Algorithm 3 details how to sample from the WF forward kernel (34).

Appendix B

B.1. Proofs of some lemmas

Lemma B.1. *The transition probabilities* (7) *are*

$$p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t,\theta) = \gamma_{|\mathbf{m}|,|\mathbf{i}|} C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t) \mathrm{MVH}(\mathbf{i}; \mathbf{m}, |\mathbf{i}|) \tag{B40}$$

Algorithm 3: Simulating from the Wright–Fisher forward kernel

---

Input: $\mathbf{x}_{i-1}$, $\mathbf{y}_{i:T}$ and the cost-to-go weights $\overleftarrow{w}_{\mathbf{k}}^{(i+1)}$
Result: A sample $\mathbf{x}_i$ from the forward kernel
Initialise
$\quad$ | Set $m \longleftarrow 0, k_0 \longleftarrow 0, \mathbf{i} \longleftarrow (i_0)$
$\quad$ | Simulate $U \sim \text{Uniform}[0, 1]$

while *TRUE* **do**
$\quad$ | Set $i_m \longleftarrow \left\lceil \tilde{C}_m^{(t,\theta)}/2 \right\rceil$ with $\tilde{C}_m^{(t,\theta)}$ as in Proposition 6.
$\quad$ | **while** $\tilde{S}_{\mathbf{i}}^-(m) < U < S_{\mathbf{i}}^+(m)$ **do**
$\quad$ | $\quad$ | Set $\mathbf{i} \leftarrow \mathbf{i} + (1, 1, \ldots, 1)$
$\quad$ | **end**
$\quad$ | **if** $S_{\mathbf{i}}^-(m) > U$ **then**
$\quad$ | $\quad$ | **return** m
$\quad$ | **else if** $S_{\mathbf{i}}^+(m) < U$ **then**
$\quad$ | $\quad$ | Set $\mathbf{i} \longleftarrow (i_0, i_1, \ldots, i_m, 0)$
$\quad$ | $\quad$ | Set $m \leftarrow m + 1$
$\quad$ | **end**
$\quad$ | Simulate $\mathbf{l} \sim \text{Categorical}\left(\{\tilde{w}_{\mathbf{l}}^m; \mathbf{l} \in \mathbb{N}^K; l_1 + \ldots + l_K = m\}\right)$
$\quad$ | Simulate $\mathbf{k} \sim \text{Categorical}\left(\{\tilde{w}_{\mathbf{l},\mathbf{k}}; \mathbf{k} \in \overleftarrow{\mathbf{M}}_{i|i+1:T}\}\right)$
$\quad$ | Simulate $\mathbf{x}_i \sim \text{Dir}(\mathbf{l} + \mathbf{x}_{i-1} + \mathbf{y}_i + \mathbf{k})$
end

---

where $\gamma_{|\mathbf{m}|,|\mathbf{i}|} = \prod_{h=0}^{|\mathbf{i}|-1} \lambda_{|\mathbf{m}|-h}$,

$$C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t) = (-1)^{|\mathbf{i}|} \sum_{k=0}^{|\mathbf{i}|} \frac{e^{-\lambda_{|\mathbf{m}|-k} \int_0^t \rho(\Theta_s)\mathrm{d}s}}{\prod_{0 \le h \le |\mathbf{i}|, h \neq k}(\lambda_{|\mathbf{m}|-k} - \lambda_{|\mathbf{m}|-h})}$$

and $\text{MVH}(\mathbf{i}; \mathbf{m}, |\mathbf{i}|)$ *is the multivariate hypergeometric pmf evaluated at* $\mathbf{i}$, *with parameters* $\mathbf{m}$ *and* $|\mathbf{i}|$.

Proof. The proof can be found in Papaspiliopoulos and Ruggiero (2014), Proposition 2.1. $\qquad\square$

Lemma B.2. *Let Assumption 1–2 above hold. Then*

$$\mathbb{E}^x\left[h(X_t, \mathbf{m}, \theta)\right] = \sum_{\mathbf{n} \le \mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t; \theta)h(x, \mathbf{n}, \Theta_t)$$

with Θ_t *being the unique solution to* (6) *with* $\Theta_0 = \theta$.

Proof. The statement follows from an application of (8) with $\Theta_0 = \theta$ and by noting that

$$\mathbb{E}^{\mathbf{m},\theta}\left[h(x, M_t, \Theta_t)\right] = \sum_{\mathbf{n} \in \mathbb{Z}_+^K} p_{\mathbf{m},\mathbf{n}}(t, \theta)h(x, \mathbf{n}, \Theta_t) = \sum_{\mathbf{n} \le \mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t; \theta)h(x, \mathbf{n}, \Theta_t)$$

where $p_{\mathbf{m},\mathbf{n}}(t; \theta)$ *are as in* (7). $\qquad\square$

B.2. Exact L_2 distances

B.2.1. L_2 distance between mixtures of gamma distributions

Lemma B.3. *Consider two gamma mixtures $g = \sum_{i=1}^{I} g_i f_i^g$ and $h = \sum_{j=1}^{J} h_j f_j^h$ where $\forall i \in \mathbb{N}, k \in \{g, h\}; f_i^k := Ga(\alpha_i^k, \beta_i^k)$. Let us further assume that $\forall i \in \mathbb{N}; \alpha_i^k > 0.5$. Then:*

$$\int_{\mathbb{R}^+} (g - h)^2 = \sum_{i,j=1}^{I} g_i g_j \frac{\beta_i^{g \ \alpha_i^g} \beta_j^{g \ \alpha_j^g}}{\Gamma(\alpha_i^g)\Gamma(\alpha_j^g)} \frac{\Gamma(\alpha_i^g + \alpha_j^g - 1)}{(\beta_i^g + \beta_j^g)^{(\alpha_i^g + \alpha_j^g - 1)}}$$

$$\sum_{i,j=1}^{J} h_i h_j \frac{\beta_i^{h \ \alpha_i^h} \beta_j^{h \ \alpha_j^h}}{\Gamma(\alpha_i^h)\Gamma(\alpha_j^h)} \frac{\Gamma(\alpha_i^h + \alpha_j^h - 1)}{(\beta_i^h + \beta_j^h)^{(\alpha_i^h + \alpha_j^h - 1)}}$$

$$- 2 \sum_{i=1}^{I} \sum_{j=1}^{J} g_i h_j \frac{\beta_i^{g \ \alpha_i^g} \beta_j^{h \ \alpha_j^h}}{\Gamma(\alpha_i^g)\Gamma(\alpha_j^h)} \frac{\Gamma(\alpha_i^g + \alpha_j^h - 1)}{(\beta_i^g + \beta_j^h)^{(\alpha_i^g + \alpha_j^h - 1)}}$$

Proof. We denote $\forall i \in \mathbb{N}; f_i := Ga(\alpha_i, \beta_i)$

$$\int_{\mathbb{R}^+} f_1 f_2 = \int_{\mathbb{R}^+} \frac{\beta_1^{\alpha_1} \beta_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{(\alpha_1 + \alpha_2 - 1) - 1} e^{-(\beta_1 + \beta_2)x}$$

$$= \frac{\beta_1^{\alpha_1} \beta_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + \alpha_2 - 1)}{(\beta_1 + \beta_2)^{(\alpha_1 + \alpha_2 - 1)}} \int_{\mathbb{R}^+} Ga(\alpha_1 + \alpha_2 - 1, \beta_1 + \beta_1)$$

Now provided that $\alpha_1 + \alpha_2 > 1$, which is implied by $\forall i \in \mathbb{N}; \alpha_i > 0.5$:

$$\int_{\mathbb{R}^+} f_1 f_2 = \frac{\beta_1^{\alpha_1} \beta_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + \alpha_2 - 1)}{(\beta_1 + \beta_2)^{(\alpha_1 + \alpha_2 - 1)}}$$

We now consider two gamma mixtures $g = \sum_{i=1}^{I} g_i f_i$ and $h = \sum_{j=1}^{J} h_j f_j$.

$$(g - h)^2 = \left(\sum_{i=1}^{I} g_i f_i - \sum_{j=1}^{J} h_j f_j \right)^2$$

$$= \left(\sum_{i=1}^{I} g_i f_i \right)^2 + \left(\sum_{j=1}^{J} h_j f_j \right)^2 - 2 \sum_{i=1}^{I} g_i f_i \sum_{j=1}^{J} h_j f_j$$

$$= \sum_{i,j=1}^{I} g_i g_j f_i f_j + \sum_{i,j=1}^{J} h_i h_j f_i f_j - 2 \sum_{i=1}^{I} \sum_{j=1}^{J} g_i h_j f_i f_j$$

$$\int_{\mathbb{R}^+} (g - h)^2 = \sum_{i,j=1}^{I} g_i g_j \int_{\mathbb{R}^+} f_i f_j + \sum_{i,j=1}^{J} h_i h_j \int_{\mathbb{R}^+} f_i f_j - 2 \sum_{i=1}^{I} \sum_{j=1}^{J} g_i h_j \int_{\mathbb{R}^+} f_i f_j$$

Let us now write $g = \sum_{i=1}^{I} g_i f_i^g = \mathrm{Ga}(\alpha_i^g, \beta_i^g)$ and $h = \sum_{i=1}^{I} h_i f_i^h = \mathrm{Ga}(\alpha_i^h, \beta_i^h)$

$$\int_{\mathbb{R}^+} (g-h)^2 = \sum_{i,j=1}^{I} g_i g_j \frac{\beta_i^{g \ \alpha_i^g} \beta_j^{g \ \alpha_j^g}}{\Gamma(\alpha_i^g)\Gamma(\alpha_j^g)} \frac{\Gamma(\alpha_i^g + \alpha_j^g - 1)}{(\beta_i^g + \beta_j^g)^{(\alpha_i^g + \alpha_j^g - 1)}}$$

$$\sum_{i,j=1}^{J} h_i h_j \frac{\beta_i^{h \ \alpha_i^h} \beta_j^{h \ \alpha_j^h}}{\Gamma(\alpha_i^h)\Gamma(\alpha_j^h)} \frac{\Gamma(\alpha_i^h + \alpha_j^h - 1)}{(\beta_i^h + \beta_j^h)^{(\alpha_i^h + \alpha_j^h - 1)}}$$

$$- 2 \sum_{i=1}^{I} \sum_{j=1}^{J} g_i h_j \frac{\beta_i^{g \ \alpha_i^g} \beta_j^{h \ \alpha_j^h}}{\Gamma(\alpha_i^g)\Gamma(\alpha_j^h)} \frac{\Gamma(\alpha_i^g + \alpha_j^h - 1)}{(\beta_i^g + \beta_j^h)^{(\alpha_i^g + \alpha_j^h - 1)}} \quad \square$$

B.2.2. L_2 distance between mixtures of Dirichlet distributions

Lemma B.4. *Consider two Dirichlet mixtures $g = \sum_{i=1}^{I} g_i f_i^g$ and $h = \sum_{j=1}^{J} h_j f_j^h$ where $\forall i \in \mathbb{N}, k \in \{g, h\}; f_i^k := Dir(\boldsymbol{\alpha}_i) = \frac{1}{B(\boldsymbol{\alpha}_i^k)} \prod_{j=1}^{K} x_j^{\alpha_{i,j}^k - 1}$ where $B(\boldsymbol{\alpha}_i) = \frac{\prod_{j=1}^{K} \Gamma(\alpha_{i,j})}{\Gamma(\sum_{j=1}^{K} \alpha_{i,j})}$. Let us denote ∇_K the K-dimensional simplex. Let us further assume that $\forall i \in \mathbb{N}; \alpha_i > 0.5$. Then:*

$$\int_{\nabla_K} (g-h)^2 = \sum_{i,j=1}^{I} g_i g_j \frac{B(\boldsymbol{\alpha}_i^g + \boldsymbol{\alpha}_j^g - 1)}{B(\boldsymbol{\alpha}_i^g) B(\boldsymbol{\alpha}_j^g)} + \sum_{i,j=1}^{I} h_i h_j \frac{B(\boldsymbol{\alpha}_i^h + \boldsymbol{\alpha}_j^h - 1)}{B(\boldsymbol{\alpha}_i^h) B(\boldsymbol{\alpha}_j^h)}$$

$$- 2 \sum_{i=1}^{I} \sum_{j=1}^{J} g_i h_j \frac{B(\boldsymbol{\alpha}_i^g + \boldsymbol{\alpha}_j^h - 1)}{B(\boldsymbol{\alpha}_i^g) B(\boldsymbol{\alpha}_j^h)}$$

Proof. We denote $\forall i \in \mathbb{N}; f_i := \mathrm{Dir}(\boldsymbol{\alpha}_i) = \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} x_j^{\alpha_{i,j} - 1}$ where $B(\boldsymbol{\alpha}_i) = \frac{\prod_{j=1}^{K} \Gamma(\alpha_{i,j})}{\Gamma(\sum_{j=1}^{K} \alpha_{i,j})}$ and ∇_K the K-dimensional simplex.

$$\int_{\nabla_K} f_1 f_2 = \int_{\nabla_K} \frac{1}{B(\boldsymbol{\alpha}_1) B(\boldsymbol{\alpha}_2)} \prod_{j=1}^{K} x_j^{\alpha_{1,j} + \alpha_{2,j} - 2}$$

$$= \frac{B(\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 - 1)}{B(\boldsymbol{\alpha}_1) B(\boldsymbol{\alpha}_2)} \int_{\nabla_K} \mathrm{Dir}(\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 - 1)$$

Now provided that $\forall j \in \{1, \dots, K\}, \alpha_{1,j} + \alpha_{2,j} > 1$:

$$\int_{\nabla_K} f_1 f_2 = \frac{B(\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 - 1)}{B(\boldsymbol{\alpha}_1) B(\boldsymbol{\alpha}_2)}$$

We now consider two Dirichlet mixtures $g = \sum_{i=1}^{I} g_i f_i$ and $h = \sum_{j=1}^{J} h_j f_j$.

Following the product expansion in the previous subsection, we can also obtain the exact expression:

$$\int_{\nabla_K} (g-h)^2 = \sum_{i,j=1}^{I} g_i g_j \frac{B(\boldsymbol{\alpha}_i^g + \boldsymbol{\alpha}_j^g - 1)}{B(\boldsymbol{\alpha}_i^g)B(\boldsymbol{\alpha}_j^g)} + \sum_{i,j=1}^{I} h_i h_j \frac{B(\boldsymbol{\alpha}_i^h + \boldsymbol{\alpha}_j^h - 1)}{B(\boldsymbol{\alpha}_i^h)B(\boldsymbol{\alpha}_j^h)}$$
$$- 2\sum_{i=1}^{I}\sum_{j=1}^{J} g_i h_j \frac{B(\boldsymbol{\alpha}_i^g + \boldsymbol{\alpha}_j^h - 1)}{B(\boldsymbol{\alpha}_i^g)B(\boldsymbol{\alpha}_j^h)} \qquad \Box$$

B.3. Additional algorithm

Algorithm B4: Filtering and likelihood

Pruning setting: ON (approximate filtering) / OFF (exact filtering)
Input: $Y_{0:n}$, $t_{0:n}$ and $\nu = h(x, \mathbf{0}, \theta_0) \in \mathcal{F}$ for some $\theta_0 \in \Theta$
Result: $\vartheta_{i|0:i}$, $\mathbf{M}_{i|0:i}$, $W_{0:n}$ with $W_i = \{w_\mathbf{m}^i, \mathbf{m} \in \mathbf{M}_{i|0:i}\}$ and $\vartheta_{i|0:i-1}$, $\mathbf{M}_{i|0:i-1}$, $W'_{1:n}$
     with $W'_i = \{w_\mathbf{m}^{i'}, \mathbf{m} \in \mathbf{M}_{i|0:i-1}\}$ and the likelihood $p(y_{0:T})$.

Initialise
> Set $\vartheta_{0|0} = T(Y_0, \theta_0)$ with T as in (12)
> Set $\mathbf{M}_{0|0} = \{t(Y_0, \mathbf{0})\} = \{\mathbf{m}^*\}$ and $W_0 = \{1\}$ with t as in Assumption 3
> Compute $\vartheta_{1|0}$ from $\vartheta_{0|0}$ as in (12)
> Set $\mathbf{M}_{1|0} = \mathcal{B}(\mathbf{M}_{0|0})$ and $W'_1 = \{p_{\mathbf{m}^*,\mathbf{n}}(\Delta, \vartheta_{0|0}), \mathbf{n} \in \mathbf{M}_{1|0}\}$ with $\mathcal{B}$ as in (5) and
> $p_{\mathbf{m},\mathbf{n}}$ as in (B40)

for *i from* 1 *to* n **do**
> **Update**
>> Set $\vartheta_{i|0:i} = T(Y_i, \vartheta_{i|0:i-1})$
>> Set $W_i = \{\frac{w_\mathbf{m}^{i'}\mu_{\mathbf{m},\vartheta_{i|0:i-1}}(Y_i)}{\sum_{\mathbf{n}\in\mathbf{M}_{i|0:i-1}} w_\mathbf{n}^{i'}\mu_{\mathbf{n},\vartheta_i}(Y_i)}, \mathbf{m} \in \mathbf{M}_{i|0:i-1}\}$ with $\mu_{\mathbf{m},\theta}$ defined as in
>> (11)
>> Set $\mathbf{M}_{i|0:i} = \{t(Y_i, \mathbf{m}), \mathbf{m} \in \mathbf{M}_{i|0:i-1}\}$ and update the labels in W_i
>
> **if** *pruning* ON **then**
>> Prune $\mathbf{M}_{i|0:i}$ and remove the corresponding weights in W_i
>> Normalise the weights in W_i
>
> **Predict**
>> Compute $\vartheta_{i+1|0:i}$ from $\vartheta_{i|0:i}$
>> Set $\mathbf{M}_{i+1|0:i} = \mathcal{B}(\mathbf{M}_{i|0:i})$ and
>> $$W'_{i+1} = \left\{ \sum_{\mathbf{m}\in\mathbf{M}_{i|0:i}, \mathbf{m}\geq\mathbf{n}} w_\mathbf{m}^i p_{\mathbf{m},\mathbf{n}}(\Delta, \vartheta_{i|0:i}), \mathbf{n} \in \mathbf{M}_{i+1|0:i} \right\}$$

end
Compute likelihood
> Compute the likelihood using $\vartheta_{i|0:i-1}$, $\mathbf{M}_{i|0:i-1}$ and $W'_{1:n}$ as in Equation (16)

end

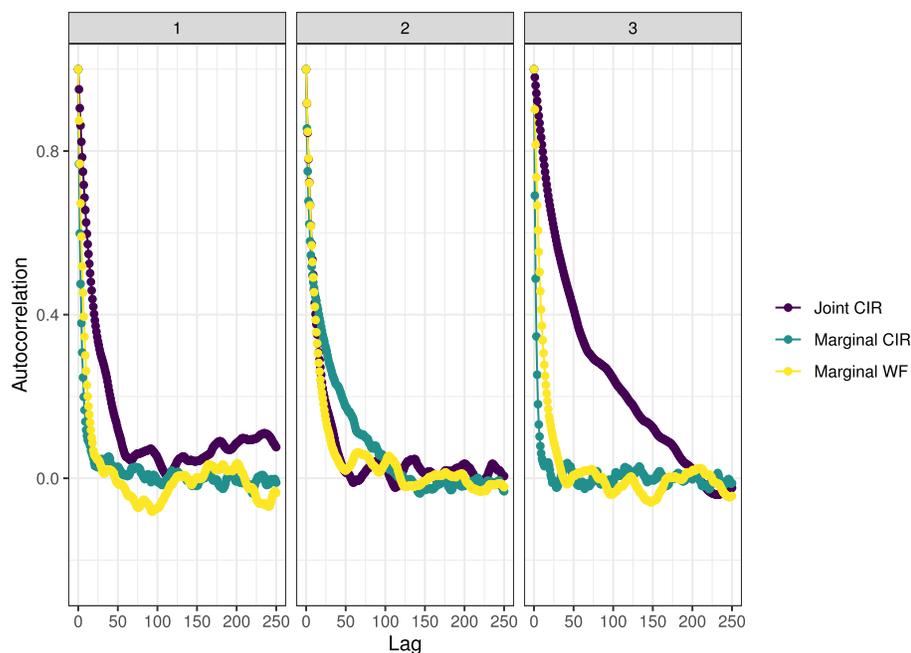
B.4. Autocorrelation function plots for all parameters

FIG B12. *Autocorrelation function of the MCMC chains for (α, β, σ) (CIR) and for $\boldsymbol{\alpha}$ (WF). The darkest line corresponds to the joint CIR inference, the intermediate line corresponds to the marginal CIR inference and the lightest line to the marginal WF inference.*

References

Ascolani, F., Lijoi, A., and Ruggiero, M. (2020). Predictive inference with Fleming–Viot-driven dependent Dirichlet processes. *Bayesian Analysis, in press*.

Bae, K., Mallick, B. K., and Elsik, C. G. (2005). Prediction of protein interdomain linker regions by a hidden Markov model. *Bioinformatics*, 21(10):2264–2270.

Beskos, A., Papaspiliopoulos, O., and Roberts, G. O. (2006). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12(6):1077–1098. MR2274855

Boone, E. L., Merrick, J. R., and Krachey, M. J. (2014). A Hellinger distance approach to MCMC diagnostics. *Journal of Statistical Computation and Simulation*, 84(4):833–849. MR3169364

Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K., and Haussler, D. (1993). Using Dirichlet mixture priors to derive hidden Markov models for protein families. In *International Conference on Intelligent Systems for Molecular Biology*, volume 1, pages 47–55.

Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models*. Springer, New York, NY, USA. MR2159833

Cérou, F., Del Moral, P., and Guyader, A. (2011). A nonasymptotic theorem for unnormalized Feynman–Kac particle models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 47(3):629–649. MR2841068

Chaleyat-Maurel, M. and Genon-Catalot, V. (2006). Computable infinite-dimensional filters with applications to discretized diffusion processes. *Stochastic processes and their applications*, 116(10):1447–1467. MR2260743

Chaleyat-Maurel, M. and Genon-Catalot, V. (2009). Filtering the Wright-Fisher diffusion. *ESAIM: Probability and Statistics*, 13:197–217. MR2518546

Chen, S. X. (2000). Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52(3):471–480. MR1794247

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75(1):79–97. MR1414504

Chopin, N. and Papaspiliopoulos, O. (2020). *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer International Publishing. MR4215639

Comte, F., Genon-Catalot, V., and Kessler, M. (2011). Multiplicative Kalman filtering. *test*, 20(2):389–411. MR2834053

Crane, H. (2017). A hidden Markov model for latent temporal clustering with application to ideological alignment in the US Supreme Court. *Computational Statistics & Data Analysis*, 110:19–36. MR3612605

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York. MR0836973

Ethier, S. N. and Griffiths, R. C. (1993). The Transition Function of a Fleming-Viot Process. *The Annals of Probability*, 21(3):1571–1590. MR1235429

Ethier, S. N. and Kurtz, T. G. (1986). *Markov processes: Characterization and convergence*. John Wiley & Sons, Inc., Publication, 1. MR0838085

Ferrante, M. (1992). On the existence of finite-dimensional filters in discrete time. *Stochastics and Stochastic Reports*, 40(3-4):169–179. MR1275131

Ferrante, M. and Runggaldier, W. J. (1990). On necessary conditions for the existence of finite-dimensional filters in discrete time. *Systems and Control Letters*, 14(1):63–69. MR1037748

Ferrante, M. and Vidoni, P. (1998). Finite dimensional filters for nonlinear stochastic difference equations with multiplicative noises. *Stochastic processes and their applications*, 77(1):69–81. MR1644614

Fieker, C., Hart, W., Hofmann, T., and Johansson, F. (2017). Nemo/Hecke: Computer algebra and number theory packages for the julia programming language. In *Proceedings of the 2017 ACM on International Symposium on Symbolic and Algebraic Computation*, ISSAC'17, pages 157–164, ACM, New York, NY, USA. MR3703682

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5(2A):1020–1056. MR2840185

Genon-Catalot, V. (2003). A non-linear explicit filter. *Statistics & probability*

letters, 61(2):145–154. MR1950665

Genon-Catalot, V. and Kessler, M. (2004). Random scale perturbation of an AR (1) process and its properties as a nonlinear explicit filter. *Bernoulli*, 10(4):701–720. MR2076070

Griffiths, R. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoretical Population Biology*, 17(1):37–50. MR0568666

Guenther, S. (1981). Finite dimensional filter systems in discrete time. *Stochastics*, 5(1-2):107–114. MR0643064

Guha, S., Li, Y., and Neuberg, D. (2008). Bayesian hidden Markov modeling of array CGH data. *Journal of the American Statistical Association*, 103(482):485–497. MR2523987

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of econometrics*, 45(1-2):39–70. MR1067230

Jansen, S. and Kurt, N. (2014). On the notion (s) of duality for Markov processes. *Probability surveys*, 11:59–120. MR3201861

Jenkins, P. A. and Spanò, D. (2017). Exact simulation of the Wright–Fisher diffusion. *The Annals of Applied Probability*, 27(3):1478–1509. MR3678477

Langrock, R., Kneib, T., Sohn, A., and DeRuiter, S. L. (2015). Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, 71(2):520–528. MR3366256

Papaspiliopoulos, O. and Ruggiero, M. (2014). Optimal filtering and the dual process. *Bernoulli*, 20(4):1999–2019. MR3263096

Papaspiliopoulos, O., Ruggiero, M., and Spano, D. (2016). Conjugacy properties of time-evolving Dirichlet and gamma random measures. *Electronic Journal of Statistics*, 10(2):3452–3489. MR3572856

Quick, N. J., Isojunno, S., Sadykova, D., Bowers, M., Nowacek, D. P., and Read, A. J. (2017). Hidden Markov models reveal complexity in the diving behaviour of short-finned pilot whales. *Scientific reports*, 7:45765.

Runggaldier, W. J. and Spizzichino, F. (2001). Sufficient conditions for finite dimensionality of filters in discrete time: A Laplace transform-based approach. *Bernoulli*, pages 211–221. MR1828503

Sarkar, A. and Dunson, D. B. (2019). Bayesian higher order hidden markov models. *arXiv preprint* arXiv:1805.12201v2.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*, volume 3. Cambridge University Press. MR3154309

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26(2):119–164. MR0770050

Titsias, M. K., Holmes, C. C., and Yau, C. (2016). Statistical inference in hidden Markov models using k-segment constraints. *Journal of the American Statistical Association*, 111(513):200–215. MR3494653

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):37–57. MR2797735