ISSN 2279-9362

Approximate filtering via discrete dual processes

Guillaume Kon Kam King, Andrea Pandolfi, Marco Piretto and Matteo Ruggiero

No. 710

December 2023

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Approximate filtering via discrete dual processes

Guillaume Kon Kam King, *Université Paris-Saclay - INRAE*

Andrea Pandolfi, *Bocconi University*

Marco Piretto, *BrandDelta*

Matteo Ruggiero*, *University of Torino and Collegio Carlo Alberto*

November 28, 2023

We consider the task of filtering a dynamic parameter evolving as a diffusion process, given data collected at discrete times from a likelihood which is conjugate to the reversible law of the diffusion, when a generic dual process on a discrete state space is available. Recently, it was shown that duality with respect to a death-like process implies that the filtering distributions are finite mixtures, making exact filtering and smoothing feasible through recursive algorithms with polynomial complexity in the number of observations. Here we provide general results for the case where the dual is a regular jump continuous-time Markov chain on a discrete state space, which typically leads to filtering distribution given by countable mixtures indexed by the dual process state space. We investigate the performance of several approximation strategies on two hidden Markov models driven by Cox–Ingersoll–Ross and Wright–Fisher diffusions, which admit duals of birth-and-death type, and compare them with the available exact strategies based on death-type duals and with bootstrap particle filtering on the diffusion state space as a general benchmark.

Keywords: Bayesian inference; Diffusion; Duality; Hidden Markov models; Particle filtering; Smoothing.

1 Introduction

Hidden Markov models are widely used statistical models for time series that assume an unobserved Markov process $(X_t)_{t\geq 0}$, or hidden *signal*, driving the process that generates the observations $(Y_{t_i})_{i=0,\dots,n}$, e.g., by specifying the dynamics of one or more parameters of the observation density $f_{X_t}(y)$, called *emission distribution*. See [11] for a general treatment of hidden Markov models. In this framework, the first task is to estimate the trajectory of the signal given observations collected at discrete times $0 = t_0 < t_1 < \cdots < t_n = T$, which amounts to performing sequential Bayesian inference by computing the so-called *filtering distributions* $p(x_{t_i}|y_{t_0}, \dots, y_{t_{i-1}})$, i.e., the marginal

*Corresponding author: ESOMAS Dept., Corso Unione Sovietica 218/bis, 10134, Torino, Italy; matteo.ruggiero@unito.it

distributions of the signal at time t_i conditional on observations collected up to time t_{i-1}. Originally motivated by real-time tracking and navigation systems and pioneered by [46, 47], classical and widely known explicit results for this problem include: the *Kalman–Bucy* filter, when both the signal and the observation process are formulated in a gaussian linear system; the *Baum–Welch* filter, when the signal has a finite state-space as the observations are categorical; the *Wonham* filter, when the signal has a finite state-space and the observations are Gaussian. These scenarios allow the derivation of so-called *finite-dimensional filters*, i.e., a sequence of filtering distributions whose explicit identification is obtained through a parameter update based on the collected observations and on the time intervals between the collection times. In such case, the resulting computational cost of the algorithm increases linearly with the number of observations. Other explicit results include [16, 26, 27, 30, 31, 55, 56]. Outside these classes, explicit solutions are difficult to obtain, and their derivation typically relies on *ad hoc* computations. This is especially true when the map $x \mapsto f_x$ is non-linear and when the signal transition kernel is known up to a series expansion, often intractable, as is the case for many widely used stochastic models. When exact solutions are not available, one must typically make use of approximate strategies, whose state of the art is most prominently based on extensions of the Kalman and particle filters. See, for example, [6, 15].

A somewhat weaker but useful notion with respect to that of a finite-dimensional filter was formulated in [13], who introduced the concept of *computable filter*. This extends the former class to a larger class of filters whose marginal distributions are *finite mixtures* of elementary kernels, rather than single kernels. Unlike the former case, such a scenario entails a higher computational cost, usually polynomial in the number of observations, but avoids the infinite-dimensionality typically implied by series expansion of the signal transition kernel. See [13, 14] for two examples.

Recently, [53] derived sufficient conditions for computable filtering based on duality. A *dual process* is a process D_t which enjoys the identity

$$\mathbb{E}[h(X_t, d)|X_0 = x] = \mathbb{E}[h(x, D_t)|D_0 = d]. \tag{1}$$

Here the expectation on the left-hand side is taken with respect to the transition law of the signal X_t, and that on the right hand side with respect to that of D_t, while the class of functions $h(x, d)$ which satisfy the above identity are called duality functions. See [44] for a review and for the technical details we have overlooked here. The use of duality is largely established in probability and statistical physics, see for example [1, 7, 8, 12, 18, 21, 23, 24, 25, 28, 33, 34, 35, 38, 42, 43, 51, 52]. The use of duality for inference was initiated by [53], who showed that for a reversible signal whose marginal distributions are conjugate to the emission distribution (i.e., the Bayesian update at a fixed t can be computed in closed-form), computable filtering is guaranteed if the stochastic part of the dual process evolves on a finite state space. Examples of such scenario were given for non-linear hidden Markov models involving signals driven by Cox–Ingersoll–Ross (CIR) and K-dimensional Wright–Fisher (WF) diffusions, for which recursive formulae for the filtering distributions were derived. Along similar lines, duality was exploited for computable *smoothing* in [50], whereby the signal is also conditioned on data collected at later times, and for nonparametric hidden Markov models driven by Fleming–Viot and Dawson–Watanabe diffusions in [2, 3, 54].

In this paper, we investigate filtering problems for hidden Markov models when the dual process takes the more general form of a continuous-time Markov chain on a discrete state space. This is of interest for example in some population genetic models with selection [7] or interaction [4, 25] whose known dual processes are of birth-and-death (B&D) type, whose specific filtering problems are currently under investigation by some of the authors. When the dual process evolves in a countable state space, the filtering distributions can in general be expected to be countably infinite mixtures. This leads to inferential procedures which are not *computable* in the sense specified above, since the computation of the filtering distribution can no longer be exact. It is thus natural to wonder how the inferential procedures obtained in such duality-based scenario, possibly aided by some suitable approximation strategies, would perform.

The paper is organized as follows. In Section 2 we identify sufficient conditions for filtering based on discrete duals and provide a general description of the filtering operations in this setting. In Section 3, we apply these results to devise practical algorithms which allow to evaluate in recursive form filtering and smoothing distributions under this formulation. Section 4 and 5 investigate hidden Markov models driven by a Cox–Ingersoll–Ross diffusion, which admits a dual given by a one-dimensional B&D process, and by a K-dimensional Wright–Fisher diffusions, which is shown to admit a dual given by a K-dimensional Moran model. The latter can be seen as a multidimensional B&D process with constant total population size. Section 6 discusses several approximation strategies used to implement the above algorithms with these dual processes, and compares their performance with exact filtering based on the results in [53] and with a bootstrap particle filter as a general benchmark. Finally, we conclude with some brief remarks.

2 Filtering via discrete dual processes

Consider a hidden signal $(X_t)_{t\geq 0}$ given by a diffusion process on $\mathcal{X} \subset \mathbb{R}^K$, for $K \geq 1$. This takes here the role of a temporally evolving parameter which is the main target of estimation. Observations $Y_{t_i} \in \mathcal{Y} \subset \mathbb{R}^D$, $D \geq 1$, are collected at discrete times $0 = t_0 < t_1 < \cdots < t_n = T$ with distribution $Y_{t_i} \overset{\text{ind}}{\sim} f_x(\cdot)$, given $X_{t_i} = x$. Knowledge of the signal state x thus makes the observations conditionally independent. Given an observation $Y = y$, define the *update operator* ϕ_y acting on densities ξ on $\mathcal{X}$ by

$$\phi_y(\xi)(x) := \frac{f_x(y)\xi(x)}{\mu_\xi(y)}, \qquad \mu_\xi(y) := \int_{\mathcal{X}} f_x(y)\xi(x). \tag{2}$$

Here and later we assume all densities of interest exist with respect to an appropriate dominating measure. In (2), ξ acts as a *prior* distribution on the signal state, which encodes the current knowledge (or lack thereof) on X_t, whereas $\mu_\xi(y)$ is interpreted as the marginal likelihood of a data point y when X_t has distribution ξ. The update operator thus amounts to an application of Bayes' theorem for conditioning ξ on a new observation y, leading to the updated density $\phi_y(\xi)$. For example, if ξ is a Beta(a, b) density on $[0, 1]$, and f_x is Bern(x), then $\phi_y(\xi)$ is Beta$(a + y, b + 1 - y)$ as in a classical Beta-Bernoulli update.

Define also the *propagation operator* ψ_t by

$$\psi_t(\xi)(x') := \int_{\mathcal{X}} \xi(x) P_t(x'|x) \mathrm{d}x'. \tag{3}$$

where P_t is the transition density of the signal. Here $\psi_t(\xi)$ is the probability density at time t obtained by propagating forward the density ξ of the signal at time 0 by means of the signal semigroup.

We will make three assumptions, the first two of which are the same as in [53].

> **Assumption 1** (Reversibility). The signal X_t is reversible with respect to the density π, i.e., $\pi(x)P_t(x'|x) = \pi(x')P_t(x|x')$.

See the discussion in [53] on the possibility of relaxing the above assumption. For $K \in \mathbb{Z}_+$, define now the space of multi-indices $\mathcal{M} = \mathbb{Z}_+^K$ to be

$$\mathcal{M} = \{\mathbf{m} = (m_1, \ldots, m_K) : \ m_j \in \mathbb{Z}_+, \text{ for } j = 1, \ldots, K\},$$

whose origin is denoted $\mathbf{0} = (0, \ldots, 0)$.

> **Assumption 2** (Conjugacy). For $\Theta \subset \mathbb{R}^l$, $l \in \mathbb{Z}_+$, let $h : \mathcal{X} \times \mathcal{M} \times \Theta \to \mathbb{R}_+$ be such that $\sup_{x \in \mathcal{X}} h(x, \mathbf{m}, \theta) < \infty$ for all $\mathbf{m} \in \mathcal{M}, \theta \in \Theta$ and $h(x, \mathbf{0}, \theta') = 1$ for some $\theta' \in \Theta$. Then $f_x(\cdot)$ is conjugate to distributions in the family
>
> $$\mathcal{F} = \{g(x, \mathbf{m}, \theta) = h(x, \mathbf{m}, \theta)\pi(x), \ \mathbf{m} \in \mathcal{M}, \theta \in \Theta\},$$
>
> i.e., there exist functions $t : \mathcal{Y} \times \mathcal{M} \to \mathcal{M}$ and $T : \mathcal{Y} \times \Theta \to \Theta$ such that if $X_t \sim g(x, \mathbf{m}, \theta)$ and $Y_t | X_t = x \sim f_x$, we have $X_t | Y_y = y \sim g(x, t(y, \mathbf{m}), T(y, \theta))$.

Here $g(x, \mathbf{m}, \theta)$ takes the role of "current" prior distribution, i.e., the prior information on the signal state, possibly based on past observations through previous conditioning and propagations, and $g(x, t(y, \mathbf{m}), T(y, \theta))$ takes the role of the posterior, i.e., $g(x, \mathbf{m}, \theta)$ conditional on a data point y observed from f_x for $X_t = x$. The functions $t(y, \mathbf{m}), T(y, \theta)$ provide the transformations that update the parameters based on y. In absence of data, the condition $h(x, \mathbf{0}, \theta') = 1$ reduces $g(x, \mathbf{m}, \theta)$ to $\pi(x)$.

The third assumption weakens Assumption 3 in [53] by assuming the dual process has finite activity on a discrete state space, and possibly has a deterministic companion.

> **Assumption 3** (Duality). Given a deterministic process $\Theta_t \in \Theta$ and a regular jump continuous-time Markov chain M_t on $\mathbb{Z}_+^K$ with transition probabilities
>
> $$p_{\mathbf{m},\mathbf{n}}(t; \theta) := \mathbb{P}(M_t = \mathbf{n} | M_0 = \mathbf{m}, \Theta_0 = \theta), \tag{4}$$
>
> equation (1) holds with $D_t = (M_t, \Theta_t)$ and h as in Assumption 2.

The following result provides a full description of the propagation and update steps which allow to compute the filtering distribution.

Proposition 2.1. *Let Assumptions 1-3 hold, and let* $\sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}} g(x,\mathbf{m},\theta)$ *be a countable mixture with* $\sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}} = 1$. *Then, for* ψ_t *as in* (3) *we have*

$$\psi_t\Big(\sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}} g(x,\mathbf{m},\theta)\Big) = \sum_{\mathbf{n}\in\mathcal{M}} w'_{\mathbf{n}}(t) g(x,\mathbf{n},\Theta_t), \tag{5}$$

where

$$w'_{\mathbf{n}}(t) = \sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{n}} p_{\mathbf{m},\mathbf{n}}(t;\theta), \tag{6}$$

and $p_{\mathbf{m},\mathbf{n}}(t;\theta)$ *are as in* (4). *Furthermore, for* ϕ_y *as in* (2), *we have*

$$\phi_y\Big(\sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}} g(x,\mathbf{m},\theta)\Big) = \sum_{\mathbf{m}\in\mathcal{M}} \hat{w}_{\mathbf{n},\theta}(y) g(x,t(y,\mathbf{m}),T(y,\theta)) \tag{7}$$

where $\hat{w}_{\mathbf{m},\theta}(y) \propto w_{\mathbf{m}}\mu_{\mathbf{m},\theta}(y)$ *and*

$$\mu_{\mathbf{m},\theta}(y) := \int_{\mathcal{X}} f_x(y) g(x,\mathbf{m},\theta)\mathrm{d}x. \tag{8}$$

Proof. First observe that $\psi_t(g(x,\mathbf{m},\theta)) = \sum_{\mathbf{n}\in\mathcal{M}} p_{\mathbf{m},\mathbf{n}}(t;\theta) g(x,\mathbf{n},\Theta_t)$, which follows similarly to Proposition 2.2 in [53]. Then the claim follows by linearity using the fact that

$$\psi_t\Big(\sum_{i\geq 1} w_i\xi_i\Big) = \sum_{i\geq 1} w_i \psi_t(\xi_i)$$

so that

$$\begin{aligned}
&\psi_t\Big(\sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}} g(x,\mathbf{m},\theta)\Big) = \sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}} \psi_t(g(x,\mathbf{m},\theta)) \\
&= \sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}} \sum_{\mathbf{n}\in\mathcal{M}} p_{\mathbf{m},\mathbf{n}}(t;\theta) g(x,\mathbf{n},\Theta_t) = \sum_{\mathbf{n}\in\mathcal{M}}\sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t;\theta) g(x,\mathbf{n},\Theta_t)
\end{aligned}$$

Using now the fact that

$$\phi_y\Big(\sum_{i\geq 1} w_i\xi_i\Big) = \sum_{i\geq 1} \frac{w_i\mu_{\xi_i}(y)}{\sum_j w_j\mu_{\xi_j}(y)} \phi_y(\xi_i), \tag{9}$$

we also find that

$$\phi_y\Big(\sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}} g(x,\mathbf{m},\theta)\Big) = \sum_{\mathbf{m}\in\mathcal{M}} \frac{w_{\mathbf{m}}\mu_{\mathbf{m},\theta}(y)}{\sum_{\mathbf{n}\in\mathcal{M}} w_{\mathbf{n}}\mu_{\mathbf{n},\theta}(y)} g(x,t(y,\mathbf{m}),T(y,\theta))$$

where

$$\mu_{\mathbf{m},\theta}(y) = \int_{\mathcal{X}} f_x(y) g(x,\mathbf{m},\theta)\mathrm{d}x = \int_{\mathcal{X}} f_x(y) h(x,\mathbf{m},\theta)\pi(x)\mathrm{d}x.$$

□

The expression (5), together with (6), provides a general recipe on how to compute the forward propagation of the current marginal distribution of the signal $g(x, \mathbf{m}, \theta)$, based on the transition probabilities of the dual continuous-time Markov chain. Since the update operator (2) can be easily applied to the resulting distribution, Proposition 2.1 then shows that under these assumptions all filtering distributions are countable mixtures of elementary kernels indexed by the state space of the dual process, with mixture weights proportional to the dual process transition probabilities $p_{\mathbf{m},\mathbf{n}}(t;\theta)$. When these transition probabilities happen to give positive mass only to points $\{\mathbf{n} \in \mathcal{M} : \ \mathbf{n} \leq \mathbf{m}\}$, as is the case for a pure-death process, then the right-hand side of (5) reduces to a finite sum, and one can construct an exact filter with a computational cost that is polynomial in the number of observations, as shown in [53].

The above approach can be seen as an alternative to deriving the filtering distribution of the signal by leveraging on a spectral expansion of the transition function P_t in (3), which typically requires ad hoc computations and does not lend itself easily to explicit update operations through (2). Note also that expressions like (5) can be used, by taking appropriate limits of $p_{\mathbf{m},\mathbf{n}}(t;\theta)$ as $t \to 0$, to identify the transition kernel of the signal P_t itself, see, e.g., [7, 38, 53].

3 Recursive formulae for filtering and smoothing

In order to translate Proposition 2.1 into practical recursive formulae for filtering and smoothing, we are going to assume for simplicity of exposition that the time intervals between successive data collections $t_i - t_{i-1}$ equal Δ for all i. For ease of reading, we will therefore use the symbol P_Δ instead of $P_{t_i - t_{i-1}}$ for the signal transition function over the interval $\Delta = t_i - t_{i-1}$. We will also use the established notation whereby $i|0:i-1$ indicates that the argument refers to time $t_i = i\Delta$, and we are conditioning on the data collected at times from 0 to $t_{i-1} = (i-1)\Delta$.

Define the *filtering density*

$$\nu_{i|0:i}(x_i) := p(x_i|y_{0:i}) \propto \int_{\mathcal{X}^i} p(x_{0:i}, y_{0:i}) \mathrm{d}x_{0:i-1}, \tag{10}$$

i.e., the law of the signal at time t_i given data up to time t_i, obtained by integrating out the past trajectory. Define also the *predictive density*

$$\nu_{i+1|0:i}(x_i) := p(x_{i+1}|y_{0:i}) = \int_{\mathcal{X}} p(x_i|y_{0:i}) P_\Delta(x_{i+1}|x_i) \mathrm{d}x_i, \tag{11}$$

i.e, the marginal density of the signal at time t_{i+1}, given data up to time t_i. This can be expressed recursively as a function of the previous *filtering density* $p(x_i|y_{0:i})$, as displayed. Finally, define the marginal *smoothing density*

$$\nu_{i|0:n}(x_i) := p(x_i|y_{0:n}) \propto \int_{\mathcal{X}^n} p(x_{0:n}, y_{0:n}) \mathrm{d}x_{0:i-1} \mathrm{d}x_{i+1:n}, \tag{12}$$

where the signal is evaluated at time t_i conditional on all available data. The first two distributions above are natural objects of inferential interest, whereas the latter is typically used to improve previous estimates once additional data become available. Finally, for Θ_Δ as in Assumption 3 and $t(\cdot,\cdot), T(\cdot,\cdot)$ as in Assumption 2, define for $i = 0, \dots, n$ the quantities

$$\vartheta_{i|0:i} := T(y_i, \vartheta_{i|0:i-1}), \qquad \vartheta_{i|0:i-1} := \Theta_\Delta(\vartheta_{i-1|0:i-1}), \qquad \vartheta_{0|0:-1} := \theta_0. \tag{13}$$

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.

The following Corollary of Proposition 2.1 extends a result of [53] (see also Theorem 1 in [50] for an easier comparison in a similar notation).

Corollary 3.1. *Let Assumptions 1-3 hold, and assume that*

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

Then (11) *can be written, through* (3)*, as*

$$\begin{aligned} \nu_{i|0:i-1}(x) &= \psi_\Delta(\nu_{i-1|0:i-1}(x)) = \sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}}^{(i-1)'} g(x, \mathbf{m}, \vartheta_{i|0:i-1}), \\ w_{\mathbf{m}}^{(i-1)'} &= \sum_{\mathbf{n}\in\mathcal{M}} w_{\mathbf{n}}^{(i-1)} p_{\mathbf{n},\mathbf{m}}(\Delta; \vartheta_{i-1|0:i-1}), \quad \mathbf{m} \in \mathcal{M}, \end{aligned} \tag{14}$$

with $p_{\mathbf{n},\mathbf{m}}(\Delta; \vartheta_{i|0:i})$ *as in* (4)*. Furthermore, given the observation* y_i*,* (10) *can be written, through* (2)*, as*

$$\begin{aligned} \nu_{i|0:i}(x) &= \phi_{y_i}(\nu_{i|0:i-1}(x)) = \sum_{\mathbf{m}\in\mathcal{M}} w_{\mathbf{m}}^{(i)} g(x, \mathbf{m}, \vartheta_{i|0:i}), \\ w_{\mathbf{m}}^{(i)} &\propto \mu_{\mathbf{n},\vartheta_{i|0:i-1}}(y_i) w_{\mathbf{n}}^{(i-1)'}, \quad \mathbf{m} = t(y_i, \mathbf{n}), \mathbf{n} \in \mathcal{M}, \end{aligned} \tag{15}$$

with $\mu_{\mathbf{m},\theta}$ *as in* (8)*.*

Algorithm 1 outlines the pseudo-code for implementing the update and propagation steps of Corollary 3.1. How to use these results efficiently can depend on the model at hand. When the transition probabilities $p_{\mathbf{m},\mathbf{n}}(t;\theta)$ are available in closed form, their use could lead to the best performance, but can also at times face numerical instability issues (as is the case pointed out in Section 4 below). When the transition probabilities $p_{\mathbf{m},\mathbf{n}}(t;\theta)$ are not available in closed form, one can approximate them by simulating N replicates of the dual component M_t, and then regroup probability masses according to the arrival states as done in (14). In our framework, the dual process is typically easier to simulate than the original process, given its discrete state space. For instance, pure-death or

Algorithm 1: Filtering

Input: $Y_{0:n}$, $t_{0:n}$

Result: $\vartheta_{i|0:i}$, $\mathbf{M}_{i|0:i}$ and $W_i = \{w^{(i)}_{\mathbf{m}}, \mathbf{m} \in \mathbf{M}_{i|0:i}\}$

Initialise

Set $\vartheta_{0|0} = T(Y_0, \theta_0)$ with T as in Assumption 2

Set $\mathbf{M}_{0|0} = \{t(Y_0, \mathbf{0})\} = \{\mathbf{m}^*\}$ and $W_0 = \{1\}$ with t as in Assumption 2

Compute $\vartheta_{1|0}$ from $\vartheta_{0|0}$ as in (13)

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 $p_{\mathbf{m},\mathbf{n}}$ as in (4)

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

Update

Set $\vartheta_{i|0:i} = T(Y_i, \vartheta_{i|0:i-1})$

Set $W_i = \left\{ \frac{w^*_{\mathbf{m}} \mu_{\mathbf{m},\vartheta_{i|0:i-1}}(Y_i)}{\sum_{\mathbf{n}\in\mathbf{M}^*} w^*_{\mathbf{n}} \mu_{\mathbf{n},\vartheta_i}(Y_i)}, \mathbf{m} \in \mathbf{M}^* \right\}$ with $\mu_{\mathbf{m},\theta}$ defined as in (8)

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

Propagation

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

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

end

Note: $\mathbf{M}_{i|0:i} = \{\mathbf{m} \in \mathcal{M} : \ w^{(i)}_{\mathbf{m}} > 0\} \subset \mathcal{M}$ is the support of the weights of $\nu_{i|0:i}$; $\mathcal{B}(\mathbf{m})$ denotes the states reached by the dual process from $\mathbf{m}$, and $\mathcal{B}(\mathbf{M})$ those reached from all $\mathbf{m} \in \mathbf{M}$.

B&D processes are easily simulated using a Gillespie algorithm [36], whereby one alternates sampling waiting times and jump transitions for the embedded chain. Depending on the dual process, there might also be more efficient simulation strategies.

A different type of approximation of the propagation step (14) in Corollary 3.1 can be based on pruning the transition probabilities or the arrival weights under a given threshold, followed by a renormalisation of the weights. Both this approximation strategy and that outlined above assign positive weights only to a finite subset of $\mathcal{M}$, hence they overcome the infinite dimensionality of the dual process state space. [50] showed that the latter strategy allows to control the approximation error while retaining almost entirely the distributional information, thus affecting the inference negligibly. In the next sections we will investigate such strategies for two hidden Markov models driven by Cox–Ingersoll–Ross and K-dimensional Wright–Fisher diffusions.

Now, in order to describe the marginal smoothing densities (12), we need an additional assumption and some further notation.

Assumption 4 For h as in Assumption 3, there exist functions $d : \mathcal{M}^2 \to \mathcal{M}$ and $e : \Theta^2 \to \Theta$ such that for all $x \in \mathcal{X}$, $\mathbf{m}, \mathbf{m}' \in \mathcal{M}$, $\theta, \theta' \in \Theta$

$$h(x, \mathbf{m}, \theta) h(x, \mathbf{m}', \theta') = C_{\mathbf{m},\mathbf{m}',\theta,\theta'} h(x, d(\mathbf{m}, \mathbf{m}'), e(\theta, \theta')), \tag{16}$$

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

Denote by $\overleftarrow{\vartheta}_i, \overleftarrow{\vartheta}'_i$ the quantities defined in (13) computed backwards. Equivalently, these are computed as in (13) with data in reverse order, i.e. using $y_{n:0}$ in place of $y_{0:n}$, namely

$$\overleftarrow{\vartheta}_{i|i+1:T} = \Theta_\Delta(\overleftarrow{\vartheta}_{i+1|i+1:T}), \qquad \overleftarrow{\vartheta}_{i|i:T} = T(y_i, \overleftarrow{\vartheta}_{i|i+1:T}), \qquad \overleftarrow{\vartheta}_{T|T} = T(y_T, \theta_0)$$

The following result extends Proposition 3 and Theorem 4 of [50]:

Proposition 3.2. *Let Assumptions 1-4 hold, and let $\nu_0 = \pi$. Then, for $0 \leq i \leq n-1$, we have*

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

with

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

$$\overleftarrow{\omega}^{(i+1)}_{\mathbf{m}} = \sum_{\mathbf{n}\in\mathcal{M}} \overleftarrow{\omega}^{(i+2)}_{\mathbf{n}} \mu_{\mathbf{n},\overleftarrow{\vartheta}_{i+1|i+2:n}}(y_{i+1}) p_{t(y_{i+1},\mathbf{n}),\mathbf{m}}(\Delta; \overleftarrow{\vartheta}_{i+1|i+1:n})$$

$w^{(i)}_{\mathbf{n}}$ *as in* (15) *and* $C_{\mathbf{m},\mathbf{n},\overleftarrow{\vartheta}_{i|i+1:n},\vartheta_{i|0:i}}$ *as in* (16).

Proof. Note that Bayes' Theorem and conditional independence allow to write (12) as

$$\nu_{i|0:n}(x_i) = p(x_i|y_{0:n}) \propto p(y_{i+1:n}|x_i)\nu_{i|0:i}(x_i)$$

where the right-hand side involves the filtering distribution, available from Corollary 3.1, and the so called *cost-to-go function* $p(y_{i+1:n}|x_i)$ (sometimes called information filter), which is the likelihood of future observations given the current signal state. Along the same lines as Proposition 3 in [50] we find that

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

with $\overleftarrow{\omega}^{(i+1)}_{\mathbf{m}}$ as in the statement. The main claim can now be proved along the same lines as Theorem 4 in [50]. □

The main difference between the above result and Theorem 4 in [50] lies in the fact that the support of the weights $\{\overleftarrow{\omega}^{(i+1)}_{\mathbf{m}}, \mathbf{m} \in \mathcal{M}\}$ (which in [50] is denoted by $\overleftarrow{M}_{i|i+1:n}$) can be countably infinite and coincide with the whole of $\mathcal{M}$. Indeed, which points of $\mathcal{M}$ have positive weight are determined by the transition probabilities of the dual process, which in the present framework is no longer assumed to make only downward moves in $\mathcal{M}$. Section 6 will deal with this possibly infinite support for a concrete implementation of the inferential strategy.

4 Cox–Ingersoll–Ross hidden Markov models

The Cox–Ingersoll–Ross diffusion, also known as the square-root process, is widely used in financial mathematics for modelling short-term interest rates and stochastic volatility, see [9, 10, 29, 37, 40]. It also belongs to the class of continuous-state branching processes with immigration, arising as the large-population scaling limit of certain branching Markov chains [49] and as the time-evolving total mass of a Dawson–Watanabe branching measure-valued diffusion [20].

Let X_t be a CIR diffusion on $\mathbb{R}_+$ that solves the one-dimensional SDE

$$\mathrm{d}X_t = \big(\delta\sigma^2 - 2\gamma X_t\big)\,dt + 2\sigma\sqrt{X_t}dB_t, \qquad X_0 \geq 0, \tag{17}$$

where $\delta, \gamma, \sigma > 0$, which is reversible with respect to the Gamma density $\pi = \mathrm{Ga}(\delta/2, \gamma/\sigma^2)$. The following proposition identifies a B&D process as dual to the CIR diffusion.

Proposition 4.1. *Let X_t be as in (17), let M_t be a* B&D *process on $\mathbb{Z}_+$ which jumps from m to $m+1$ at rate $\lambda_m = 2\sigma^2(\delta/2+m)(\theta-\gamma/\sigma^2)$ and to $m-1$ at rate $\mu_m = 2\sigma^2\theta m$, and let*

$$h(x,m,\theta) = \frac{\Gamma(\delta/2)}{\Gamma(\delta/2+m)}\Big(\frac{\gamma}{\sigma^2}\Big)^{-\delta/2}\theta^{\delta/2+m}x^m e^{-(\theta-\gamma/\sigma^2)x}. \tag{18}$$

Then (1) holds with $D_t = M_t$.

Proof. The infinitesimal generator associated to (17) is

$$\mathcal{A}f(x) = (\delta\sigma^2 - 2\gamma x)f'(x) + 2\sigma^2 x f''(x),$$

for $f : \mathbb{R}_+ \to \mathbb{R}$ vanishing at infinity. Letting $h(x,m) = h(x,m,\theta)$ denote (18) omitting the dependence on θ to make notations lighter, a direct computation yields

$$\begin{aligned}
\mathcal{A}h(\cdot,m)(x) =& (\delta\sigma^2 - 2\gamma x)\Big(mx^{m-1} - x^m(\theta-\gamma/\sigma^2)\Big)\frac{\Gamma(\delta/2)}{\Gamma(\delta/2+m)}\Big(\frac{\gamma}{\sigma^2}\Big)^{-\delta/2}\theta^{\delta/2+m}e^{-(\theta-\gamma/\sigma^2)x}\\
&+ 2\sigma^2 x\Big(m(m-1)x^{m-2} + x^m(\theta-\gamma/\sigma^2)^2 - 2mx^{m-1}(\theta-\gamma/\sigma^2)\Big)\\
&\times \frac{\Gamma(\delta/2)}{\Gamma(\delta/2+m)}\Big(\frac{\gamma}{\sigma^2}\Big)^{-\delta/2}\theta^{\delta/2+m}e^{-(\theta-\gamma/\sigma^2)x}\\
=& \frac{\delta\sigma^2 m\theta}{\delta/2+m-1}h(x,m-1) + 2\gamma(\theta-\gamma/\sigma^2)\frac{\delta/2+m}{\theta}h(x,m+1)\\
&- [2\gamma m + \delta\sigma^2(\theta-\gamma/\sigma^2)]h(x,m) + 2\sigma^2 m(m-1)\frac{\theta}{\delta/2+m-1}h(x,m-1)\\
&+ 2\sigma^2(\theta-\gamma/\sigma^2)^2\frac{\delta/2+m}{\theta}h(x,m+1) - 4\sigma^2 m(\theta-\gamma/\sigma^2)h(x,m)\\
=& 2\sigma^2\theta m h(x,m-1) + 2\sigma^2(\delta/2+m)(\theta-\gamma/\sigma^2)h(x,m+1)\\
&- [2\gamma m + \sigma^2(\delta+4m)(\theta-\gamma/\sigma^2)]h(x,m).
\end{aligned}$$

where it can be checked that

$$2\sigma^2\theta m + 2\sigma^2(\delta/2+m)(\theta-\gamma/\sigma^2) = 2\gamma m + \sigma^2(\delta+4m)(\theta-\gamma/\sigma^2).$$

Hence the r.h.s. equals

$$\mathcal{B}g(m) = \lambda_m[g(m+1)-g(m)] + \mu_m[g(m-1)-g(m)]$$

with $g(\cdot) := h(x,\cdot)$, $\lambda_m = 2\sigma^2(\delta/2+m)(\theta-\gamma/\sigma^2)$, and $\mu_m = 2\sigma^2\theta m$, which is the infinitesimal generator of a B&D process with rates λ_m, μ_m. The claim now follows from Proposition 1.2 in [44]. □

Assign now prior $\nu_0 = \mathrm{Ga}(\delta/2, \gamma/\sigma^2)$ to X_0, and assume Poisson observations are collected at equally spaced intervals of length Δ. Specifically, $Y|X_t = x \overset{\text{iid}}{\sim} \mathrm{Po}(\tau x)$, for some $\tau > 0$. By the well-known conjugacy to Gamma priors, we have that $X_t|Y = y \sim \mathrm{Ga}(\delta/2+y, \gamma/\sigma^2+\tau)$. Without loss of generality, we can set $\tau = 1$, which allows to interpret the update of the gamma rate parameter as the size of the conditioning data set. The filtering algorithm starts by first updating the prior ν_0 to $\nu_{0|0} := \phi_{Y_0}(\nu_0)$. If we observe $Y_0 = (Y_{0,1}, \dots, Y_{0,k})$ at time 0, then $\nu_{0|0}$ is the law of $X_0|\sum_{j=1}^k y_{0,j} = m \sim \mathrm{Ga}(\delta/2+m, \gamma/\sigma^2+k)$. Then $\nu_{0|0}$ is propagated forward for a Δ time interval, yielding $\nu_{1|0} := \psi_\Delta(\nu_{0|0})$. In light of Proposition 4.1, an application of (5) to $\nu_{0|0}$ yields the infinite mixture

$$\psi_\Delta\left(\mathrm{Ga}(\delta/2+m, \gamma/\sigma^2+k)\right) = \sum_{n\geq 0} p_{m,n}(\Delta)\mathrm{Ga}(\delta/2+n, \gamma/\sigma^2+k), \tag{19}$$

where $p_{m,n}(t)$ are the transition probabilities of M_t in Proposition 4.1. Hence, the law of the signal is indexed by $\mathbb{Z}_+$, the state space of the dual process. While after the update at time 0 mass one is assigned to the sum of the observations $\sum_{j=1}^k y_{0,j} = m$, after the propagation the mass is spread over the whole $\mathbb{Z}_+$ by the effect of the dual process. We then observe $Y_1 \sim f_{X_1}$ at time 1, which is used to update $\nu_{1|0}$ to $\nu_{1|1}$ and has the effect of shifting the probability masses of the mixture weights. For example, the weight $p_{m,n}(\Delta)$ in (19) is assigned to $n \in \mathbb{Z}_+$, but after the update based on $Y_1 = (Y_{1,1}, \dots, Y_{1,k'})$ it will be assigned to $n+m'$ if $\sum_{j=1}^{k'} y_{1,j} = m'$, on top of being transformed according to (9). We then propagate forward again and proceed analogously.

When the current distribution of the signal, after the update, is given by a mixture of type $\sum_{m\in\mathbb{Z}_+} w_m \mathrm{Ga}(\delta/2+m, \gamma/\sigma^2+k)$, it is enough to rearrange the mixture weights after the propagation step as done in (6).

The main difference with qualitatively similar equations found in [50] is now given by the transition probabilities $p_{m,n}(t)$ in (19), which are those of the B&D process in Proposition 4.1. Before tackling the problem of how to use the above expressions for inference, we try to provide further intuition of the extent and implications of such differences. To this end, consider the simplified parameterization $\alpha = \delta/2, \beta = \gamma/\sigma^2, \sigma^2 = 1/2, \tau = 1$, whereby one can check that the embedded chain of the B&D process of Proposition 4.1 has jump probabilities

$$p_{m,m+1} = \frac{k(\alpha+m)}{k(\alpha+m)+m(\beta+k)}, \qquad p_{m,m-1} = 1 - p_{m,m+1}.$$

Here m, k are the same as in the left-hand side of (19), so m/k is the sample mean. It is easily verified that $p_{m,m+1} < p_{m,m-1}$ if $m/k > \alpha/\beta$ and viceversa. Therefore, the dual evolves on $\mathbb{Z}_+$ so that it reverts m/k to the prior mean α/β. Indeed, the dual has Negative Binomial ergodic distribution $\mathrm{NBin}\,(\alpha, \beta/(\beta + k))$, whose mean is $k\alpha/\beta$, i.e., such that m/k on average coincides with α/β.

Recall now that the dual process elicited in [53] for the CIR model is $D_t = (M_t, \Theta_t)$, with M_t a pure-death process with rates from m to $m-1$ equal to $2\sigma^2\theta$ and Θ_t a deterministic process that solves $\mathrm{d}\Theta_t/\mathrm{d}t = -2\sigma^2\Theta_t(\Theta_t - \gamma/\sigma^2)$, $\Theta_0 = \theta$. This dual has a single ergodic state given by $(0, \beta)$ (note that [53] uses a slightly different parameterization, where the ergodic state $(0, \beta)$ means that, in the limit for $t \to \infty$, the gamma parameters are the prior parameters). In particular, as $t \to \infty$, this entails the convergence of $p_{m,n}(t)$ in (19) to 1 if $n = 0$ and 0 elsewhere. Whence the strong ergodic convergence $\psi_t(g(x, m, \theta)) \to \pi$ as $t \to \infty$, whereby the effect of the observed data become progressively negligible as t increases. One could then argue that in the long run, the filtering strategy based on the pure-death dual process in [53] completely forgets the collected data. As a consequence, one could expect filtering with long-spaced observations (relative to the forward process autocorrelation) to be similar to using independent priors at each data collection point. On the other hand, the B&D dual can be thought as not forgetting but rather spreading around the probability mass in such a way as to preserve convergence of the empirical mean to the prior mean. It is not obvious a priori which of these two scenarios could be more beneficial in terms of filtering, hence in Section 6.1 we provide numerical experiments for comparing the performance of strategies based on these different duals.

In view of such experiments, note that the transition probabilities of the above B&D dual are in principle available in closed form (cf. [5, 17]), but their computation is prone to numerical instability. Alternatively, we can approximate the transition probabilities $p_{m,n}(t)$ in (19) by drawing N sample paths of the dual started in m and use the empirical distribution of the arrival points. This can in principle be done through the Gillespie algorithm [36], which alternates sampling waiting times and jumps of the embedded chain. A faster strategy can be achieved by writing the B&D rates in Proposition 4.1 as $\lambda_m = \lambda m + \beta$ and $\mu_m = \mu m$ with

$$\lambda = 2\sigma^2(\theta - \gamma/\sigma^2), \qquad \beta = \sigma^2\delta(\theta - \gamma/\sigma^2), \qquad \mu = 2\sigma^2\theta,$$

where λ, μ represent the per capita birth and death rate and β is the immigration rate. Then write $M_t = A_t + B_t$ where A_t is the population size of the descendant of autochthonous individuals (already in the population at $t = 0$), and B_t the descendants of the immigrants. These rates define a linear B&D process, whereby [57] suggests simulating A_t by drawing, given $A_0 = i$,

$$F \sim \mathrm{Bin}(i, g(t)), \qquad A_t \sim \mathrm{NBin}(F, h(t)) + F, \tag{20}$$

with $h(t) = (\lambda - \mu)/(\lambda \exp\{(\lambda - \mu)t\} - \mu)$ and $g(t) = h(t)\exp\{(\lambda - \mu)t\}$, with the convention $\mathrm{NBin}(0, p) = \delta_0$. Let now N_s be the number of immigrants up to time s, which follows a simple Poisson process with rate β, so given N_t the arrival times are uniformly distributed on $[0, t]$. Once in the population, the lineage of each immigrating individual follows again a B&D process and can

be simulated using (20) starting at $i = 1$. Summing the numerosity of each immigrant family at time t yields B_t.

5 Wright–Fisher hidden Markov models

The K-dimensional WF diffusion is a widely studied classical model in population genetics (see [19, 21, 22, 41] and references therein), recently used also in a statistical framework [50, 53]. See also [12] for connections with statistical physics. It takes values in the simplex

$$\Delta_K = \left\{ \mathbf{x} \in [0,1]^K : \sum_{1 \le i \le K} x_i = 1 \right\}$$

and, in the population genetics interpretation, it models the temporal evolution of K proportions of types in an underlying large population. Its infinitesimal generator on $C^2(\Delta_K)$ is

$$\mathcal{A} = \frac{1}{2} \sum_{i,j=1}^{K} x_i(\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j} + \frac{1}{2} \sum_{i=1}^{K} (\alpha_i - \theta x_i) \frac{\partial}{\partial x_i} \tag{21}$$

for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$, $\theta = \sum_{i=1}^K \alpha_i$, and its reversible measure is the Dirichlet distribution whose density with respect to Lebesgue measure is

$$\pi_{\boldsymbol{\alpha}}(\mathbf{x}) = \frac{\Gamma(\theta)}{\prod_{i=1}^K \Gamma(\alpha_i)} x_1^{\alpha_1 - 1} \cdots x_K^{\alpha_K - 1}, \quad x_K = 1 - \sum_{i=1}^{K-1} x_i.$$

See for example [21], Chapter 10. The transition density of this model is (cf., e.g., [19], eqn. (1.27))

$$p_t(\mathbf{x}, \mathbf{x}') = \sum_{m=0}^{\infty} d_m(t) \sum_{\mathbf{m} \in \mathbb{Z}_+^K : |\mathbf{m}| = m} \mathrm{MN}(\mathbf{m}; m, \mathbf{x}) \pi_{\boldsymbol{\alpha} + \mathbf{m}}(\mathbf{x}'), \tag{22}$$

where $\mathrm{MN}(\mathbf{m}; m, \mathbf{x}) = \binom{m}{m_1, \ldots, m_K} \prod_{i=1}^K x_i^{m_i}$, and where $d_m(t)$ are the transition probabilities of the block counting process of Kingman's coalescent on $\mathbb{Z}_+$, which has an entrance boundary at ∞. Cf., e.g., [19], eqn. (1.12).

It is well known that a version of Kingman's typed coalescent with mutation is dual to the WF diffusion. This can be seen as a death process on Z_+^K which jumps from $\mathbf{m}$ to $\mathbf{m} - \mathbf{e}_i$ at rate

$$q_{\mathbf{m}, \mathbf{m} - \mathbf{e}_i} = m_i(\theta + |\mathbf{m}| - 1)/2. \tag{23}$$

Here $\mathbf{e}_i$ is the canonical vector in the i-th direction. See, for example, [23, 24, 39]; see also [53], Section 3.3. The above death process with transitions $d_m(t)$ is indeed the process that counts the surviving blocks of the typed version without keeping track of which types have been removed.

Recall now that a Moran model with N individuals of K types is a particle process with overlapping generations whereby at discrete times a uniformly chosen individual is removed and another, uniformly chosen from the remaining individuals, produces one offspring of its own type, leaving the total population size constant. See, e.g., [22]. In the presence of mutation, upon reproduction, the offspring can mutate to type j at parent-independent rate α_j. The generator of such process on the set $B(\mathbb{Z}_+^K)$ of bounded functions on $\mathbb{Z}_+^K$ can be written in terms of the multiplicities of types $\mathbf{n} \in \mathbb{Z}_+^K$ as

$$\mathcal{B}f(\mathbf{n}) = \frac{1}{2} \sum_{1 \le i \ne j \le K} n_i(\alpha_j + n_j) f(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) - \frac{1}{2} \sum_{1 \le i \ne j \le K} n_i(\alpha_j + n_j) f(\mathbf{n}), \tag{24}$$

where an individual of type i is removed at rate n_i, the number of individuals of type i, and is replaced by an individual of type j at rate $\alpha_j + n_j$.

The following proposition extends a result in [12] (cf. Section 5) and shows that the above Moran model is dual to the WF diffusion with generator (21).

Proposition 5.1. *Let X_t have generator (21), let $N_t \in \mathbb{Z}_+^K$ be a Moran model which from $\mathbf{n}$ jumps to $\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j$ at rate $n_i(\alpha_j + n_j)/2$, and let*

$$h(\mathbf{x}, \mathbf{n}) = \frac{\Gamma(\theta + |\mathbf{n}|)}{\Gamma(\theta)} \prod_{i=1}^K \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + n_i)} x_i^{n_i}, \qquad \theta = \sum_{i=1}^K \alpha_i.$$

Then (1) holds with $D_t = N_t$ and h as above.

Proof. From (21), since $\theta = \sum_{i=1}^K \alpha_i$, we can write

$$\begin{aligned} 2\mathcal{A} &= \sum_{1 \le i \le K} x_i(1 - x_i) \frac{\partial^2}{\partial x_i^2} - \sum_{1 \le i \ne j \le K} x_i x_j \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{1 \le i \le K} (\alpha_i(1 - x_i) - x_i \sum_{1 \le j \le K, j \ne i} \alpha_j) \frac{\partial}{\partial x_i} \\ &= \sum_{1 \le i \ne j \le K} x_i x_j \frac{\partial^2}{\partial x_i^2} - \sum_{1 \le i \ne j \le K} x_i x_j \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{1 \le i \le K} \alpha_i \sum_{1 \le j \le K, j \ne i} x_j \frac{\partial}{\partial x_i} - \sum_{1 \le i \ne j \le K} \alpha_j \frac{\partial}{\partial x_i}. \end{aligned}$$

Then one can check that

$$\begin{aligned} 2\mathcal{A}h(\mathbf{x}, \mathbf{n}) &= \sum_{1 \le i \ne j \le K} n_i(\alpha_i + n_i - 1) \frac{\Gamma(\theta + |\mathbf{n}|)}{\Gamma(\theta)} \mathbf{x}^{\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j} \prod_{h=1}^K \frac{\Gamma(\alpha_h)}{\Gamma(\alpha_h + n_h)} \\ &\quad - \sum_{1 \le i \ne j \le K} n_i(\alpha_j + n_j) \mathbf{x}^{\mathbf{n}} \frac{\Gamma(\theta + |\mathbf{n}|)}{\Gamma(\theta)} \prod_{h=1}^K \frac{\Gamma(\alpha_h)}{\Gamma(\alpha_h + n_h)} \\ &= \sum_{1 \le i \ne j \le K} n_i(\alpha_j + n_j) h(\mathbf{x}, \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) - \sum_{1 \le i \ne j \le K} n_i(\alpha_j + n_j) h(\mathbf{x}, \mathbf{n}). \end{aligned}$$

Hence we have

$$(\mathcal{A}h(\cdot,\mathbf{n}))(\mathbf{x}) = (\mathcal{B}h(\mathbf{x},\cdot))(\mathbf{n})$$

where the right hand side is (24) applied to $h(\mathbf{x},\mathbf{n})$ as a function of $\mathbf{n}$. The claim now follows from Proposition 1.2 in [44]. □

Assign now prior $\nu_0 = \pi_{\boldsymbol{\alpha}}$ to X_0, and assume categorical observations so that $\mathbb{P}(Y = j|X_t = x) = x_j$. By the well-known conjugacy to Dirichlet priors, we have $X_t|Y = y \sim \pi_{\boldsymbol{\alpha}+\delta_y}$, where $\boldsymbol{\alpha} + \delta_y = (\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_K)$ if $y = j$. When multiple categorical observations with vector of multiplicities $\mathbf{m} \in \mathbb{Z}_+^K$ are collected, we write $\pi_{\boldsymbol{\alpha}+\mathbf{m}}$. The filtering algorithm then proceeds by first updating ν_0 to $\nu_{0|0} := \phi_{Y_0}(\nu_0) = \pi_{\boldsymbol{\alpha}+\mathbf{m}}$, if $Y_0 = (Y_{0,1}, \ldots, Y_{0,k})$ yields multiplicities $\mathbf{m}$, then propagating $\nu_{0|0}$ to $\nu_{1|0} := \psi_\Delta(\nu_{0|0})$. In light of the previous result, an application of (5) to $\nu_{0|0} = \pi_{\boldsymbol{\alpha}+\mathbf{m}}$ yields the mixture

$$\psi_\Delta\left(\pi_{\boldsymbol{\alpha}+\mathbf{m}}\right) = \sum_{\mathbf{n}:|\mathbf{n}|=|\mathbf{m}|} p_{\mathbf{m},\mathbf{n}}(\Delta)\pi_{\boldsymbol{\alpha}+\mathbf{n}}, \tag{25}$$

where $p_{\mathbf{m},\mathbf{n}}(\Delta)$ are the transition probabilities of N_t in Proposition 5.1 over the interval Δ. We then observe $Y_1|X_1$, which is in turn used to update $\nu_{1|0}$ to $\nu_{1|1}$, as so forth. We refer again the reader to [50], Section 2.4.2, for details on qualitatively similar recursive formulae.

In (25), the overall multiplicity $|\mathbf{n}|$ equals the original $|\mathbf{m}|$, as an effect of the population size preservation provided by the Moran model. The space $\{\mathbf{n} : |\mathbf{n}| = |\mathbf{m}|\}$ is finite, which shows that Assumption 3 need not require the presence of a death-like process to have filtering distributions being finite mixtures. However, it is not obvious *a priori* how (25) compares in terms of practical implementation with the different representation obtained in [53], namely

$$\psi_\Delta\left(\pi_{\boldsymbol{\alpha}+\mathbf{m}}\right) = \sum_{\mathbf{n}:|\mathbf{n}|\leq|\mathbf{m}|} \hat{p}_{\mathbf{m},\mathbf{n}}(\Delta)\pi_{\boldsymbol{\alpha}+\mathbf{n}} \tag{26}$$

where $\hat{p}_{\mathbf{m},\mathbf{n}}(\Delta)$ are the transition probabilities of the death process on $\mathbb{Z}_+^K$ with rates (23). Similarly to what has already been discussed for the CIR case, the death process dual has a single ergodic state given by the origin $(0, \ldots, 0)$, which entails the convergence of $\hat{p}_{\mathbf{m},\mathbf{n}}(t)$ to 1 if $\mathbf{n} = (0, \ldots, 0)$ and 0 elsewhere, implying the strong convergence $\psi_t\left(\pi_{\boldsymbol{\alpha}+\mathbf{m}}\right) \to \pi_{\boldsymbol{\alpha}}$ in (26). This is ultimately determined by the fact that Kingman's coalescent removes lineages by coalescence and mutation until absorption to the empty set.

At first glance, a similar convergence is seemingly precluded to (25). However, we note in the first sum of (24) that the new particle's type is either resampled from the survived particles or drawn from the baseline distribution, in which case the new particle is of type j with (parent-independent) probability $\alpha_j / \sum_{i=1}^K \alpha_i$. Hence each particle will be resampled from the baseline distribution in finite time. Together with the fact that $\sum_{j=1}^K \pi_{\boldsymbol{\alpha}+\delta_j}\alpha_j / \sum_{i=1}^K \alpha_i = \pi_{\boldsymbol{\alpha}}$, which follows from Proposition G.9 in [32], and considering that the number of particles is finite, we can therefore expect that, as $t \to \infty$, we have the convergence $\psi_t\left(\pi_{\boldsymbol{\alpha}+\mathbf{m}}\right) \to \pi_{\boldsymbol{\alpha}}$ in (25) also for this case.

The transition probabilities $p_{\mathbf{m},\mathbf{n}}(t)$ in (25), induced by the Moran model, are not available in closed form. This poses a limit on the direct applicability of the presented algorithms for numerical experiments. The first alternative is then to approximate them by drawing N points from the discrete distribution on the dual space before the propagation, making use of the Gillespie algorithm to draw as many paths, and evaluating the empirical distribution of the arrival points. Alternative approximations are suggested by the fact that an appropriately rescaled version of the Moran model converges in distribution to a WF diffusion (see, e.g., [22], Lemma 2.39). Indeed, a spatial rescaling of the Moran model in (24) to get proportions in place of multiplicities of types results in the generator

$$\mathcal{C}_{|\mathbf{n}|}f(\mathbf{x}) = \sum_{1\leq i\neq j\leq K} \frac{n_i}{|\mathbf{n}|}\frac{\alpha_j + n_j}{|\mathbf{n}|}\left[f\left(\mathbf{x} - \frac{\mathbf{e}_i}{|\mathbf{n}|} + \frac{\mathbf{e}_j}{|\mathbf{n}|}\right) - f(\mathbf{x})\right],$$

where $x_i := n_i/|\mathbf{n}|$. A classical argument based on a Taylor expansion for f now leads to write $|\mathbf{n}|^2\mathcal{C}_{|\mathbf{n}|}f = \mathcal{A}f + O(|\mathbf{n}|^{-1})$, with $\mathcal{A}$ as in (21) and where $O(|\mathbf{n}|^{-1})$ represent a remainder term which goes to zero with $|\mathbf{n}|^{-1}$. The claim then be based on classical arguments following, e.g., [21], Theorem 4.8.7. We could therefore use a WF diffusion to approximate the Moran dual transitions in (25). Since the spatially rescaled Moran model takes values $\mathbf{m}/|\mathbf{n}|$ with $\mathbf{m} \in \mathcal{M}$ such that $0 \leq |\mathbf{m}| \leq |\mathbf{n}|$, to the above end it suffices to discretize the states of the WF diffusion through binning, e.g., given a state $\mathbf{x}$ of the approximating WF diffusion, we take as state of the Moran model the point $[|\mathbf{n}|\mathbf{x}] := ([|\mathbf{n}|x_1], \ldots, [|\mathbf{n}|x_K])$, where $[|\mathbf{n}|x_i]$ is the approximation of $|\mathbf{n}|x_i$ to the closest integer in $\{0, \ldots, |\mathbf{n}|\}$. The functionals of interest can thus be evaluated through the same procedure which uses the original Moran model, i.e., through (25), based on the WF diffusion transition probabilities. This strategy in principle has the drawback of having to deal with the intractable terms $d_m(t)$ in the transition function expansion (22) of the diffusion, hurdle overcome by adopting the solution proposed by [45].

It is also known that one could also construct a sequence of WF discrete Markov chains with non-overlapping generations indexed by the population size which, upon appropriate rescaling, converge weakly to the desired WF diffusion (see, e.g., [48], Sec. 15.2.F or [22], Sec 4.1). Since two sequences that converge to the same limit can to some extent be considered close to each other, one could then consider a WF discrete chain indexed by $|\mathbf{n}|$ with a parameterization that would make it converge to (21), and use it to approximate the Moran transition probabilities. This would permit a straightforward implementation, given WF discrete chains have multinomial transitions. In Section 6.2 we compare the performance of all the above mentioned strategies.

6 Numerical experiments

To illustrate how the above results can be used in practice and how they perform in comparison with other methods, we are going to consider particle approximations of the dual processes for evaluating their transition probabilities, which in turn are used in (6) in place of the true transition probabilities to evaluate the predictive distributions for the signal, denoted here $\hat{p}(x_{k+1}|y_{1:k})$. We compare

these distributions with the exact predictive distribution obtained through the results in [53] and those obtained through *bootstrap particle filtering* which make use of the signal transition function. Particle filtering can be considered the state of the art for this type of inferential problems, a general reference being [15]. The notable difference between these two approaches is that bootstrap particle filtering operates on the original state space of the signal, whereas filtering based on dual processes index the filtering mixtures using the dual state space, which in the present framework is discrete.

We first briefly describe the specific particle approximation on the dual space we are going to use. To approximate a predictive distribution $\nu_{i|0:i-1}(x_i)$, the classical particle approximation used in bootstrap particle filtering can be described as follows:

- sample $X_{i-1}^{(m)} \overset{\text{iid}}{\sim} \nu_{i-1|0:i-1}$, $m = 1, \dots, N$;
- propagate the particles by sampling $X_i^{(m)} \sim p_t(X_{i-1}^{(m)}, \cdot)$, with p_t the signal transition density;
- estimate $\nu_{i|0:i-1}$ with $\hat{\nu}_{i|0:i-1} := N^{-1} \sum_{m=1}^{N} \delta_{X_i^{(m)}}$.

For what concerns the use of dual processes, we are going to operate similarly to bootstrap particle filtering but on the dual space. The filtering distributions considered in this work are mixtures of the form $\nu_{i-1|0:i-1}(x_{i-1}) = \sum_{\mathbf{m}} w_{\mathbf{m}} h(x_{i-1}, \mathbf{m}) \pi(x_{i-1})$. An estimate of these can be obtained through a particle approximation of the discrete mixing measure, that is we draw $\mathbf{m}^{(n)} \overset{\text{iid}}{\sim} \sum_{\mathbf{m}} w_{\mathbf{m}} \delta_{\mathbf{m}}$, $n = 1, \dots, N$, to obtain $\hat{\nu}_{i-1|0:i-1}(x_{i-1}) := N^{-1} \sum_{n=1}^{N} h(x_{i-1}, \mathbf{m}^{(n)}) \pi(x_{i-1})$. The natural approximation of $\nu_{i|0:i-1}(x_i)$ is therefore as follows:

- sample $\mathbf{m}^{(n)} \overset{\text{iid}}{\sim} \sum_{\mathbf{m}} w_{\mathbf{m}} \delta_{\mathbf{m}}$;
- propagate the particles by sampling $\mathbf{n}^{(n)} \sim p_{\mathbf{m}^{(n)},\cdot}(t)$, with $p_{\mathbf{m}^{(n)},\cdot}(t)$ the transition probabilities of the dual process;
- estimate $\nu_{i|0:i-1}(x_i)$ with $\hat{\nu}_{i|0:i-1}(x_i) := N^{-1} \sum_{n=1}^{N} h(x_i, \mathbf{n}^{(n)}) \pi(x_i)$.

Here some important remarks are in order. The above dual particle approximation is a finite mixture approximation of a mixture which can be either finite or infinite. Hence the above strategy can be applied both to filtering given death-like duals but also given general duals on discrete state spaces. The quality of the dual particle approximation, in general, may differ from that obtained through the particle filtering approximation since the particles live on a discrete space in the first case and on a continuous space in the second. This is the object of the following sections, at least for two specific examples. Finally, the ease of implementation of the two approximations may be very different because simulating from the original Markov process may be much harder than simulating from the dual process. An example is the simulation of Kingman's typed coalescent, immediate as compared to the simulation from (22), which would be unfeasible without [45].

6.1 Cox–Ingersoll–Ross numerical experiments

The CIR diffusion admits two different duals:

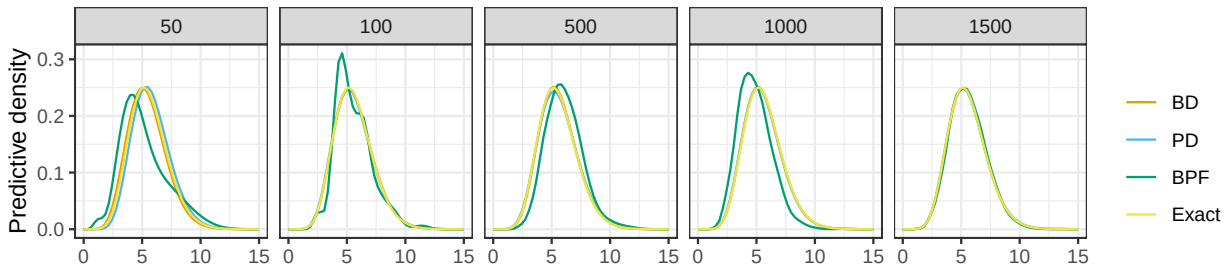

Figure 1: Comparison of the signal predictive distribution $\hat{p}(x_{k+1}|y_{1:k})$ obtained through the approximation approach to the death-process dual and the B&D dual, and through the bootstrap particle filter, with the exact predictive. The number of particles used for the approximations are 50, 100, 500, 1000, 1500 and are indicated in the panel labels. The acronyms are BD: Birth-and-Death, PD: Pure-Death, BPF: Bootstrap Particle Filter.

- the death-like dual given by $D_t = (M_t, \Theta_t)$, with M_t a pure-death process on $\mathbb{Z}_+$ with rates $2\sigma^2\theta m$ from m to $m-1$ and Θ_t a deterministic process that solves the ODE $\mathrm{d}\Theta_t/\mathrm{d}t = -2\sigma^2\Theta_t(\Theta_t - \gamma/\sigma^2)$, $\Theta_0 = \theta$. Cf. [53], Section 3.1.
- the B&D dual M_t on $\mathbb{Z}_+$ with birth rates from m to $m+1$ given by $\lambda_m = 2\sigma^2(\delta/2+m)(\theta-\gamma/\sigma^2)$ and death rates from m to $m-1$ given by $\mu_m = 2\sigma^2\theta m$ respectively. Cf. Proposition 4.1.

Note that the latter is time-homogeneous, the former is not. In general, temporal homogeneity is to be preferred since a direct simulation with a Gillespie algorithm in the inhomogeneous case would require a time-rescaling. However, for this specific case, there is a convenient closed-form expression for the transition density of the first dual, which can be used to simulate for arbitrary time transitions (see the third displayed equation at page 2011 in [53]). The second dual, by virtue of the temporal homogeneity, can be simulated directly using a Gillespie algorithm. This may be slow if the event rate becomes large, but as suggested in Section 4 we can see it as a linear B&D process, and a convenient closed-form expression can be used to simulate arbitrary time transitions.

We compare these two particle approximations with an exact computation of the predictive distribution following [53] and to a bootstrap particle filtering approach on the original state space of the signal, which is easy to implement for arbitrary time transitions thanks to the Gamma-Poisson expansion of the CIR transition density (see details in [50], Section 5).

Figure 1 shows the comparison of the above-illustrated strategies, with prediction performed for a forecast time horizon of 0.05. The CIR parameters were specified to $\delta = 11, \sigma = 1, \gamma = 1.1$. The starting distribution for the prediction is a filtering distribution for a dataset whose last Poisson observation equals 4, so the starting distribution is a mixture of Gamma densities roughly centred around this point. The density estimates for the bootstrap particle filter were obtained from a Gamma kernel density estimator with bandwidth estimated by cross-validation. This is

expected to induce a negligible error because the target distribution is a finite mixture of Gamma distributions.

The figure suggests that the bootstrap particle filter is slower to converge to the exact predictive distribution. Instead, with only 50 particles, both dual approximations that use a pure-and a B&D dual (respectively PD and BD in the Figure legend) are already almost indistinguishable from the exact predictive distribution. This shows that accurately approximating the mixing measure on the discrete dual space seems to require fewer particles than approximating the continuous distribution on the original continuous state space. Shorter and longer time horizons than that used in Figure 1 were also tested and provided qualitatively similar results.

Next, we turn to investigating the error on the filtering distributions, which combines successive particle approximations. Since the update operation can be performed exactly through (7), particle filtering using the dual process is conveniently implemented like a bootstrap particle approximation to a Baum-Welch filter with systematic resampling. We quantify the error on the filtering distributions by measuring the absolute error on the first moment and the standard deviation of the filtering distributions (with respect to the exact computation). We also include the error on the signal retrieval, measured as the absolute difference between the first moment of the filtering distributions and the value of the hidden signal to be retrieved. The mean filtering error is averaged over the second half of the sequence of observations to avoid possible transient effects at the beginning of the observation sequence and estimated over 50 different simulated datasets. The parameter specification is again $\delta = 11, \sigma = 1, \gamma = 1.1$, with a single Poisson observation at each of 200 observation times, and intervals between consecutive observations equal to 0.1. Figure 2 shows that the pure-death particle approximation performs better than the B&D particle approximation, but the latter performs comparably to the bootstrap particle filter, possibly with a modest advantage.

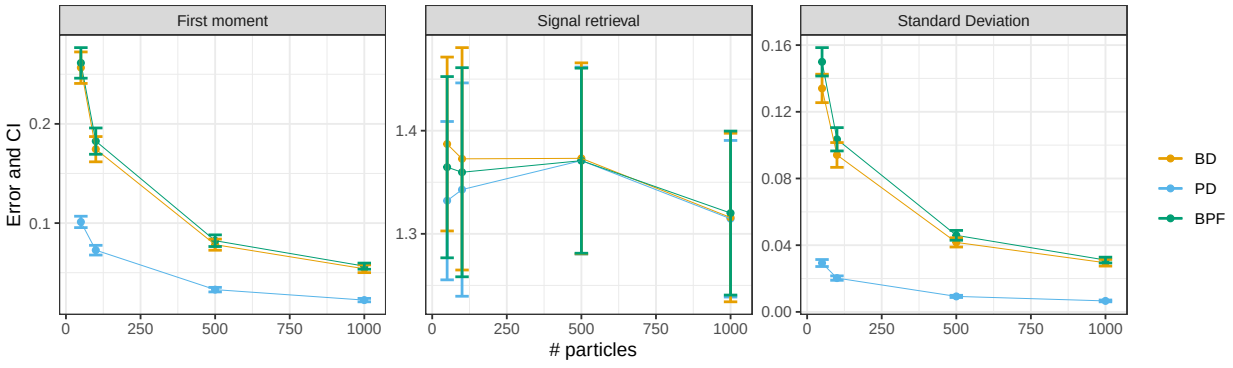

Figure 2: Mean filtering error as a function of the number of particles for the various particle approximation methods. The error bars represent the confidence interval on the error estimate from the 50 repetitions. The acronyms are BD: Birth-and-Death, PD: Pure-Death, BPF: Bootstrap Particle Filter.

6.2 Wright–Fisher numerical experiments

The WF diffusion admits two different duals:

- Kingman's typed coalescent with mutation dual, given by a pure-death process on $\mathbb{Z}_+^K$ with rates $\lambda_{\mathbf{m},\mathbf{m}-\mathbf{e}_i} = m_i(|\boldsymbol{\alpha}| + |\boldsymbol{m}| - 1)/2$ from $\mathbf{m}$ to $\mathbf{m} - \mathbf{e}_i$. Cf. [53], Section 3.3.
- a Moran dual process, given a homogeneous B&D process on $\mathbb{Z}_+^K$ with rates $\lambda_{\mathbf{m},\mathbf{m}-\mathbf{e}_i+\mathbf{e}_j} = m_i(\alpha_j + m_j)/2$ from $\mathbf{m}$ to $\mathbf{m} - \mathbf{e}_i + \mathbf{e}_j$. Cf. Proposition 5.1.

Here both processes are temporally homogeneous and can thus be easily simulated using a Gillespie algorithm, with the only caveat that the simulation can be inefficient when the infinitesimal rates are large. Similar to the CIR case, there is a closed-form expression for the transition probabilities in the first case, which can be used for simulation purposes for arbitrary time transitions (see Theorem 3.1 in [54]). Unlike the one-dimensional CIR case, handling this expression is challenging in the multi-dimensional WF case, with significant numerical stability issues raised by the need to compute the sum of alternated series with terms that can both overflow and underflow. In [50], these hurdles were addressed using arbitrary precision computation libraries and careful re-use of previous computations applicable when data is equally spaced. The Gillespie simulation strategy presents no such restriction and may be significantly faster when the event rates remain low.

As mentioned in Section 5, no closed-form expression is available for the Moran dual and the Gillespie algorithm approach is the main option, likely resulting in a slow algorithm. Alternatively, as argued in Section 5, we can approximate the Moran dual process by a finite population Wright–Fisher chain, with the quality of approximation increasing with the population size. The interest in this approximation is that the event rate for the latter is lower than for the Moran process. This is related to the fact that weak convergence of a sequence of WF chains to a WF diffusion occurs when time is rescaled by a factor of N (cf. [48], Sec. 15.2.F), whereas a Moran model whose individual updates occur at the times of a Poisson process with rate 1, needs a rescaling by a factor N^2 to obtain a similar convergence. In other words, in order to establish weak convergence to the diffusion, time must be measured in units of N generations in the WF chain and in units of N^2 generations in the Moran model. See discussion in Section 5. For this reason, the resulting Gillespie simulation is expected to be faster using a WF chain approximation to the Moran model.

The above considerations also suggest another possibility. Since the Moran process converges weakly to a Wright–Fisher diffusion, the latter could also be used as a possible approximation instead of a WF chain. In this case, it is possible to sample directly from (22) for arbitrary time transitions using the algorithm in [45]. Hence we would be using a WF diffusion to approximate the dual Moran transitions in (25).

A standard bootstrap particle filter performed directly on the Wright–Fisher diffusion state space also crucially relies on the algorithm of [45] for the prediction step, without which approximate sampling recipes from the transition density would be needed.

In Figure 3, we compare prediction strategies for a WF diffusion with $K = 4$ types using:

- the closed-form transition of the pure death dual ("Exact" in Fig. 3 legend);
- an approximation of the pure death dual using a Gillespie algorithm ("PD");
- an approximation of the Moran dual using a Gillespie algorithm ("BD Gillespie Moran");
- a WF chain approximation of the Moran dual using a Gillespie algorithm ("BD Gillespie WF");
- a WF diffusion approximation of the Moran dual using [45] ("BD diffusion WF");
- a bootstrap particle filtering approximation using [45] ("Bootstrap PF").

In Figure 3, prediction was performed for a forecast time horizon equal to 0.1, with WF parameters $\boldsymbol{\alpha} = (3, 3, 3, 3)$. The starting distribution for the prediction is a filtering distribution for a dataset whose last multinomial observation is equal to $(4, 0, 9, 2)$ (so the starting distribution is a mixture of Dirichlet distributions roughly centred around $(4/15, 0, 9/15, 2/15)$). Various values for parameter $\boldsymbol{\alpha}$ were also tested and provided results qualitatively similar to Figure 1. The density estimates for the bootstrap particle filter are obtained from a Dirichlet kernel density estimator with bandwidth estimated by cross-validation (using Julia package KernelEstimators `https://github.com/panlanfeng/KernelEstimator.jl`). This is expected to induce a negligible error because the target distribution is a finite mixture of Dirichlet distributions. Figure 3 shows that among these particle approximations of $p(x_{k+1}|y_{1:k})$, the Wright–Fisher diffusion approximation of the Moran dual seems to converge slowest, followed by the bootstrap particle filter, whereas the other strategies based on the dual process converge quickly to the exact distribution.

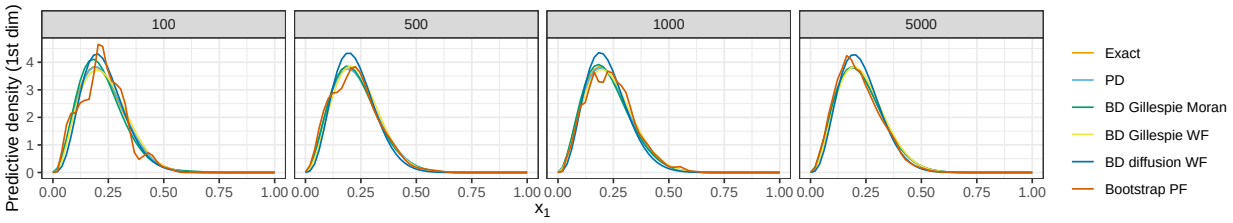

Figure 3: Convergence of the WF predictive distribution (only the first dimension) with the number of particles for the various particle approximations. The acronyms are PD: Pure-Death, BD: Birth-and-Death, WF: Wright-Fisher, PF: Particle Filter.

Figure 4 evaluates the filtering error for a WF process with $K = 3$ and parameters $\boldsymbol{\alpha} = (1.1, 1.1, 1.1)$, given 20 categorical observations collected at each time, over 10 collection times spaced by intervals equal to 1. We consider increasing numbers of particles and use 100 replications to estimate the error. The figure shows that the particle approximation of the pure death dual process using the closed-form transition exhibits better performance. The bootstrap particle approximation has the fastest improvement relative to increasing the number of particles. Overall, the Moran dual performs better or comparably to bootstrap particle filtering.

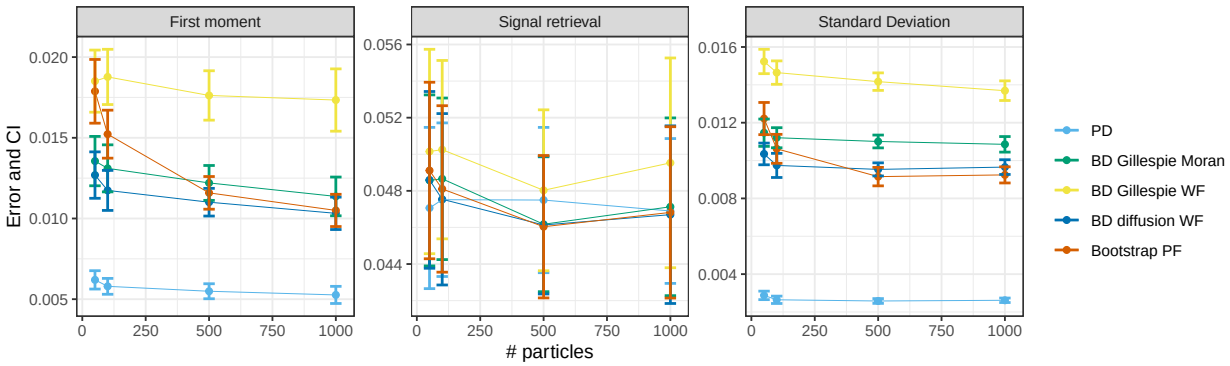

Figure 4: Mean filtering error as a function of the number of particles for the various particle approximation methods. The error bars represent the confidence interval on the error estimate from the 100 repetitions. The acronyms are PD: Pure-Death, BD: Birth-and-Death, WF: Wright-Fisher, PF: Particle Filter.

7 Concluding remarks

We have provided conditions for filtering diffusion processes on multidimensional continuous spaces which avoid computations on the state space of the forward process when a dual process given by a discrete Markov chain is available. Motivated by certain diffusion models for which only duals with a countable state space are known (e.g., B&D-like duals for WF diffusions with selection), we have investigated the performance of filtering based on a B&D dual for the CIR diffusion and based on a Moran process dual for the WF diffusion. All approximation methods proposed appear to be valuable strategies, despite resting on different simulation schemes. The optimal strategy is bound to depend on the application at hand, together with several other details like the interval lengths between data collection times, and possibly be constrained by which of these tools are available. For example, the transition function of coupled WF diffusions [4] is not available, whereas a discrete dual was found in [25]. Overall, approximate filtering using B&D-like duals may perform better or comparably to bootstrap particle filtering, with the advantage of operating on a discrete state space. The computational effort for each of these strategies is also bound to depend on a series of factors the identification of which is beyond the scope of this contribution.

The code to reproduce the analyses illustrated above will be made available in the Supporting Material and is based on the package freely available at `https://github.com/konkam/DualOptimalFiltering.jl`.

8 Acknowledgements

The authors are grateful to two anonymous referees for carefully reading the manuscript and for providing constructive suggestions that led to improving the paper. They also gratefully acknowl-

edge insightful conversations with Omiros Papaspiliopoulos on performing particle filtering on the dual space.
The last author is partially supported by MUR, PRIN project 2022CLTYP4.

References

[1] Arthreya, S. and Swart, J. Branching-coalescing particle systems. *Probab. Theory Relat. Fields* **131**, 376–414.
[2] Ascolani, F., Lijoi, A. and Ruggiero, M. (2021). Predictive inference with Fleming–Viot-driven dependent Dirichlet processes. *Bayesian Anal.* **16**, 371–395.
[3] Ascolani, F., Lijoi, A. and Ruggiero, M. (2023). Smoothing distributions for conditional Fleming–Viot and Dawson–Watanabe diffusions. *Bernoulli* **29**, 1410-1434.
[4] Aurell, E., Ekeberg, M. and Koski, T. (2019). On a multilocus Wright–Fisher model with mutation and a Svirezhev-Shahshahani gradient-like selection dynamics. *arXiv:1906.00716*.
[5] Bailey, N.T.J. (1964). *The elements of stochastic processes with applications to the natural sciences*. Wiley, New York
[6] Bain, A. and Crisan, D. (2009). *Fundamentals of stochastic filtering*. Springer.
[7] Barbour, A.D., Ethier, S.N. and Griffiths, R.C. (2000). A transition function expansion for a diffusion model with selection. *Ann. Appl. Probab.* **10**, 123–162.
[8] Birkner, M.C., Blath, J., Moehle, M., Steinruecken, M. and Tams, J. (2008). A modified lookdown construction for the Xi-Fleming–Viot process with mutation and populations with recurrent bottlenecks. *Alea* **6**, 25–61.
[9] Chen, R. and Scott, L. (1992). Pricing interest rate options in a two-factor Cox–Ingersoll–Ross model of the term structure. *Rev. Financial Stud.* **5**, 613–636.
[10] Cox, J.C., Ingersoll, J.E. and Ross, S.A. (1985). A theory of the term structure of interest rates. *Econometrica* **53**, 385–407.
[11] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in hidden Markov models.* Springer.
[12] Carinci, C., Giardinà, C., Giberti, C. and Redig, F. (2015). Dualities in population genetics: A fresh look with new dualities. *Stoch. Proc. Appl.* **125**, 941–969.
[13] Chaleyat-Maurel, M. and Genon-Catalot, V. (2006). Computable infinite-dimensional filters with applications to discretized diffusion processes. *Stoch. Proc. Appl.* **116**, 1447–1467.
[14] Chaleyat-Maurel, M. and Genon-Catalot, V. (2009). Filtering the Wright–Fisher diffusion. *ESAIM Probab. Stat.* **13**, 197–217.
[15] Chopin, N. and Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo.* Springer.
[16] Comte, F., Genon-Catalot, V. and Kessler, M. (2011). Multiplicative Kalman filtering. *Test* **20**, 389–411.
[17] Crawford, F.W. and Suchard, M.A. (2012). Transition probabilities for general birth-death processes with applications in ecology, genetics, and evolution. *J. Math. Biol.* **65**, 553–580.

[18] Depperschmidt, A., Greven, A. and Pfaffelhuber, P. (2019). Duality and the well-posedness of a martingale problem. *arXiv:1904.01564*.
[19] Ethier, S.N. and Griffiths, R.C. (1993a). The transition function of a Fleming–Viot process. *Ann. Probab.* **21**, 1571–1590.
[20] Ethier, S.N. and Griffiths, R.C. (1993b). The transition function of a measure-valued branching diffusion with immigration. In *Stochastic Processes. A Festschrift in Honour of Gopinath Kallianpur* (S. Cambanis, J. Ghosh, R. L. Karandikar and P. K. Sen, eds.) 71–79. Springer, New York.
[21] Ethier, S.N. and Kurtz, T.G. (1986). *Markov processes: characterization and convergence*. Wiley, New York.
[22] Etheridge, A.M. (2009). *Some mathematical models from population genetics*. École d'été de Probabilités de Saint-Flour XXXIX. Lecture Notes in Math. **2012**. Springer.
[23] Etheridge, A.M. and Griffiths, R.C. (2009). A coalescent dual process in a Moran model with genic selection. *Theor. Pop. Biol.* **75**, 320–330.
[24] Etheridge, A.M., Griffiths, R.C. and Taylor, J.E. (2010). A coalescent dual process in a Moran model with genic selection and the lambda coalescent limit. *Theor. Popn. Biol.* **78**, 77–92.
[25] Favero, M., Hult, H. and Koski, T. (2021). A dual process for the coupled Wright–Fisher diffusion. *J. Math. Biol.* **82:6**.
[26] Ferrante, M. and Runggaldier, W.J. (1990). On necessary conditions for the existence of finite-dimensional filters in discrete time. *Systems Control Lett.* **14**, 63–69.
[27] Ferrante, M. and Vidoni, P. (1998). Finite dimensional filters for nonlinear stochastic difference equations with multiplicative noises. *Stoch. Proc. Appl.* **77**, 69–81.
[28] Franceschini, C., Giardinà, C. and Groenevelt, W. (2018). Self-duality of Markov processes and intertwining functions. *Math. Phys. Anal. Geom.* **21**, 29.
[29] Frydman, H. (1994). Asymptotic inference for the parameters of a discrete-time square-root process. *Math. Finance* **4**, 169–181.
[30] Genon-Catalot, V. (2003). A non-linear explicit filter. *Statist. Probab. Lett.* **61**, 145–154.
[31] 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**, 701–720.
[32] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.
[33] Giardinà, C., Kurchan, J. Redig, F. (2007). Duality and exact correlations for a model of heat conduction. *J. Math. Phys.* **48**, 033301.
[34] Giardinà, C., Kurchan, J., Redig, F. and Vafayi, K. (2009a). Duality and hidden symmetries in interacting particle systems. *J. Stat. Phys.* **135**, 25–55.
[35] Giardinà, C., Redig, F. and Vafayi, K. (2009b). Correlation inequalities for interacting particle systems with duality, *J. Stat. Phys.* **141**, 242–263.
[36] Gillespie, D.T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58**, 35–55.
[37] Göing-Jaeschke, A. and Yor, M. (2003). A survey and some generalizations of Bessel processes. *Bernoulli*, **9**, 313–349.

[38] Griffiths, R.C., Ruggiero, M, Spanó, D. and Zhou, Y. (2022). Dual process in the two-parameter Poisson-Dirichlet Petrov diffusion. *arXiv:2102.08520*.

[39] Griffiths, R.C. and Spanó, D. (2009). Diffusion processes and coalescent trees. In *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*, ed. N. H. Bingham and C. M. Goldie. London Mathematical Society Lecture Notes Series, Cambridge University Press 2010.

[40] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility, with applications to bond and currency options. *Rev. Financial Stud.* **6**, 327–343.

[41] Huillet, T. (2007). On Wright?Fisher diffusion and its relatives. *J. Stat. Mech.*, P11006.

[42] Huillet, T. and Martinez, S. (2011). Duality and intertwining for discrete Markov kernels: relations and examples. *Adv. Appl. Probab.* **43**, 437–460.

[43] Hutzenthaler, M. and Wakolbinger, A. (2007). Ergodic behavior of locally regulated branching populations. *Ann. Appl. Probab.* **17**, 474–501.

[44] Jansen, S. and Kurt, N. (2014). On the notion(s) of duality for Markov processes. *Probab. Surv.* **11**, 59–120.

[45] Jenkins, P. A. and Spanò, D. (2017). Exact simulation of the Wright–Fisher diffusion. *Ann. Appl. Probab.* **27(3)**, 1478–1509.

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

[47] Kalman, R.E. and Bucy, R.S. (1961). New results in linear filtering and prediction theory. *J. Basic Eng.* **83** 95–108.

[48] Karlin, S. and Taylor, H.M. (1981). *A second course in stochastic processes*. Academic Press, New York.

[49] Kawazu, K. and Watanabe, S. (1971). Branching processes with immigration and related limit theorems. *Theory Prob. Appl.* **16**, 36–54.

[50] Kon Kam King, G., Papaspiliopoulos, O. and Ruggiero, M. (2021). Exact inference for a class of hidden Markov models on general state spaces. *Electron. J. Stat.* **15**, 2832–2875.

[51] Möhle, M. (1999). The concept of duality and applications to Markov processes arising in neutral population genetics models. *Bernoulli* **5**, 761–777.

[52] Ohkubo, J. (2010). Duality in interacting particle systems and boson representation. *J. Stat. Phys.* **139**, 454–465.

[53] Papaspiliopoulos, O. and Ruggiero, M. (2014). Optimal filtering and the dual process. *Bernoulli* **20**, 1999–2019.

[54] Papaspiliopoulos, O., Ruggiero, M. and Spanò, D. (2016). Conjugacy properties of time-evolving Dirichlet and gamma random measures. *Electron. J. Stat.* **10**, 3452–3489.

[55] Runggaldier, W.J. and Spizzichino, F. (2001). Sufficient conditions for finite dimensionality of filters in discrete time: a Laplace transform-based approach. *Bernoulli* **7**, 211-221.

[56] Sawitzki, G. (1981). Finite-dimensional filter systems in discrete time. *Stochastics* **5**, 107–114.

[57] Tavaré, S. (2018). The linear birth-and-death process: an inferential retrospective. *Adv. Appl. Probab.* **50**, 253–269.