ISSN 2279-9362

Collegio Carlo Alberto

Optimal filtering and the dual process

Omiros Papaspiliopoulos
Matteo Ruggiero

No. 299
May 2013

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Optimal filtering and the dual process

Omiros Papaspiliopoulos
ICREA and Universitat Pompeu Fabra
Matteo Ruggiero*
University of Torino and Collegio Carlo Alberto

May 20, 2013

We link optimal filtering for hidden Markov models to the notion of duality for Markov processes. We show that when the signal is dual to a process that has two components, one deterministic and one a pure death process, and with respect to functions that define changes of measure conjugate to the emission density, the filtering distributions evolve in the family of finite mixtures of such measures and the filter can be computed at a cost that is polynomial in the number of observations. Hence, for models in this framework, optimal filtering reduces to a version of the Baum-Welch filter. Special cases of our framework are the Kalman filter, but also models where the signal is the Cox-Ingersoll-Ross process and the one-dimensional Wright-Fisher process, which have been investigated before in the literature. The duals of these two processes that we identify in this paper appear to be new in the literature. We also discuss the extensions of these results to an infinite-dimensional signal modelled as a Fleming-Viot process, and the connection of the duality framework we develop here and Kingman's coalescent.

Keywords: auxiliary variables, Bayesian conjugacy, Dirichlet process, finite mixture models, Cox-Ingersoll-Ross process, hidden Markov model, Kalman filters.

1 Introduction

A hidden Markov model (HMM) for a sequence of observations $\{Y_n, n \geq 0\}$, where $Y_n \in \mathcal{Y}$, is a discrete-time stochastic process with dynamics depicted in Figure 1. It is defined in terms of a hidden Markov chain, the so-called *signal*, which in this paper will be taken to be the discrete-time sampling of a time-homogeneous continuous-time Markov process X_t, with state-space $\mathcal{X}$, transition kernel $P_t(x, dx')$, and initial distribution $\nu(\mathrm{d}x)$. The observations relate to the signal by means of conditional distributions, assumed to be given by the kernel $F(x, \mathrm{d}y)$. We will assume that

$$F(x, \mathrm{d}y) = f_x(y)\mu(\mathrm{d}y), \tag{1}$$

*Email: matteo.ruggiero@unito.it

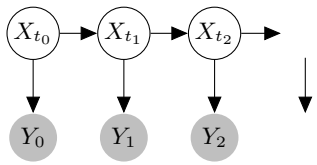

Figure 1: Hidden Markov model represented as a graphical model.

for some measure $\mu(\mathrm{d}y)$, in which case the corresponding densities are known as the *observation* or *emission* densities; Section 4 discusses results for models where such densities do not exist. The optimal filtering problem is the derivation of the conditional distributions $\mathcal{L}(X_{t_n}|Y_0,\ldots,Y_n)$ of the unobserved signal given the observations collected up to time t_n, henceforth denoted $\nu_n(\mathrm{d}x)$. These filtering distributions are the backbone of all statistical estimation problems in this framework, such as the prediction of future observations, the derivation of smoothing distributions (i.e., the conditional distribution of X_{t_n} given past and future observations) and the calculation of the likelihood function, i.e., the marginal density of the observations when the emission distributions are dominated. See Cappé, Moulines and Rydén (2005) for details and applications.

Throughout the paper we will assume that the signal is stationary and reversible with respect to a probability measure π. It is also appealing, from a modelling point of view, to assume that the signal evolves in continuous time, since there is a rich family of such models with a pre-specified stationary measure π. In addition, this assumption will give us a powerful tool to study optimal filtering by using the generator of the process, as we show in Section 2. In the examples of Section 3, the state space $\mathcal{X}$ of the signal will either be a subset of $\mathbb{R}$ or the K-dimensional simplex Δ_K; Section 4 discusses extensions to infinite-dimensional signals, where $\mathcal{X}$ is the space of probability measures on some compact Polish space $\mathcal{Y}$.

Mathematically, optimal filtering is the solution of the recursion

$$\nu_0 = \phi_{Y_0}(\nu)\,, \quad \nu_n = \phi_{Y_n}(\psi_{t_n - t_{n-1}}(\nu_n)), \quad n > 0,$$

which involves the following two operators acting on probability measures:

$$\begin{aligned} &\text{update:} && \phi_y(\nu)(\mathrm{d}x) = \frac{f_x(y)\nu(\mathrm{d}x)}{p_\nu(y)}, && p_\nu(y) = \int_{\mathcal{X}} f_x(y)\nu(\mathrm{d}x)\,. \\ &\text{prediction:} && \psi_t(\nu)(\mathrm{d}x') = \nu P_t(\mathrm{d}x') = \int_{\mathcal{X}} \nu(\mathrm{d}x) P_t(x,\mathrm{d}x') \end{aligned} \tag{2}$$

The "update" is the application of Bayes theorem, and the "prediction" gives the distribution of

the next step of the Markov chain initiated from ν. These operators have the following property when applied to *finite mixtures of distributions*:

$$(3) \qquad \phi_y\Big(\sum_{i=1}^n w_i\nu_i\Big)(\mathrm{d}x) = \sum_{i=1}^n \frac{w_i p_{\nu_i}(y)}{\sum_j w_j p_{\nu_j}(y)}\phi_y(\nu_i)\,, \quad \psi_t\Big(\sum_{i=1}^n w_i\nu_i\Big)(\mathrm{d}x) = \sum_{i=1}^n w_i\psi_t(\nu_i).$$

This implies that when $\mathcal{X}$ is a finite set, there is a simple algorithm for the sequential computation of the filtering probabilities. To see this, note that we can think of a distribution ν on a finite set $\mathcal{X}$, specified in terms of probabilities $\alpha_x, x \in \mathcal{X}$, as a finite mixture of point masses, $\nu = \sum_x \alpha_x\delta_x$; it is easy to compute $\phi_y(\delta_x), \psi_t(\delta_x)$ and then use the above result to obtain the probabilities associated with the distributions $\phi_y(\nu)$ and $\psi_t(\nu)$. This yields a popular algorithm for inference in HMMs, commonly known as the Baum-Welch filter, whose complexity is easily seen to be $\mathcal{O}(n|\mathcal{X}|^2)$, where $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$.

Outside the finite state-space case, the iteration of these two operators typically leads to analytically intractable distributions. However, there are notable exceptions to this rule. The classic example is the linear Gaussian state-space model, for which the filtering distributions are Gaussian with mean and covariance that can be iteratively computed using the so-called Kalman filter, at cost that grows linearly with n. Recent work by Genon-Catalot and collaborators uncovered that there exist interesting non-Gaussian models for which the filtering distributions are *finite mixtures* of parametric distributions. See Genon-Catalot and Kessler (2004); Chaleyat-Maurel and Genon-Catalot (2006; 2009), where the authors show how to compute the corresponding parameters sequentially in these models. We revisit their findings in Section 3. However, the number of mixture components increases with n in a way such that the cost of computing the filters grows polynomially with n (see Section 2 for details). Borrowing and adapting the terminology from Chaleyat-Maurel and Genon-Catalot (2006), we will refer to filters with such computational cost as *computable*, whereas filters whose cost grows linearly with n as *finite-dimensional*.

The work by Genon-Catalot and collaborators raises four important questions, which we address in this paper: are there more models which admit computable filters; do they share some basic structure; is there a general methodology to identify such models and to obtain the algorithm which computes the sequence of parameters; what is the computational complexity of such schemes and how can we obtain faster approximate filtering algorithms? We show that the answer to all these questions relates to an important probabilistic object: the *dual process*. Duality methods have a long history in Probability, dating back to the work of Lévy (1948) (see Jansen and Kurt (2013) for a recent review). These have been widely applied to the study of interacting particle systems Liggett (2005) and proven to be a powerful method which provides alternative, and often simpler, tools for investigating the sample path properties of the process

at hand. For example, the existence of a dual for a certain Markov process (and for a sufficiently large class of functions) implies that the associated martingale problem is well defined, hence that the process is unique; see Section 4.4 of Ethier and Kurtz (1986). See also Dawson (1993) and Etheridge (2000) for applications of duality to population genetics.

In this paper we illustrate that dual processes play a central role in optimal filtering and to a great extent can be used to settle the four questions posed above. We also uncover their potential as auxiliary variables in Monte Carlo schemes for stochastic processes (and, hence, as a variance reduction scheme). In our framework, the dual will in general be given by two components: a *deterministic process*, driven by an ordinary differential equation, and a (multidimensional) *death process* with countable state-space. We show how to derive an explicit, recursive filtering scheme once the dual is identified, and apply this methodology to three cases of fundamental interest. In doing so, we identify some duals that, to the best of our knowledge, have not appeared before in the literature, and we solve the filtering problem for some new classes of HMMs.

The rest of the paper is organized as follows. In Section 2 we link optimal filtering to a specific type of duality, we show how to identify the dual in terms of the *generator* of X_t, and study the complexity of the resulting filtering algorithm. Section 3 analyzes three interesting models for which the dual process is derived: the Cox-Ingersoll-Ross model, the Ornstein-Uhlenbeck process and the K-dimensional Wright-Fisher diffusion. These models are reversible with respect to the gamma, Gaussian and Dirichlet distribution respectively, and for the Gaussian case the computable filter reduces to the Kalman filter. Section 4 deals with the extension of the results to the nonparametric filtering. In particular, we consider the case where the signal is infinite-dimensional and evolves according to a Fleming-Viot process, which is a measure-valued diffusion reversible with respect to the law of a Dirichlet process (Ferguson, 1973), and the observations are random samples from the underlying measures. The solution to this problem is based on our results on Wright-Fisher processes, their connection with Fleming-Viot processes and the characterization of Dirichlet process mixtures by Antoniak (1974). Finally, Section 5 discusses several connections, including how our work interacts with previous work on dual processes.

2 Methodology: filtering the dual process

2.1 Linking optimal filtering to duality

Before presenting the main results, we introduce three fundamental assumptions which provide the general framework under which the results are derived. First, we will assume that X is reversible with respect to a probability measure π:

A1 (Reversibility): $\pi(\mathrm{d}x)P_t(x,\mathrm{d}x') = \pi(\mathrm{d}x')P_t(x',\mathrm{d}x)$.

In order to state the second assumption, we need to introduce a certain amount of notation. Define, for $K \in \mathbb{Z}_+ = \mathbb{N} \cup \{0\}$, the space of multi-indices

$$\mathcal{M} = \mathbb{Z}_+^K = \{\, \mathbf{m} = (m_1, \ldots, m_K) : m_j \in \mathbb{Z}_+,\ j = 1, \ldots, K \,\}. \tag{4}$$

We will use the symbol $\mathbf{0}$ to denote the vector of zeros, $\mathbf{e}_j$ for the vector in $\mathcal{M}$ whose only non-zero element is found at the jth coordinate and equals 1, and let $|\mathbf{m}| = \sum_i m_i$. Furthermore, we will use the product order on $\mathcal{M}$, according to which for $\mathbf{m}, \mathbf{n} \in \mathcal{M}$, $\mathbf{m} \leq \mathbf{n}$ if an only if $m_j \leq n_j$ for all j. Then, for $\mathbf{i} \leq \mathbf{m}$, $\mathbf{m} - \mathbf{i}$ is the vector with jth element $m_j - i_j$. Additionally, if $\Lambda \subset \mathcal{M}$, define

$$G(\Lambda) = \{\mathbf{n} : \mathbf{n} \leq \mathbf{m}\,, \mathbf{m} \in \Lambda\}\,. \tag{5}$$

The notation for $\mathcal{M}$ does not reflect its dependence on the dimension K, but we will reserve boldface for elements of $\mathcal{M}$ when $K > 1$ (or unspecified), whereas normal typeface will be used for elements of $\mathbb{Z}_+$. Finally, the following notations will be used to denote conditional expectations

$$(P_t f)(x) = \mathbb{E}^x[f(X_t)] = \mathbb{E}[f(X_t) \mid X_0 = x] = \int_{\mathcal{X}} f(x')P_t(x,\mathrm{d}x')\,.$$

The first denotes the action on f of the semigroup operator associated to the transition kernel, where with some abuse of notation the same symbol is used both for the semigroup and the kernel.
The second assumption is concerned with models where $\pi(\mathrm{d}x)$ is *conjugate* to the emission density $f_x(y)$:

A2 (Conjugacy): For $\Theta \subseteq \mathbb{R}^l$, $l \in \mathbb{Z}_+$, let $h : \mathcal{X} \times \mathcal{M} \times \Theta \to \mathbb{R}_+$ be such that $\sup_x h(x, \mathbf{m}, \theta) < \infty$ for all $\mathbf{m} \in \mathcal{M}, \theta \in \Theta$, and $h(x, \mathbf{0}, \tilde{\theta}) = 1$ for some $\tilde{\theta} \in \Theta$. Then $\mathcal{F} = \{h(x, \mathbf{m}, \theta)\pi(\mathrm{d}x),\ \mathbf{m} \in \mathcal{M}, \theta \in \Theta\}$ is assumed to be a family of probability measures such that there exist functions $t : \mathcal{Y} \times \mathcal{M} \to \mathcal{M}$ and $T : \mathcal{Y} \times \Theta \to \Theta$ with $\mathbf{m} \to t(y, \mathbf{m})$ increasing and such that

$$\phi_y(h(x, \mathbf{m}, \theta)\pi(\mathrm{d}x)) = h(x, t(y, \mathbf{m}), T(y, \theta))\pi(\mathrm{d}x)\,.$$

Hence here with conjugacy we intend the fact that the family $\mathcal{F}$ of measures, which includes π, is closed under the update operation. The assumption that h is bounded in x will be discussed after the statement of Assumption A4.

For $p_\nu(y)$ as in (2), it is easy to check that in the context of A2, we have

$$p_{h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)}(y) =: c(\mathbf{m},\theta,y) = \frac{f_x(y)h(x,\mathbf{m},\theta)}{h(x,t(y,m),T(y,\theta))}\,, \tag{6}$$

which, despite its appearance, does not depend on x.
Note that our definitions of $\mathcal{M}$ and Θ allow the possibility that $K = 0$ or $l = 0$, in which case h in A2 is function only of the variables with non-zero dimension, whereas the case $K = l = 0$ is not of interest here. In the setting of Assumption A2 and for the trivial Markov dynamics $X_t \equiv X_0$, with $X_0 \sim \pi$, the filtering problem collapses to conjugate Bayesian inference for the unknown parameter x of the sampling density $f_x(y)$. See Section 5.2 and Appendix A.2 of Bernardo and Smith (1994) for an exposition of conjugate Bayesian inference and stylized conjugate Bayesian models, and Section 3 in this paper for examples within our framework.
The third main assumption for our results concerns the existence of a certain type of dual process for the signal.

A3 (Duality): We assume that $r : \Theta \to \Theta$ is such that the differential equation

$$\mathrm{d}\Theta_t/\mathrm{d}t = r(\Theta_t), \quad \Theta_0 = \theta_0\,, \tag{7}$$

has a unique solution for all θ_0. Let $\lambda : \mathbb{Z}_+ \to \mathbb{R}_+$ be an increasing function, $\rho : \Theta \to \mathbb{R}_+$ be a continuous function, and consider a two-component Markov process (M_t, Θ_t) with state-space $\mathcal{M} \times \Theta$, where Θ_t evolves autonomously according to (7), and when at $(M_t, \Theta_t) = (\mathbf{m}, \theta)$, the process jumps down to state $(\mathbf{m} - \mathbf{e}_j, \theta)$ with instantaneous rate

$$\lambda(|\mathbf{m}|)\rho(\theta)m_j\,. \tag{8}$$

We assume (M_t, Θ_t) is *dual* to X_t with respect to the family of functions h defined in A2, in the sense that

$$\mathbb{E}^x[h(X_t,\mathbf{m},\theta)] = \mathbb{E}^{(\mathbf{m},\theta)}[h(x,M_t,\Theta_t)], \quad \forall x \in \mathcal{X}, \mathbf{m} \in \mathcal{M}, \theta \in \Theta, t \geq 0\,. \tag{9}$$

When $K = 0$ or $l = 0$ in A2, the dual process is just Θ_t or M_t respectively, and we adopt the convention that

$$\rho(\theta) \equiv 1 \quad \text{whenever} \quad l = 0.$$

As mentioned in Section 1, the notion of duality for Markov processes with respect to a given function is well known. See for example Section II.4 in Liggett (2005). Among the most common type of duality relations we mention *moment duality*, that is duality with respect to functions of type $h(x, y) = x^y$, and *Laplace duality*, with respect to functions of type $h(x, y) = e^{-axy}$.

See for example Jansen and Kurt (2013). In our framework, the duality functions are Radon-Nikodym derivatives between measures that are conjugate to the emission density, and this setup is perfectly tailored to optimal filtering. See Section 4.4 for more comments on this aspect. Furthermore, A3 specifies that we are interested in dual processes which can be decomposed into two parts: one purely *deterministic* and the other given by a K-dimensional *pure death process*, whose death rates are subordinated by the deterministic process. Note that M_t can only jump to "smaller" states according to the partial order on $\mathcal{M}$, and $\mathbf{0}$ is an absorbing state. The transition probabilities of the death process, conditional on the initial state $\Theta_0 = \theta$, will be denoted by

$$p_{\mathbf{m},\mathbf{n}}(t;\theta) = \mathbb{P}[M_t = \mathbf{n} | M_0 = \mathbf{m}, \Theta_0 = \theta], \quad \mathbf{n}, \mathbf{m} \in \mathcal{M}, \ \mathbf{n} \leq \mathbf{m}. \tag{10}$$

It is worth mentioning that the requirements on the structure of the dual processes prescribed by Assumption A3, with particular reference to the intensity (8), are justified by the three main reasons. The first is that, as shown in Section 3, they define a framework general enough to identify duals of processes of interest, the incorporation of a deterministic component being necessary in this respect. The second reason is that the transition probabilities (10) are analytically available, as provided by the following result, whose proof can be found in the Appendix.

Proposition 2.1. *Let (M_t, Θ_t) be as in A3, with $(M_0, \Theta_0) = (\mathbf{m}, \theta) \in \mathcal{M} \times \Theta$, and let $\lambda_{|\mathbf{m}|} = |\mathbf{m}|\lambda(|\mathbf{m}|)$. Then the transition probabilities for M_t are $p_{\mathbf{m},\mathbf{m}}(t;\theta) = \exp\{-\lambda_{|\mathbf{m}|}\int_0^t \rho(\Theta_s)\mathrm{d}s\}$ and, for any $\mathbf{0} \leq \mathbf{i} \leq \mathbf{m}$,*

$$p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t;\theta) = \left(\prod_{h=0}^{|\mathbf{i}|-1} \lambda_{|\mathbf{m}|-h}\right) C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t) \frac{\binom{|\mathbf{i}|}{i_1,\ldots,i_K}\binom{|\mathbf{m}|-|\mathbf{i}|}{m_1-i_1,\ldots,m_K-i_K}}{\binom{|\mathbf{m}|}{m_1,\ldots,m_K}},$$

where

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

This result can be interpreted as follows. The probability that a one-dimensional death process with inhomogeneous rates $\lambda_{|\mathbf{m}|}\rho(\Theta_s)$ decreases from $|\mathbf{m}|$ to $|\mathbf{m}| - |\mathbf{i}|$ in the interval $[0, t]$ is $(\prod_{h=0}^{|\mathbf{i}|-1} \lambda_{|\mathbf{m}|-h}) C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t)$, where the second factor is related to the convolution of the waiting times in an inhomogeneous Poisson process (cf. Section 19.10 in Johnson, Kotz and Balakrishnan (1994), and Sen and Balakrishnan (1999)). For a K-dimensional death process, the same quantity is the probability associated to all paths leading from level $|\mathbf{m}|$ to level $|\mathbf{m}| - |\mathbf{i}|$. The

hypergeometric-type combinatorial factor is then the probability of the subset of those leading from $\mathbf{m}$ to $\mathbf{m}-\mathbf{i}$.

Note that the special case of Proposition 2.1 yielded by $K=1$ and $\rho(\Theta_s)\equiv 1$ relates to the result obtained in Proposition 4.5. in Chaleyat-Maurel and Genon-Catalot (2009). Note also that when $\rho(\Theta_s)\equiv 1$ and $\lambda_m = m(\theta+m-1)/2$, $C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t)$ is the transition probability of the block-counting process of Kingman's coalescent with mutation, provided in Tavaré (1984) (cf. also Griffiths (2006)). See Section 4.4 for further comments on the links with Kingman's coalescent.

The third motivation behind the type of duality required by A3 is that if it holds, the prediction operator maps measures as in A2 into finite mixtures.

Proposition 2.2. *Let ψ_t be as in (2) and assume A1-A2-A3 hold. Then*

$$\psi_t(h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)) = \sum_{\mathbf{0}\le\mathbf{i}\le\mathbf{m}} p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t;\theta)h(x,\mathbf{m}-\mathbf{i},\Theta_t)\pi(\mathrm{d}x), \tag{11}$$

with $p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t;\theta)$ as in Proposition 2.1 and where Θ_t is the value in t of the process in (7) started from $\Theta_0=\theta$.

Proof. From (2) we have

$$\begin{aligned}\psi_t(h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)) &= \int_{\mathcal{X}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)P_t(x,\mathrm{d}x') = \int_{\mathcal{X}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x')P_t(x',\mathrm{d}x)\\ &= \pi(\mathrm{d}x')\mathbb{E}^{x'}[h(X_t,\mathbf{m},\theta)] = \pi(\mathrm{d}x')\mathbb{E}^{(\mathbf{m},\theta)}[h(x',M_t,\Theta_t)]\\ &= \sum_{\mathbf{n}\le\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t;\theta)h(x',\mathbf{n},\Theta_t)\pi(\mathrm{d}x')\end{aligned}$$

where we have used A1 in the second equality, A3 in the second fourth, the last follows from the definition in (10). □

The above result states that reversibility and the existence of the required duality jointly guarantee that the prediction operator can be computed with a finite effort. The reduction of the operator to a sum is due to the fact that X_t is dual to a Markov process with discrete state-space, but it is precisely the fact that M_t is a pure death process that makes the number of terms in the sum being finite. The next result shows that computable filtering is available in the framework we have outlined.

Proposition 2.3. *Consider the family of finite mixtures*

$$\bar{\mathcal{F}}_f = \left\{ \sum\nolimits_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x) \,:\, \Lambda\subset\mathcal{M}\,,\ |\Lambda|<\infty\,,\ w_{\mathbf{m}}\ge 0\,,\ \sum\nolimits_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} = 1 \right\}. \tag{12}$$

Then, under Assumptions A1-A2-A3, $\bar{\mathcal{F}}_f$ is closed under the application of the prediction and update operators (2), and specifically

$$\phi_y\Big(\sum\nolimits_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)\Big) = \sum\nolimits_{\mathbf{n}\in t(y,\Lambda)} \widehat{w}_{\mathbf{n}} h(x,\mathbf{n},T(y,\theta))\pi(\mathrm{d}x)$$

with

$$\begin{aligned} t(y,\Lambda) &:= \{\mathbf{n} : \mathbf{n} = t(y,\mathbf{m}), \mathbf{m}\in\Lambda\} \\ \widehat{w}_{\mathbf{n}} &\propto w_{\mathbf{m}} c(\mathbf{m},\theta,y) \quad \textit{for} \quad \mathbf{n} = t(y,\mathbf{m})\,, \sum\nolimits_{\mathbf{n}\in t(y,\Lambda)} \widehat{w}_{\mathbf{n}} = 1\,, \end{aligned} \tag{13}$$

and

$$\psi_t\Big(\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)\Big) = \sum_{\mathbf{n}\in G(\Lambda)} \Big(\sum_{\mathbf{m}\in\Lambda,\mathbf{m}\geq\mathbf{n}} w_{\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t;\theta)\Big) h(x,\mathbf{n},\theta_t)\pi(\mathrm{d}x). \tag{14}$$

The above Proposition shows that under Assumption A1 to A3, and provided the starting state belongs to the family $\bar{\mathcal{F}}_f$ of finite mixtures with components as in A2, then the filtering distributions evolve within $\bar{\mathcal{F}}_f$. Furthermore, the explicit reweighing of the mixture components is provided, thus allowing to concretely implement the recursive filtering scheme. Note also that this result generalizes Theorem 2.1 in Chaleyat-Maurel and Genon-Catalot (2006), which states a similar result for $K = 1$ under the Assumption A2 and the result in Proposition 2.2. The proof of Proposition 2.3 follows from (3), A2 and Proposition 2.2 by direct computation, and is thus omitted. Later in this section we will derive filtering algorithms based on this result. However, we first address in the next subsection the most important aspect of the approach described in this section, which is how to find a dual process that satisfies A2.

2.2 Local duality as a sufficient condition

It is typically easier to identify a process that satisfies the duality relation (9) for infinitesimal t. Formally, this requires studying the *generator* of X_t, which we will denote by $\mathcal{A}$. This is a linear operator, with domain denoted $\mathscr{D}(\mathcal{A})$, linked to the semigroup operator via the Kolmogorov backward equation

$$\frac{\partial}{\partial t} P_t f(x) = (\mathcal{A} P_t f)(x), \qquad f \in \mathscr{D}(\mathcal{A}),$$

where on the left hand side $P_t h(x)$ is differentiated in t for given x, whereas on the right hand side, $\mathcal{A}$ acts on $P_t h(x)$ as a function of x for given t. See for example Proposition 1.1.5 in Ethier and Kurtz (1986).

Suppose now X_t is a diffusion processes which solves an SDE on $\mathbb{R}^d$ of the form

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t\,.$$

In this case, $\mathcal{A}$ is the second-order differential operator given by

$$(\mathcal{A}f)(x) = \sum_{i=1}^{d} b_i(x)\frac{\partial f(x)}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{d} a_{i,j}(x)\frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \qquad f \in \mathscr{D}(\mathcal{A}), \tag{15}$$

for an appropriate domain $\mathscr{D}(\mathcal{A})$ and where $a_{i,j}(x) := (\sigma(x)\sigma(x)^T)_{i,j}$.
Let now A denote the generator of the dual process defined in A3, which can be easily checked to be

$$(Ag)(\mathbf{m},\theta) = \lambda(|\mathbf{m}|)\rho(\theta)\sum_{i=1}^{K} m_i g(\mathbf{m}-\mathbf{e}_i,\theta) + \sum_{i=1}^{l} r_i(\theta)\frac{\partial g(\mathbf{m},\theta)}{\partial \theta}\,, \quad g \in \mathscr{D}(A), \tag{16}$$

with r as in (7). The main idea is then to identify the dual process from the generator, instead of the semigroup operator.

A4 (Local duality): The function $h(x,\mathbf{m},\theta)$ defined in A2 is such that $h(x,\mathbf{m},\theta)$, as a function of x belongs to $\mathscr{D}(\mathcal{A})$ for all $(\mathbf{m},\theta) \in \mathcal{M}\times\Theta$, as a function of $(\mathbf{m},\theta)$ belongs to $\mathscr{D}(A)$ for all $x \in \mathcal{X}$, and

$$(\mathcal{A}h(\cdot,\mathbf{m},\theta))(x) = (Ah(x,\cdot,\cdot))(\mathbf{m},\theta)\,, \quad \forall x \in \mathcal{X}, \mathbf{m} \in \mathcal{M}, \theta \in \Theta\,. \tag{17}$$

It will typically be the case that in order to have $h(\cdot,\mathbf{m},\theta) \in \mathscr{D}(\mathcal{A})$, one needs h to be bounded in x, requirement included in Assumption A2. Proposition 1.2 of Jansen and Kurt (2013) shows that A4, together with the further assumption

$$\mathbb{E}^x[h(X_t,\mathbf{m},\theta)] \in \mathscr{D}(A), \qquad \mathbb{E}^{(\mathbf{m},\theta)}[h(x,M_t,\Theta_t)] \in \mathscr{D}(\mathcal{A}),$$

implies (9), whose argument can be sketched as follows. From (17) we can write

$$(\beta I - \mathcal{A})h = (\beta I - A)h, \qquad \beta \in \mathbb{R},$$

where I denotes the identity operator. Since $\mathcal{A}$ and A generate strongly continuous contraction semigroups, say on L_1 and L_2, their ranges are dense in L_1 and L_2 respectively. Moreover, the resolvents $\mathcal{R}_\beta = (\beta I - \mathcal{A})^{-1}, R_\beta = (\beta I - A)^{-1}$ are one-to-one for all $\beta > 0$, so the previous implies

$$\mathcal{R}_\beta h = R_\beta h, \qquad h \in L_1 \cap L_2, \quad \beta > 0.$$

Since the resolvent of an operator is the Laplace transform of the associated semigroup, and because of the uniqueness of Laplace transforms, the previous expression in turn implies (9). The approach sketched above for identifying the dual process by means of the local condition (17) will be implemented in Section 3, where we will identify the duals for some interesting relevant models.

2.3 The filtering algorithm

Typically, the initial distribution of the signal process belongs to $\mathcal{F}$, and most often equals the invariant measure π. Thus, without loss of generality and in order to simplify the exposition below, we make the following additional assumption.

A5 (Initialisation): The initial distribution of the signal is $\nu = h(x, \mathbf{m}_0, \theta_0)\pi(\mathrm{d}x) \in \mathcal{F}$, for some $\mathbf{m}_0 \in \mathcal{M}, \theta_0 \in \Theta$.

Proposition 2.3 provides a probabilistic interpretation of the weights involved in the finite mixtures in terms of the transition probabilities of the dual death process M_t. This interpretation can be elaborated further, in order to facilitate the development of filtering algorithms. With a little abuse of notation, denote by $\{D_n = (M_n, \Theta_n), n \geq 0\}$ a discrete-time process with state-space $\mathcal{M} \times \Theta$ constructed as follows. Consider a partially observed Markov process, where the signal is now D_n and the conditional independence structure, given in Figure 2 graphically, is as follows. Let $D_0 = (M_0, \Theta_0) = (\mathbf{m}_0, \theta_0)$ be the initial state of the chain, with $(\mathbf{m}_0, \theta_0)$ defined

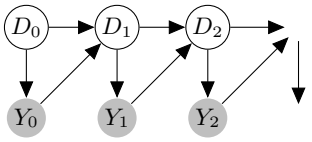

Figure 2: The partially observed Markov process dual to the hidden Markov model in Figure 1, where $D_i = (M_i, \Theta_i)$.

in A5. Then $\mathcal{L}(Y_n | D_n = (\mathbf{m}, \theta)) = c(\mathbf{m}, \theta, y)\mu(\mathrm{d}y)$, with $c(\mathbf{m}, \theta, y)$ as in (6) and μ in (1), and for $n \geq 1$, $\mathcal{L}(D_n | Y_{n-1} = y, D_{n-1} = (\mathbf{m}, \theta))$ is the law of $(M_{t_n - t_{n-1}}, \Theta_{t_n - t_{n-1}})$ in A3 started from $(t(y, \mathbf{m}), T(y, \theta))$ at time 0. Then, the connection between duality and optimal filtering can be expressed as

$$\mathcal{L}(X_{t_n} | Y_0, \ldots, Y_n) = \int h(x, t(Y_n, M_n), T(Y_n, \Theta_n))\pi(\mathrm{d}x)\mathrm{d}\mathcal{L}(D_n | Y_0, \ldots, Y_{n-1})\,. \tag{18}$$

Thus, filtering X_{t_i} in the original model (cf. Figure 1) can be achieved by filtering D_i in the dual model (cf. Figure 2). Since Θ_n evolves deterministically, optimal filtering for X reduces to filtering M_n, which has finite support with probabilities that can be computed recursively using an algorithm similar to the Baum-Welch filter, as we now describe. $\mathcal{L}(M_0, \Theta_0)$ has support on the single point $\{(\mathbf{m}_0, \theta_0)\}$; if $\mathcal{L}(M_n, \Theta_n | Y_0, \ldots, Y_{n-1})$ has support on $\Lambda_n \times \{\theta_n\}$ for $\Lambda_n \subset \mathcal{M}$ and $\theta_n \in \Theta$, and assignes probability $w_{\mathbf{m}}$ to state $(\mathbf{m}, \theta_n)$, then $\mathcal{L}(M_{n+1}, \Theta_{n+1} | Y_0, \ldots, Y_n)$ has support on $\Lambda_{n+1} \times \{\theta_{n+1}\}$, where $\Lambda_{n+1} = G(t(Y_n, \Lambda_n))$, for G and $t(y, \cdot)$ defined in (5) and (13) respectively, θ_{n+1} the solution of (7) at time $t_{n+1} - t_n$ started from $\Theta_0 = \theta_n$, and the probability associated to state $(\mathbf{n}, \theta_{n+1}) \in \Lambda_{n+1} \times \{\theta_{n+1}\}$ is

$$\mathbb{P}[M_{n+1} = \mathbf{n}, \Theta_{n+1} = \theta_{n+1} | Y_0, \ldots, Y_n] = \sum_{\substack{\mathbf{m} \in \Lambda_n \\ t(Y_n, \mathbf{m}) \geq \mathbf{n}}} w_{\mathbf{m}} \, p_{t(Y_n, \mathbf{m}), \mathbf{n}}(t_{n+1} - t_n; \theta_n) \,. \tag{19}$$

Therefore, the optimal filtering reduces to the sequential computation of the parameters θ_n, the supports Λ_n and the probabilities on each support point in Λ_n, for $n = 0, 1, \ldots$.
The computation of the probabilities (19) for all $(\mathbf{n}, \theta_{n+1}) \in \Lambda_{n+1} \times \{\theta_{n+1}\}$ can be done at a cost that is at most of order $|\Lambda_{n+1}|^2$. Therefore, the overall cost of computing the filters up to the nth observation is bounded from above by $\sum_{i=0}^{n} |\Lambda_i|^2$. If $|\Lambda_i|$ were constant with i, we would recover the complexity of the Baum-Welch filter, discussed in Section 1. However, $|\Lambda_i|$ increases with i, as a result of the successive operation of G and $t(y, \cdot)$ defined in (5) and (13) respectively. Clearly, it is hard to make further analysis on the computational complexity without some information on $t(y, \cdot)$. Here we will assume that $t(y, \mathbf{m}) = \mathbf{m} + N(y)$, where $N : \mathcal{Y} \to \mathcal{M}$, a structure that is found in all the examples we study in this paper. We then have the following key result. The proof of the Lemma is omitted.

Lemma 2.4. *For any $\Lambda \subset \mathcal{M}$ and $\mathbf{m} \in \mathcal{M}$, we define $\Lambda + \mathbf{m} = \{\mathbf{n} + \mathbf{m}; \mathbf{n} \in \Lambda\}$. Then,*

$$G(G(\Lambda) + \mathbf{m}) = G(\Lambda + \mathbf{m}) \,.$$

Proposition 2.5. *Under the assumption that $t(y, \mathbf{m}) = \mathbf{m} + N(y)$, where $N : \mathcal{Y} \to \mathcal{M}$, we have that*

$$|\Lambda_n| = G\left(\mathbf{m}_0 + \sum_{i=1}^{n} N(Y_i)\right) \leq \left(1 + \frac{d_n}{K}\right)^K$$

where $d_n = |\mathbf{m}_0 + \sum_{i=1}^{n} N(Y_i)|$.

Proof. The equality follows by successive application of Lemma 2.4. For the inequality, notice that $\mathbf{n} := \mathbf{m}_0 + \sum_{i=1}^{n} N(Y_i) \in \mathcal{M}$, with $|\mathbf{n}| = d_n$. Then, by (5), $|G(\mathbf{n})| = \prod_{i=1}^{K}(n_i + 1)$. Then, apply Jensen's inequality to $\log |G(\mathbf{n})|$ to obtain the result. □

When the observations follow a stationary process, d_n will be of order n. Therefore, the complexity of carrying out the computations involved in the filtering recursions up to iteration n, will be $\mathcal{O}(n^{2K}+1)$, where the constant depends on K but not n. We return to the issue of complexity in Section 5.

3 The dual of some stochastic processes

Following the local duality approach outlined in the previous section, here we identify the dual processes for the Cox-Ingersoll-Ross model, diffusion processes with linear coefficients and K-dimensional Wright-Fisher diffusions. Recall that d, K, l denote the state space dimension for X_t, M_t and Θ_t respectively.

3.1 CIR processes

The so-called Cox-Ingersoll-Ross (CIR) model is a nonnegative one-dimensional diffusion, that solves the SDE

$$\mathrm{d}X_t = (\delta\sigma^2 - 2\gamma X_t)\mathrm{d}t + 2\sigma\sqrt{X_t}\mathrm{d}B_t\,.$$

This name is due to Cox, Ingersoll and Ross (1985) who introduced the model in mathematical finance, although this model had been studied long before in the literature, see for example the population growth model in Section 13.C of Karlin and Taylor (1981) and the process described in Section 5 of Feller (1951). From a broader perspective, the CIR model can also be seen as a special case of a continuous-state branching process with immigration Kawazu and Watanabe (1971).

The generator of the CIR process is

$$\mathcal{A} = (\delta\sigma^2 - 2\gamma x)\frac{\mathrm{d}}{\mathrm{d}x} + 2\sigma^2 x\frac{\mathrm{d}^2}{\mathrm{d}x^2}, \qquad \delta,\gamma,\sigma > 0\,, \tag{20}$$

with domain defined as follows. With the above parametrization, and using Feller's terminology, the boundary point $+\infty$ is natural for all choices of parameters, while 0 is regular if $\delta < 2$ and entrance if $\delta \geq 2$. Define

$$\mathscr{D}_0(\mathcal{A}) = \Big\{ f \in C_0([0,\infty)) \cap C^2((0,\infty)) :\ \mathcal{A}f \in C_0([0,\infty)) \Big\},$$

where $C_0([0,\infty))$ is the space of continuous functions vanishing at infinity, and

$$\mathscr{D}(\mathcal{A}) = \begin{cases} f \in \mathscr{D}_0(\mathcal{A}), & \text{if } \delta \geq 2, \\ f \in \mathscr{D}_0(\mathcal{A}) : \lim_{x\to 0} x^{\delta/2} e^{-(\gamma/\sigma^2)x} f'(x) = 0, & \text{if } 0 < \delta < 2. \end{cases}$$

Then $\{(f, \mathcal{A}f) : f \in \mathscr{D}(\mathcal{A})\}$ generates a Feller semigroup on $C_0([0,\infty))$. See Theorems 8.1.2 and 8.2.1 in Ethier and Kurtz (1986).

Previous results on duality for the CIR model include a Laplace duality, that is with respect to a function of type $h(x,y) = e^{-axy}$. See, e.g., Hutzenthaler and Wakolbinger (2007). Here however we identify a new, gamma-type duality relation, which has as special cases a moment and a Laplace duality. First, note that the CIR process is reversible with respect to

$$\pi \equiv \text{Gamma}(\delta/2, \gamma/\sigma^2)\,.$$

Now, let $d = 1$, $K = 1$ and $l = 1$, and define, for $\theta > 0$, the function

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

This function can be identified as the Radon-Nikodym derivative of a Gamma$(\delta/2 + m, \theta)$ distribution with respect to π. The family of gamma distributions that arises by varying $m \in \mathbb{Z}_+$ and $\theta > 0$, defines a subset of the family of gamma distributions that is conjugate to emission densities that as a function of x are proportional to

$$x^n e^{-\lambda x}\,, \qquad n \in \mathbb{Z}_+, \lambda > 0\,,$$

in which case t and T in A2 coincide with

$$t(y, m) = n + m\,, \qquad T(y, \theta) = \theta + \lambda\,.$$

Such type of emission density arises, for example, for observations $Y_n = n$ distributed as Poisson with intensity λX_{t_n}, giving rise to a dynamic version of the Poisson-gamma conjugate Bayesian model.

On the other hand, $h(\cdot, m, \theta)$ belongs to the domain of $\mathcal{A}$ only when $\theta \geq \gamma/\sigma^2$, in which case $h \in C_0^2([0,\infty))$. In order to be able to use local duality as in A4 we will assume that the family is defined as

$$\mathcal{F} = \{h(x, m, \theta)\pi(\mathrm{d}x), m \in \mathbb{Z}_+, \theta \geq \gamma/\sigma^2\}$$

but we will return to the case $\theta < \gamma/\sigma^2$ at the end of this subsection. Then a simple computation yields

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

Motivated by this structure, and with view to achieving the local duality in (17) we consider a two-component process (M_t,Θ_t) with generator A as in (16), where

$$\lambda(m) = 2\sigma^2 m\,, \quad r(\theta) = 2\sigma^2\theta(\gamma/\sigma^2-\theta)\,, \quad \rho(\theta)=\theta\,.$$

It is then easy to check that local duality holds, namely

$$\mathcal{A}h(\cdot,m,\theta)(x) = Ah(x,\cdot,\cdot)(m,\theta)\,.$$

Additionally, the conditions that are required to derive (9) from this local duality are satisfied. In this example, the solution of the dynamical system (7) for $\Theta_0=\theta$ is given by

$$\Theta_t = \frac{\gamma}{\sigma^2}\frac{\theta e^{2\gamma t}}{\theta e^{2\gamma t}+(\gamma/\sigma^2-\theta)}\,,$$

which in conjunction with Proposition 2.1 implies that the transition probabilities for the death process simplify to binomial probabilities

$$p_{m,m-i}(t;\theta) = \mathrm{Bin}\left(m-i\,;\,m,\,\frac{\gamma}{\sigma^2}(\theta e^{2\gamma t}+(\gamma/\sigma^2-\theta))^{-1}\right)\,.$$

Therefore, we have all the ingredients necessary to implement the filtering algorithm. Finally, note that if $\theta_0 \geq \gamma/\sigma^2$, then $\theta_n \geq \gamma/\sigma^2$ for all n.
Notice that the result on the transition probabilities above, together with Proposition 2.2, implies the following interesting property of the CIR process:

$$\begin{aligned}&\psi_t(\mathrm{Gamma}(m+\delta/2,\theta)) =\\ &\sum_{k=0}^{m}\mathrm{Bin}\left(k\,;\,m,\,\frac{\gamma}{\sigma^2}(\theta e^{2\gamma t}+(\gamma/\sigma^2-\theta))^{-1}\right)\mathrm{Gamma}\left(k+\delta/2,\,\frac{\gamma}{\sigma^2}\frac{\theta e^{2\gamma t}}{\theta e^{2\gamma t}+(\gamma/\sigma^2-\theta)}\right)\,.\end{aligned} \tag{21}$$

This result has been obtained before, using a completely different approach; the case $\delta=1$ can be shown directly by elementary calculations using a change of variables and binomial expansion of the left-hand-side; the general case was proved in Chaleyat-Maurel and Genon-Catalot (2006),

see Proposition 3.4 and the associated Lemma 3.1, after some rather heavy calculations. The result in (21) leads to a computable filter, as we showed in Proposition 2.3, which is precisely the result also obtained in Chaleyat-Maurel and Genon-Catalot (2006) for the CIR process. It is neat that using duality and the generic result in Proposition 2.2, this result can be obtained in a straightforward manner. The proof in Chaleyat-Maurel and Genon-Catalot (2006) is based on the following known series expansion of the CIR transition kernel, see expression (80) in Chaleyat-Maurel and Genon-Catalot (2006) and p. 334 of Karlin and Taylor (1981), which can be re-expressed as a Poisson mixture of gamma distributions as follows:

$$P_t(x, \mathrm{d}x') = \sum_{k\geq 0} \text{Poisson}\left(k\,;\, \frac{\gamma}{\sigma^2}\frac{1}{e^{2\gamma t}-1}x\right) \text{Gamma}\left(k+\delta/2, \frac{\gamma}{\sigma^2}\frac{e^{2\gamma t}}{e^{2\gamma t}-1}\right). \tag{22}$$

It is interesting that instead of deriving (21) from (22), which in any case is laborious, one can prove the latter using duality and then obtain (22) by taking $\theta = (m+\delta/2)/x$ and letting $m \to \infty$ in (22).

In view of the arguments of Section 2, it follows that, for $\theta < \gamma/\sigma^2$, $h \notin \mathscr{D}(\mathcal{A})$, hence duality in the sense of A3 cannot be established using local duality. However, in view of the result (21) that has already been obtained in Chaleyat-Maurel and Genon-Catalot (2006), it is obvious that duality still holds in this case. This also shows the limitation of the functional analytic method for establishing duality: it is a very powerful when all formal requirements are met, but there will be examples, like this one, where (9) would have to be established by alternative arguments. Nevertheless, a formal calculation using the generator reveals the dual even when $\theta < \gamma/\sigma^2$.

3.2 Linear diffusion processes

We consider the scalar Ornstein-Uhlenbeck process that solves an SDE of the form

$$\mathrm{d}X_t = -\frac{\sigma^2}{\alpha}(X_t-\gamma)\mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_t\,,$$

which is reversible with respect to the Gaussian distribution,

$$\pi(\mathrm{d}x) \equiv \text{Normal}(\gamma, \alpha)\,.$$

The generator is given by

$$\mathcal{A} = (\sigma^2\gamma/\alpha - \sigma^2 x/\alpha)\frac{\mathrm{d}}{\mathrm{d}x} + \sigma^2\frac{\mathrm{d}^2}{\mathrm{d}x^2}$$

with domain $C_0^2((-\infty,\infty))$. In this model, we have $d=1, K=0, l=2$, where $\theta=(\mu,\tau)\in\mathbb{R}\times\mathbb{R}_+$, and

$$h(x,\mu,\tau)=\left(\frac{\alpha}{\tau}\right)^{1/2}\exp\left\{-\frac{(x-\mu)^2}{2\tau}+\frac{(x-\gamma)^2}{2\alpha}\right\},$$

which can be easily recognised as the Radon-Nikodym derivative between a Normal(μ,τ) and π. The measures $h(x,\mu,\tau)\pi(\mathrm{d}x)$ are conjugate to emission densities that as a function of x are proportional to

$$\exp\left\{-\frac{1}{2\lambda}(x-c)^2\right\}, \quad \lambda>0, c\in\mathbb{R},$$

with $T(y,\theta)=((\lambda\mu+\tau c),\lambda\tau)/(\lambda+\tau)$. Such density arises for example with data $Y_n=c$ that is Gaussian with mean X_{t_n} and variance λ. As with the CIR process, we have the technical problem that this function belongs to $\mathscr{D}(\mathcal{A})$ only for $\tau<\alpha$, hence we will restrict to this case and define $\Theta=\{(\mu,\tau):\mu\in\mathbb{R},0<\tau<\alpha\}$. A direct calculation gives that

$$\mathcal{A}h(\cdot,\mu,\tau)(x)=\frac{\sigma^2}{\alpha}(\gamma-\mu)\frac{\partial}{\partial\mu}h(x,\mu,\tau)+2\sigma^2(1-\tau/\alpha)\frac{\partial}{\partial\tau}h(x,\mu,\tau).$$

This suggests that the dual is purely deterministic and described in terms of the ODEs:

$$\mathrm{d}\mu_t/\mathrm{d}t=\frac{\sigma^2}{\alpha}(\gamma-\mu_t)\mathrm{d}t, \quad \mathrm{d}\tau_t/\mathrm{d}t=2\sigma^2(1-\tau_t/\alpha)\mathrm{d}t.$$

Duality with respect to this deterministic process implies that the filter evolves within the Gaussian family and the computational cost is linear in n, that is we are dealing with a finite-dimensional filter.

Of course, all this is known: the ODEs above are the well known equations for the first two moments of linear SDEs, and the filter is the Kalman filter. Thus, within the assumptions we have made in this article, the finite dimensional filter corresponds to the special case where the dual is purely deterministic. We considered $d=1$ for simplicity, but the results carry over to multi-dimensional stationary linear SDEs. The same discussion as for the CIR applies here regarding the restrictions posed by needing that $h\in\mathscr{D}(\mathcal{A})$. We return to this issue in Section 5.

3.3 Wright-Fisher diffusions

Wright-Fisher (WF) processes are K-dimensional diffusions with paths confined in the $(K-1)$-dimensional simplex

$$\Delta_K=\left\{x\in[0,1]^K:\sum\nolimits_{i=1}^K x_i=1\right\}. \tag{23}$$

These processes approximate large-population discrete Wright-Fisher reproductive models with non overlapping generations, and describe the time-evolution of the species abundancies when the individuals in the underlying population are subject to random genetic drift and, possibly, mutation, selection and recombination. See for example Dawson (1993) for a review. Here we are interested in the case without selection nor recombination, and with *parent-independent mutation*. That is we consider a WF diffusion driven by the SDEs

$$\mathrm{d}X_{t,i} = \frac{1}{2}(\alpha_i - |\boldsymbol{\alpha}|X_{t,i})\mathrm{d}t + \sum_{j=1}^{K}\sqrt{X_{t,i}(\delta_{ij} - X_{t,j})}\mathrm{d}B_j\,, \quad i = 1,\dots,K,$$

where δ_{ij} denoting Kronecker delta, $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_K) \in \mathbb{R}_+^K$ and $|\boldsymbol{\alpha}| = \sum_{i=1}^K \alpha_i$. Note that this is a hypoelliptic diffusion, i.e., the square of the diffusion matrix is not full rank, as a result of the constraint $\sum_i x_i = 1$. Even though we could work with an elliptic diffusion for the $K-1$ variables, it is the formulation above that is desirable for identifying the dual, as we will show. The generator of the process is

$$\mathcal{A} = \frac{1}{2}\sum_{i=1}^{K}(\alpha_i - |\boldsymbol{\alpha}|x_j)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{K} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j}, \tag{24}$$

acting on $C^2(\Delta_K)$ functions, whose closure generates a strongly continuous contractive semigroup on $C(\Delta_K)$. See Ethier and Kurtz (1981) for details. Such diffusion is reversible with respect to the Dirichlet distribution

$$\pi(\mathrm{d}x_1,\dots,\mathrm{d}x_K) = \frac{\Gamma(|\boldsymbol{\alpha}|)}{\prod_{j=1}^K \Gamma(\alpha_j)} x_1^{\alpha_1-1}\dots x_K^{\alpha_K-1}\mathrm{d}x_1\cdots\mathrm{d}x_K, \qquad x \in \Delta_K \tag{25}$$

It is well know that a moment duality, that is with respect to functions of type $h(x,m) = x^m$, holds between the one-dimensional WF diffusion without mutation and the block-counting process of Kingman's coalescent Kingman (1982). See, e.g., Section 1.2 in Jansen and Kurt (2013). This type of duality can be thought of as a special case of a much more general duality framework, which holds between the Fleming-Viot process and Kingman's coalescent. See Section 4.4. However, the dual process we elicit for the K-dimensional WF diffusion with mutation seems to be new, and it is the one required for optimal filtering.

In this model we have $d = K \geq 2$ and $l = 0$, therefore there is no deterministic component in the dual process. We denote

$$x^{\mathbf{m}} = x_1^{m_1}\cdots x_K^{m_K}, \qquad x \in \Delta_K,\ \mathbf{m} \in \mathcal{M}\,,$$

and define

$$h(x,\mathbf{m}) = \frac{\Gamma(|\boldsymbol{\alpha}|+|\mathbf{m}|)}{\Gamma(|\boldsymbol{\alpha}|)}\prod_{j=1}^{K}\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j+m_j)}x^{\mathbf{m}}, \tag{26}$$

whence clearly $h(\cdot,\mathbf{m}) \in \mathscr{D}(\mathcal{A})$. This can be identified with the Radon-Nikodym derivative between a Dirichlet distribution with parameters $(\alpha_1+m_1,\dots,\alpha_K+m_K)$ and π, and it is conjugate to emission densities that as a function of x are proportional to

$$x_1^{n_1}\cdots x_K^{n_K}, \qquad n_i \in \mathbb{Z}_+, i=1,\dots,K$$

in which case t in A2 coincides with $t(y,m) = n+m$. Such type of emission density arises, for example, for observations $Y_n = (n_1,\dots,n_K)$ distributed as Multinomial with parameters $X_{t_n} = (X_{t_n,1},\dots,X_{t_n,K})$, giving rise to a dynamic version of the Dirichlet-Multinomial conjugate Bayesian model.
Then we have

$$\begin{aligned}\mathcal{A}h(x,\mathbf{m}) &= \sum_{i=1}^{K}\left(\frac{\alpha_i m_i}{2}+\binom{m_i}{2}\right)\frac{\Gamma(|\boldsymbol{\alpha}|+|\mathbf{m}|)}{\Gamma(|\boldsymbol{\alpha}|)}\prod_{j=1}^{K}\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j+m_j)}x^{\mathbf{m}-\mathbf{e}_i}\\ &\quad -\sum_{i=1}^{K}\left(\frac{|\boldsymbol{\alpha}|m_i}{2}+\binom{m_i}{2}+\frac{1}{2}m_i\sum_{j\neq i}m_j\right)\frac{\Gamma(|\boldsymbol{\alpha}|+|\mathbf{m}|)}{\Gamma(|\boldsymbol{\alpha}|)}\prod_{j=1}^{K}\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j+m_j)}x^{\mathbf{m}}\\ &= \frac{|\boldsymbol{\alpha}|+|\mathbf{m}|-1}{2}\sum_{i=1}^{K}m_i h(x,\mathbf{m}-\mathbf{e}_i) - \frac{|\mathbf{m}|(|\boldsymbol{\alpha}|+|\mathbf{m}|-1)}{2}h(x,\mathbf{m}).\end{aligned}$$

This suggests considering a one-component dual process, with M_t a Markov jump process with generator A obtained by letting

$$\lambda(|\mathbf{m}|) = (|\boldsymbol{\alpha}|+|\mathbf{m}|-1)/2, \quad \rho(\theta)\equiv 1,$$

in (16). Since $h(x,\cdot)\in\mathscr{D}(A)$, it is then easy to check that the local duality condition

$$\mathcal{A}h(\cdot,\mathbf{m})(x) = Ah(x,\cdot)(\mathbf{m})$$

holds. Hence the WF diffusion with parent-independent mutation X_t and the death process M_t on $\mathbb{Z}_+^K$, which jumps from $\mathbf{m}$ to $\mathbf{m}-\mathbf{e}_j$ at rate $m_j(|\boldsymbol{\alpha}|+|\mathbf{m}|-1)/2$, are dual with respect to the above h in the sense of A3. The transition probabilities of M_t are as in Proposition 2.1.

Filtering the WF model when $K = 2$ on the basis of binomial data was studied in Chaleyat-Maurel and Genon-Catalot (2009). One can appreciate the strength of the approach we introduce here, since it is straightforward to obtain the filtering recursion using the dual and Proposition 2.1 for any K. It has to be noted that, in our opinion, one of the reasons why the results are harder to obtain using the approach in Chaleyat-Maurel and Genon-Catalot (2009), is because they decide to work with the elliptic WF model, which is a scalar diffusion since $K = 2$. Working with the elliptic model hides the structure of duality, which is immediately apparent in the hypoelliptic model.

4 Nonparametric filtering

Our results for the WF can be easily extended to the case $K = \infty$. In terms of filtering, the signal has a countable state space and coincides with the infinite-dimensional diffusion studied by Ethier (1981), for which absolute continuity is not lost and a transition density is still available. In particular, the reversible measure of the process, given by the GEM distribution, is conjugate to emission densities that as a function of x are proportional to

$$\prod_{i=1}^{\infty} x_i^{n_i}, \qquad n_i \in \mathbb{Z}_+, \ \sum_i n_i = n, \ n \in \mathbb{N}.$$

See Ethier and Griffiths (1993). In this section instead we will obtain a much more general, and far less straightforward result, where the signal is "fully" nonparametric, that is the discrete-time sampling of a diffusion process taking values in the space of discrete probability measures. Extending the intuition from the previous sections, the filtering problem can be seen as the dynamic version of a Bayesian nonparametric inferential procedure. Given that conjugate Bayesian nonparametric inference might be unfamiliar to some readers, we quickly summarise the main ideas below.

4.1 The Bayesian nonparametric model

Let $\mathcal{Y}$ be a Polish space. The Bayesian approach assumes the observations $Z_i \in \mathcal{Y}$, $i \geq 1$, are infinitely exchangeable, that is their joint law is invariant with respect to any finite permutation. The nonparametric approach to a Bayesian inferential problem consists in assigning a *prior* distribution Q to an infinite-dimensional subset of the space $\mathscr{P}(\mathcal{Y})$ of Borel probability measures on $\mathcal{Y}$, and evaluate the *posterior* (or conditional) distribution $Q \mid Z_1, \ldots, Z_n$, given observations

$Z_1, \ldots, Z_n$. A challenge in this nonparametric framework is the lack of "likelihood", which makes standard Bayesian calculations non-trivial. The most important example of nonparametric model is given by the Dirichlet prior (Ferguson, 1973), which is the law induced on $\mathscr{P}(\mathcal{Y})$ by a Dirichlet process, defined as follows. Denote by $\mathscr{M}(\mathcal{Y})$ the space of finite Borel measures on $\mathcal{Y}$. A discrete random probability measure $P \in \mathscr{P}(\mathcal{Y})$ is said to be a *Dirichlet process* with parameter $\alpha \in \mathscr{M}(\mathcal{Y})$, denoted $P \sim \Pi_\alpha$, if for any $K \geq 2$ and measurable partition $A_1, \ldots, A_K$ of $\mathcal{Y}$, the projection vector $(P(A_1), \ldots, P(A_K))$ has Dirichlet distribution (25) with $\alpha_j = \alpha(A_j)$. Two features of Dirichlet priors are crucial for the development of the remainder of the section. The first is the *conjugacy* property. That is, the family of Dirichlet priors is closed under the operation of conditioning to the observations, and if $P \sim Q$ and $Z_i \mid P \sim^{iid} P$, then the posterior is still Dirichlet with updated parameter measure $\alpha + \sum_{i=1}^n \delta_{z_i}$. Here δ_y denotes a point mass at y. The second property of interest here is *mutual singularity*, more specifically given by the fact that Dirichlet priors with different non atomic parameter measures are mutually singular. See, e.g., Section 3.2. of Ghosh and Ramamoorthi (2003).
An extension of the Dirichlet process which is crucial for our purposes is given by a *mixture of Dirichlet processes* (MDP) (Antoniak, 1974), which, roughly speaking, corresponds to a Dirichlet process with randomized parameter measure α.

Definition 4.1 (Antoniak, 1974)**.** *Let $(\mathcal{Y}, \mathscr{Y})$ and $(\mathcal{U}, \mathscr{U})$ be two measurable spaces. Then P is said to be an MDP with transition measure $\gamma : \mathcal{U} \times \mathscr{Y} \to [0, \infty)$ and mixing distribution $H : \mathcal{U} \to [0, 1]$, denoted*

$$P \sim \int_{\mathcal{U}} \Pi_{\gamma(u,\cdot)} \mathrm{d}H(u),$$

if, for every measurable partition $A_1, \ldots, A_K$ of $\mathcal{Y}$, one has

$$(P(A_1), \ldots, P(A_K)) \sim \int_{\mathcal{U}} \pi_{K,u}(\cdot) \mathrm{d}H(u),$$

where $\pi_{K,u}$ denotes (25) with parameters $(\gamma(u, A_1), \ldots, \gamma(u, A_K))$.

That is, an MDP is defined (and characterized) as a Kolmogorov-consistent family of probability measures, identified with mixtures of Dirichlet distributions when projected on a finite-dimensional subset. Similarly to the Dirichlet process, an MDP enjoys a conjugacy property, which for ease of reference we state in the following Proposition.

Proposition 4.2 (Antoniak, 1974)**.** *Let P be an MDP as in Definition 4.1, with $\gamma_u = \gamma(u, \cdot)$, and let $\boldsymbol{z} = (z_1, \ldots, z_n)$ be an n-sized sample observed from P. Then*

$$P \mid z_1, \ldots, z_n \sim \int_{\mathcal{U}} \Pi_{\gamma_u + \sum_{i=1}^n \delta_{z_i}} \mathrm{d}H_{\boldsymbol{z}}(u),$$

that is the posterior mixture $P \mid z_1, \ldots, z_n$ is an MDP with transition measure $\gamma_u + \sum_{i=1}^n \delta_{z_i}$ and mixing distribution $H_{\boldsymbol{z}}(\cdot) = H(\cdot \mid z_1, \ldots, z_n)$.

In general, the conditional distribution $H_{\mathbf{z}}$ has a non trivial-expression, due to its dependence on the presence or absence of atoms in some or all of the $\{\gamma_u, u \in \mathcal{U}\}$. In Section 4.3 we will see that considering a simplified expression is sufficient in our setting.

4.2 Neutral Fleming-Viot processes

If the simplest Bayesian nonparametric approach to a static problem concerning a set of discrete random probability measures is to use a Dirichlet prior, when the object of inference $X_n = X_{t_n}$ is the discrete-time sampling of a purely-atomic measure-valued diffusion it is natural to assign a prior to the space of continuous sample paths which is marginally Dirichlet. The best known diffusion satisfying this requirement is the Fleming-Viot (FV) process with neutral mutations and without selection nor recombination. FV processes form a large class of measure-valued diffusions taking values in the space of purely atomic Borel probability measures. Exhaustive reviews can be found in Ethier and Kurtz (1993) and Dawson (1993). These describe the time evolution of an infinite population whose individuals are born with the same chromosomal information of their parents, experience spatial displacement according to some Feller operator called *mutation process*, and possibly other evolutionary mechanisms such as selection or recombination, described by appropriate transition kernels. The subclass of FV processes we are interested in is known as the infinitely-many-neutral-alleles model with parent-independent mutation (henceforth for simplicity we will refer to this specific class also as FV process). In this sub-case, mutation is modeled as a jump to a location sampled from a diffuse probability measure defined on an uncountable space $\mathcal{Y}$, which for simplicity we assume to be compact, whereby every mutant is almost surely of a new type. Furthermore, mutations occur at constant rate, which is thus independent of the parent type. This mechanism gives rise to the drift of the process, whereas the volatility is produced by the so-called random genetic drift, given by the *resampling* of existing types caused by births and deaths.

Let $\varphi \in B(\mathcal{Y}^m)$ be real-valued, measurable and bounded on $\mathcal{Y}^m$, and define

$$f_m(x) = \langle \varphi, x^m \rangle := \int_{\mathcal{Y}^m} \varphi(z_1, \dots, z_m) x(\mathrm{d}z_1) \dots x(\mathrm{d}z_m), \qquad x \in \mathscr{M}(\mathcal{Y}).$$

The generator of the neutral FV process can be written

$$\mathcal{A} f_m(x) = \sum_{i=1}^{m} \langle M_i \varphi, x^m \rangle + \frac{1}{2} \sum_{1 \le j \ne i \le m} \langle \Phi_{ij} \varphi - \varphi, x^m \rangle \tag{27}$$

where $M_i \varphi$ denotes the mutation operator

$$M g(z) = \frac{|\boldsymbol{\alpha}|}{2} \int_{\mathcal{Y}} [g(z') - g(z)] P_0(\mathrm{d}z'), \qquad g \in B(\mathcal{Y}), \tag{28}$$

applied to the ith coordinate of φ. Here $P_0 \in \mathscr{P}(\mathcal{Y})$ is the mutant offspring distribution, the distribution from which new types are sampled, and $|\boldsymbol{\alpha}|/2$ is the rate at which mutations occur. A property of the above FV process which is of particular interest for the purpose of this paper is its stationarity and reversibility with respect to the law Π_α of a Dirichlet process with parameter $|\boldsymbol{\alpha}| P_0$. For this result and for more details on the infinite-alleles model we refer the reader to Ethier and Kurtz (1986), Chapter 10, or Ethier and Kurtz (1993), Section 9.2.
Note that the K-dimensional WF process with operator (24) is obtained as a special case of the infinitely-many-alleles model X_t by considering its projection $(X_t(A_1), \dots, X_t(A_K))$ onto a measurable partition of $\mathcal{Y}$, and letting $\alpha_j = |\boldsymbol{\alpha}| P_0(A_j)$. The same result can be equivalently obtained by identifying the type space $\mathcal{Y}$ with $\{1, \dots, K\}$ and letting $\alpha_j = |\boldsymbol{\alpha}| P_0(j)$. See, e.g., Dawson (2009).

4.3 Filtering the Fleming-Viot process

Without loss of generality we assume single observations $Y_1, Y_2, \dots$ are collected at discrete time points, where $Y_n \mid X_n \sim X_n$, so that the emission distribution is the current hidden state $X_n = X_{t_n}$ of the FV signal. The crucial observation here is that there is no common dominating measure for the states X_n, which implies that Bayes' theorem cannot be applied and the update operator in (2) is not defined. If, on the one hand, the problem of domination for the update operator can be easily overcome be exploiting the conjugacy property of Dirichlet and mixtures of Dirichlet processes, on the other hand the mutual singularity of Dirichlet priors with different non atomic parameter measures (cf. Section 4.1) makes the strategy used in Sections 2 and 3.3 for devising a computable filter unfeasible in this setting. More specifically, the derivation of a

finite expression for the prediction operator applied to some suitable class of measures does not help devising a recursive formula, as is done for example in Proposition 2.3, since there is no likelihood to recombine with the distribution of the signal.

It turns out, though, that the results obtained in the previous section can be used for studying the measure-valued case by exploiting the fact that a K-dimensional WF diffusion can be obtained from a FV process as a projection (cf. Section 4.2). Here the parent-independence of the mutation process is crucial for the projected process to retain the Markov property, since the rate at which individuals mutate from i to j, that is jump from A_i to A_j, only depends on a constant and on the mass $P_0(A_j)$ the mutant offspring distribution puts on the arrival class, and not on the specific position of the particle in A_i, which is lost in the binning.

Before presenting the results of this section, we need to adapt the notation used so far. Given that a sample from a Dirichlet process will almost surely feature ties among the observations, due to the discrete nature of the random measure, define

$$\mathcal{Y}^* = \{\, \mathbf{y}_m = (y_1, \ldots, y_{K_m}, 0, 0, \ldots),\ y_j \in \mathcal{Y},\ y_1 \neq \cdots \neq y_{K_m},\ K_m \leq m \in \mathbb{Z}_+ \,\}$$

to be the space of infinite vectors which record in order of appearance the $K_m \leq m$ distinct values observed in an m-sized sample, and

$$\mathcal{M}^* = \{\, \mathbf{m} = (m_1, \ldots, m_{K_m}, 0, 0, \ldots),\ m_j \in \mathbb{Z}_+,\ |\mathbf{m}| = m,\ K_m \leq m \in \mathbb{Z}_+ \,\}$$

to be the space of associated multiplicities. The peculiarity with respect to the WF case is that the effective dimension of the index space increases with the number of observations collected. In particular, the growth rate of K_m is showed in Korwar and Hollander (1973) to be $K_m \approx \log m$. Define also the class of finite mixtures of Dirichlet priors which share the diffuse part α of the parameter measure, given by

$$\bar{\mathscr{F}}_f = \Big\{ \sum\nolimits_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} \Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}}(\mathrm{d}x) : \Lambda \subset \mathcal{M}^*\,,\ |\Lambda| < \infty\,,\ w_{\mathbf{m}} \geq 0\,, \\ \sum\nolimits_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} = 1,\ \mathbf{m} \in \mathcal{M}^*, \mathbf{y}_m \in \mathcal{Y}^* \Big\},$$

where we have set $\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}} = \sum_{j=1}^{K_m} m_j \delta_{y_j}$. The following result provides the update step of the algorithm for the nonparametric case. See the Appendix for a proof.

Proposition 4.3. *Given* $(\mathbf{m}, \mathbf{y}_m) \in \mathcal{M}^* \times \mathcal{Y}^*$ *and a further observation* $y \in \mathcal{Y}$*, define* $\mathbf{y}_{m+1} = \mathbf{y}_m$ *and* $t(y, \mathbf{m}) = \mathbf{m} + \mathbf{e}_j$ *if* $y = y_j$ *for some* $j = 1, \ldots, K_m$*, and* $\mathbf{y}_{m+1} = (y_1, \ldots, y_{K_m}, y, 0, 0, \ldots)$

and $t(y,\mathbf{m}) = \mathbf{m} + \mathbf{e}_{K_m+1}$ *otherwise. Let also* p_0 *be the Radon-Nikodyn derivative of* P_0 *with respect to the Lebesgue measure. Then*

$$\phi_y\Big(\sum\nolimits_{\mathbf{m}\in\Lambda} w_{\mathbf{m}}\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}}\Big) = \sum_{\mathbf{n}\in t(y,\Lambda)} \hat{w}_{\mathbf{n}}\Pi_{\alpha+t(y,\mathbf{m})\cdot\boldsymbol{\delta}_{\mathbf{y}_{m+1}}}, \tag{29}$$

where, for $\mathbf{n} = t(y,\mathbf{m})$,

$$\hat{w}_{\mathbf{n}} \propto \begin{cases} w_{\mathbf{m}}\dfrac{m_j}{|\boldsymbol{\alpha}|+|\mathbf{m}|}, & \text{if } y = y_j, \\ w_{\mathbf{m}}\dfrac{|\boldsymbol{\alpha}|p_0(y)}{|\boldsymbol{\alpha}|+|\mathbf{m}|}, & \text{if } y = y_{K_m+1}. \end{cases}$$

As already anticipated, the update step simply exploits the conjugacy property of MDPs in the present framework.

We now turn to prediction. First, in order to make their formal representation independent of the underlying partition of $\mathcal{Y}$, write the transition probabilities in Proposition 2.1 as

$$p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t) = C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t)\Big(\prod_{h=0}^{|\mathbf{i}|-1}\lambda_{|\mathbf{m}|-h}\Big)\binom{|\mathbf{m}|}{\mathbf{m}}^{-1}\binom{|\mathbf{i}|}{\mathbf{i}}\binom{|\mathbf{m}|-|\mathbf{i}|}{\mathbf{m}-\mathbf{i}}, \tag{30}$$

where $\mathbf{m},\mathbf{i} \in \mathcal{M}^*$, $\mathbf{0} \le \mathbf{i} \le \mathbf{m}$, and the dependence on θ has been suppressed since there is no deterministic dual here. Note that the multinomial coefficients in (30) depend only on finitely-many components. The following result, whose proof can also be found in the Appendix, provides the prediction step of the recursive algorithm and the closure of the class $\bar{\mathscr{F}}_f$ with respect to the time propagation.

Proposition 4.4. *For any finite mixture in* $\bar{\mathscr{F}}_f$*, we have*

$$\psi_t\Big(\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}}\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}}\Big) = \sum_{\mathbf{n}\in G(\Lambda)}\Big(\sum_{\mathbf{m}\in\Lambda,\,\mathbf{m}\ge\mathbf{n}} w_{\mathbf{m}}p_{\mathbf{m},\mathbf{n}}(t)\Big)\Pi_{\alpha+\mathbf{n}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}}$$

with $p_{\mathbf{m},\mathbf{n}}(t)$ *as in (30).*

The previous proposition can be interpreted as follows. Roughly speaking, during the interval of length t the prediction operator removes pieces of information collected in the past in a random amount, guaranteeing, in the long run, convergence to the stationary measure Π_α. The weight associated to the mixture component indexed by $\mathbf{n}$ at time t, where $\mathbf{n}$ is not greater than the starting index $\mathbf{m}$, is then given by the sum of weights of those mixture components that carry

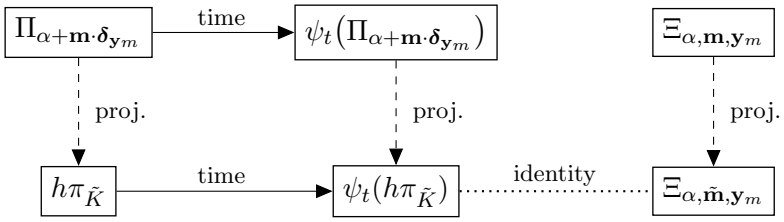

Figure 3: The projection onto a $\tilde{K}$-dimensional partition of the time propagation $\psi_t\big(\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}^*_m}}\big)$ of the initial distribution $\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}^*_m}}$ of the FV process equals the time propagation $\psi_t(h\pi_{\tilde{K}})$ of the projection $h\pi_{\tilde{K}}$ of $\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}^*_m}}$. Since the projection $\Xi_{\alpha,\tilde{\mathbf{m}},\mathbf{y}_m}$ onto the same partition of the mixture of Dirichlet processes $\Xi_{\alpha,\mathbf{m},\mathbf{y}_m}$ equals $\psi_t(h\pi_{\tilde{K}})$, and the projections are consistent, we have that $\psi_t\big(\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}^*_m}}\big) = \Xi_{\alpha,\mathbf{m},\mathbf{y}_m}$.

more information, i.e. are indexed by $\mathbf{m} \geq \mathbf{n}$, rescaled by the probability of going from $\mathbf{m}$ to $\mathbf{n}$, that is the probability of losing "on the way" the exact amount of information $\mathbf{m} - \mathbf{n}$.
With the aim of emphasizing the different strategy followed with respect to those described in Section 2 and applied in Section 3, we sketch here the main argument of the proof, detailed in the Appendix and stylized in Figure 3. If one shows that the prediction operator applied to a single Dirichlet prior yields

$$\psi_t\Big(\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}^*_m}}\Big) = \sum_{\mathbf{n}\in G(\mathbf{m})} p_{\mathbf{m},\mathbf{n}}(t)\Pi_{\alpha+\mathbf{n}\cdot\boldsymbol{\delta}_{\mathbf{y}m}}, \tag{31}$$

with G as in (5), then Proposition 4.4 follows by direct computation similarly to (14). Denote by $\tilde{\mathbf{m}}$ the vector of multiplicities resulting from binning $(\mathbf{m}, \mathbf{y}_m)$ into an arbitrary but fixed measurable partition of $\mathcal{Y}$. Since the corresponding projection of $\Pi_{\alpha+\mathbf{m},\mathbf{y}_m}$ is a Dirichlet distribution $\pi_{\tilde{K}}$ as in (25), with $\tilde{K}$ being the number of classes in the partition, and the corresponding projection of the FV process with generator (27) is a $\tilde{K}$-dimensional WF process, from Section 3.3 we have that

$$\psi_t\Big(h(\mathbf{x},\tilde{\mathbf{m}})\pi_{\tilde{K}}(\mathrm{d}\mathbf{x})\Big) = \sum_{\mathbf{n}\in G(\tilde{\mathbf{m}})} p_{\tilde{\mathbf{m}},\mathbf{n}}(t)h(\mathbf{x},\mathbf{n})\pi_{\tilde{K}}(\mathrm{d}\mathbf{x}). \tag{32}$$

Given now that an MDP is characterized by its finite-dimensional projections (cf. Definition 4.1), the projection of the left-hand side of (31) equals the left-hand side of (32). To prove (31)

then it suffices to show that the projection onto the same partition of the right hand side of (31) equals the right hand side of (32), and that the mixture weights are consistent with respect to fragmentation and merging of classes. It is worth emphasizing again that here a key factor is that binning the FV process yields the WF process with generator (24) only if mutations are parent-independent and driven by (28).

4.4 Nonparametric signals, duality and Kingman's coalescent

Although the derivation of the recursive algorithm for the FV case does not use duality arguments, there are some interesting connections to duality worth pointing out. The FV process possesses a *function-valued dual* which lives in the space $\cup_{m\geq 1} B(\mathcal{Y}^m)$, whichis linked to Kingman's coalescent with mutation. See Section 3 in Ethier and Kurtz (1993) and Section 5.6 in Dawson (1993). This of course provides a much greater level of generality than the type of dual process considered in Section 3.3. On the one hand, such level of generality for the dual is extremely useful for characterizing the process and deriving other sample path properties. For example, the existence of the dual for a sufficiently rich class of functions implies uniqueness of the original process, via the Stroock-Varadhan's martingale characterization. See Section 4.4 in Ethier and Kurtz (1986) and Section 1.6 in Etheridge (2000). On the other hand, the restriction of the FV dual to the WF setting does not yield the dual process derived in Section 3.3, since the binning hides some important information about the coalescence events at the level of particles. It is only with a specific parametrization of the function spaces of interest, which is therefore crucial in this case, that one is able to manipulate the resulting objects in order to exploit the conjugacy of the associated statistical model. The lack of likelihood for the nonparametric signal does not allow, at such level of generality, to recombine the likelihood with the prior and derive the recursive filter.

5 Discussion

5.1 Local duality, reversibility and spectral representations

We have demonstrated that computable filtering follows from duality, in the sense described in Assumptions A2 and A3 in Section 2. A sufficient condition to establish duality is the local duality described in Assumption A4, which is based on the properties of the generator of the signal process and its relation to the semigroup operator via the Kolmogorov backward equation.

The use of this functional analytic machinery places some constraints on the duality function in A2, such as for example that as a function of x it has to vanish at infinity. Therefore, even when duality holds in the sense of A3 for functions that do not satisfy such constraints, the local duality cannot be used to prove this. On the other hand, the local duality can still be used formally to identify the dual. Both in the CIR process when $\theta < \gamma/\sigma^2$ and for the OU process when $\tau < \alpha$ (see Section 3 for details) the formal application of the generator identifies the dual correctly.

We have assumed reversibility with respect to a probability measure π, in order to transform the problem of studying how measures act on the left of the transition kernel to that of how the semigroup acts on functions; this technique was suggested in Section 2.3 of Chaleyat-Maurel and Genon-Catalot (2006) and it was put in practice in Chaleyat-Maurel and Genon-Catalot (2009) for filtering the WF process when $K = 2$. However, we already know that in the CIR the same family of gamma distributions is propagated to a mixture of gammas in the CIR process, and the Gaussian to a Gaussian in linear diffusions, independently of the assumption of stationarity. Therefore, it is interesting to study to what extent the ideas in this paper can be extended to non-stationary signals.

Another topic of investigation is the connection of the duality, as used in this paper, and results about the spectral representation of the transition kernel of the signal, for example the type of expression in (22). There are classic results about such expressions, see e.g. Chapter 13 of Karlin and Taylor (1981), and their existence seems to be related to computable filtering, see Section 6.4 of Chaleyat-Maurel and Genon-Catalot (2009), but the connection is not well understood.

5.2 Nominal complexity, stability and approximations

Our results in Section 2 show that for observations generated by a stationary process the computational cost associated with the identification of the filtering distributions grows polynomially with the number of observations, unless $K = 0$ in which case the growth is linear. However, it might be the case that most of the components in the mixture representations have negligible weight. Previous simulation studies show that after a few iterations the filter might concentrate all its mass in two or three components. See for example Table 1 in Genon-Catalot and Kessler (2004). We believe that the connection to the dual process might be very helpful in studying the effective number of components. However, there are subtleties in this line of research. Note that when X_t is ergodic, and $t_i - t_{i-1}$ is large relative to its mixing time, practically all mass of the filtering distribution will be concentrated on a single component, the "root" $(\mathbf{0}, \tilde{\theta})$ (see

A2) that corresponds to the invariant measure π. Therefore, the time evolution of the number of states with non-negligible filtering probabilities (say above a given $\epsilon \approx 0$) will depend on the number of observations per unit of time in the X process. This aspect deserves careful study.

5.3 Nonparametric filtering and dependent processes in Bayesian nonparametrics

We conclude with a remark on a relatively recent approach developed in Bayesian nonparametrics that is having considerable success in the literature and has partial analogies with the present approach. A research line initiated by MacEachern (1999; 2000) deals with the construction of time- or, more generally, covariate-dependent random discrete distributions and their implementation for inference in a Bayesian framework. See for example Hjort, Holmes, Müller and Walker (2010), Chapters 5, 7, 8, and Ghoshal and van der Vaart (2013), Chapter 14. Besides the clear analogy given by using nonparametric priors in a temporal framework, the inferential procedures elaborated following this approach substantially differ from the results we presented in Section 4. The foremost reason is that the former heavily rely on simulation techniques, such as Markov Chain Monte Carlo or Sequential Monte Carlo methods for computing posterior quantities of interest. Here instead we are primarily interested in the closed-form evaluation of the filtering distributions of the nonparametric signal.

Ackowledgements

The first author acknowledges financial support from the Spanish government via grant MTM2012-37195. The second author is supported by the European Research Council (ERC) through StG "N-BNP" 306406.

Appendix

Proof of Proposition 2.1

Before stating the result, we recall a useful lemma, whose proof can be found in Chaleyat-Maurel and Genon-Catalot (2009).

Lemma 5.1.

$$\sum_{j=0}^{l} \frac{(-1)^j}{(\lambda_n - \lambda_{n-1-j}) \prod_{0\le h\le l, h\neq j} |\lambda_{n-1-j} - \lambda_{n-1-h}|} = \frac{1}{\prod_{1\le h\le l+1}(\lambda_n - \lambda_{n-h})}.$$

Proof of Proposition 2.1. Consider first the one-dimensional case, i.e. $\mathbf{m} = m$, and denote for brevity $\theta[s,t] = \int_s^t \rho(\theta_u)\mathrm{d}u$, and define for $i \ge 1$

$$I_{1,\dots,i} = \int_0^t \cdots \int_{t_{i-1}}^t e^{-\lambda_m \theta[0,t_1]} \prod_{k=1}^{i-1} \theta_{t_k} e^{-\lambda_{m-k}\theta[t_k,t_{k+1}]}\mathrm{d}t_k \theta_{t_i} e^{-\lambda_{m-i}\theta[t_i,t]}\mathrm{d}t_i$$

$$I_{1,\dots,j-1,j,\dots,i} = \int_0^t \cdots \int_{t_{i-1}}^t e^{-\lambda_m \theta[0,t_1]} \prod_{k=1,k\neq j}^{i-1} \theta_{t_k} e^{-\lambda_{m-k}\theta[t_k,t_{k+1}]}\mathrm{d}t_k \theta_{t_i} e^{-\lambda_{m-i}\theta[t_i,t]}\mathrm{d}t_i$$

where $t_j := t_{j+1}$ in $I_{1,\dots,j-1,j+1,\dots,i}$. It can be easily seen that

$$I_i = -\frac{e^{-\lambda_m\theta[0,t]} - e^{-\lambda_{m-i}\theta[0,t]}}{\lambda_m - \lambda_{m-i}} \tag{33}$$

Then we have

$$(\prod_{h=0}^{i-1} \lambda_{m-h})^{-1} p_{m,m-i}(t) = I_{1,\dots,i}, \tag{34}$$

where $p_{m,m-i}(t)$ is the transition probability associated the one-dimensional death process. By integrating twice we obtain

$$\begin{aligned} I_{1,\dots,i} &= \frac{(-1)(I_{1,\dots,i-1} - I_{1,\dots,i-2,i})}{\lambda_{m-(i-1)} - \lambda_{m-i}} \\ &= \frac{(-1)^2}{\lambda_{m-(i-1)} - \lambda_{m-i}} \left[\frac{(I_{1,\dots,i-2} - I_{1,\dots,i-3,i-1})}{\lambda_{m-(i-2)} - \lambda_{m-(i-1)}} - \frac{(I_{1,\dots,i-2} - I_{1,\dots,i-3,i})}{\lambda_{m-(i-2)} - \lambda_{m-i}} \right]. \end{aligned}$$

The iteration of the successive integrations can be represented as a binary tree with root $(i,0) := I_{1,\dots,i}$, whose node $(i-j, i-k) := I_{1,\dots,i-j,i-k}$ branches into $(i-j,0) := I_{1,\dots,i-j}$ and $((i-j-1)^+, i-k) = I_{1,\dots,i-j-1,i-k}$, with both branches weighed $1/(\lambda_{m-(i-j)} - \lambda_{m-(i-k)})$, determined by the parent node's indices. The leaves correspond to nodes where the left coordinate touches zero if the right coordinate is already zero, or where the left crosses zero if the right coordinate

is positive. The term associated to the leaf $(0, i-k)$ will be $(-1)^i e^{-\lambda_{m-(i-k)}\theta[0,t]}$ weighed by some appropriate coefficient. The level before the leaves can be seen as the sequence

$$\underbrace{I_1I_2}_{2^1}\underbrace{I_1I_3}_{2^2}\underbrace{I_1I_2I_1I_4}_{2^3}\underbrace{I_1I_2I_1I_3I_1I_2I_1I_5}_{2^4}\ldots$$

where every sequence of 2^i terms is repeated with the last index augmented by one, and each I_i produces the leaves $e^{-\lambda_m}$ and $e^{-\lambda_{m-i}}$. Hence given i, there are 2^{i-2} terms I_1, 2^{i-3} terms I_2, $\ldots$, 2^1 terms I_{i-2}, 2^0 terms I_{i-1} and I_i. Note also that I_1 has 2^0 paths in common with I_i, 2^0 paths in common with I_{i-1}, 2^1 paths in common with I_{i-2}, $\ldots$, 2^{i-3} paths in common with I_2. The correct coefficient for I_k is computed by collecting some constants related to the paths that have the same last coefficient and simplifying. In particular, given i, the paths to be grouped for I_k are those whose constants change for indices greater than k change, since according to the rule above, when k is the rightmost index in ${}_{1,\ldots,k}$, there is only one path down to I_k. Hence, given i, term I_k has coefficient

$$\frac{(-1)^{i-1}}{\prod_{1\le h<k}(\lambda_{m-h}-\lambda_{m-k})} \times \sum_{j=0}^{i-(k+1)} \frac{1}{(\lambda_{m-k}-\lambda_{m-k-1-j})\prod_{0\le h\le i-(k+1),h\neq j}(\lambda_{m-k-1-j}-\lambda_{m-k-1-h})}.$$

By taking moduli and applying Lemma 5.1 to the sum above, we obtain

$$\frac{(-1)^{i-1}}{\prod_{1\le h\le i,h\neq k}(\lambda_{m-k}-\lambda_{m-h})}.$$

The result now follows from (33) and (34), and from the fact that in the K-dimensional case, the probability of going from $\mathbf{m}$ to $\mathbf{m}-\mathbf{i}$ is

$$\frac{\binom{|\mathbf{i}|}{i_1,\ldots,i_K}\binom{|\mathbf{m}|-|\mathbf{i}|}{m_1-i_1,\ldots,m_K-i_K}}{\binom{|\mathbf{m}|}{m_1,\ldots,m_K}}.$$

□

Proof of Proposition 4.3

The result follows by applying Proposition 4.2 and choosing a discrete mixing measure $H(\cdot) = \sum_{\mathbf{0}\le\mathbf{i}\le\mathbf{m}} w_{\mathbf{m}-\mathbf{i}}\delta_{\mathbf{i}}(\cdot)$ on $\mathcal{M}$ as in (4) restricted to vectors $\mathbf{0} \le \mathbf{i} \le \mathbf{m}$, a transition measure $\alpha_{\mathbf{m}-\mathbf{i}} = \alpha + \sum_{j=1}^{K_m}(m_j - i_j)\delta_{y_j}$, together with the fact that $H(\mathbf{i} \mid y) = \hat{w}_{\mathbf{m}+\mathbf{y}-\mathbf{i}}$, which can be deduced by Equation (5) in Antoniak (1974).

Proof of Proposition 4.4

Let $\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}}$ be as in $\bar{\mathscr{F}}_f$. If we show that (31) holds for any such distribution, then the statement follows for elements of $\bar{\mathscr{F}}_f$ from the fact that the locations $\mathbf{y}_m$ play no role in the mixture weights, and we have

$$\psi_t\Big(\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}}\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}}\Big) = \sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} \sum_{\mathbf{n}\in G(\mathbf{m})} p_{\mathbf{m},\mathbf{n}}(t)\Pi_{\alpha+\mathbf{n}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}} \tag{35}$$

$$= \sum_{\mathbf{n}\in G(\Lambda)} \Big(\sum_{\mathbf{m}\in\Lambda,\,\mathbf{m}\ge\mathbf{n}} w_{\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t)\Big)\Pi_{\alpha+\mathbf{n}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}}. \tag{36}$$

In order to show (31), use (30) to write it as

$$\sum_{i=0}^{m} C_{m,m-i}(t)\Big(\prod_{h=0}^{i-1}\lambda_{m-h}\Big) \sum_{\mathbf{0}\le\mathbf{i}\le\mathbf{m}:\,|\mathbf{i}|=i} \frac{\binom{i}{\mathbf{i}}\binom{m-i}{\mathbf{m}-\mathbf{i}}}{\binom{m}{\mathbf{m}}}\Pi_{\alpha+(\mathbf{m}-\mathbf{i})\cdot\boldsymbol{\delta}_{\mathbf{y}_m}} \tag{37}$$

and similarly for (32) with $\tilde{\mathbf{m}}$ and $h(\cdot,\tilde{\mathbf{m}})\pi_{\tilde{K}}$ in place of $\mathbf{m}$ and $\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}}$. Since $h(\cdot,\tilde{\mathbf{m}})\pi_{\tilde{K}}$ is the projection of $\Pi_{\alpha+\mathbf{m}\cdot\boldsymbol{\delta}_{\mathbf{y}_m}}$ by assumption, the right hand sides of (31) and (32) coincide if we show that the mixing distribution is the same. Hence we only have to show that the inner sums are consistent with respect to aggregation of sets. To this end, and without loss of generality, let $\mathbf{m}$ be given and fix $(i_1,\ldots,i_{K_m-2})$ and $i_{K_m-1}+i_{K_m}$ in $\mathbf{i}$. Then it suffices to show that the distinguished components corresponding to varying i_{K_m-1}, i_{K_m} have the same aggregate weights as the component of the mixture obtained by binning the last two classes, which corresponds to the configuration $(m_1,\ldots,m_{K_m-2},m_{K_m-1}+m_{K_m})$ and $(i_1,\ldots,i_{K_m-2},i_{K_m-1}+i_{K_m})$. Denote $\mathbf{m}' = (m_1,\ldots,m_{K_m-2})$ and $\mathbf{m}'' = (m_{K_m-1},m_{K_m})$ and similarly for $\mathbf{i}',\mathbf{i}''$, and let $a = m_{K_m-1}+$

m_{K_m} and $b = i_{K_m-1} + i_{K_m}$. Then the consistency condition is

$$\sum_{\mathbf{0}\leq\mathbf{i}''\leq\mathbf{m}'':|\mathbf{i}''|=b} \frac{\binom{i}{\mathbf{i}',\mathbf{i}''}\binom{m-i}{\mathbf{m}'-\mathbf{i}',\mathbf{m}''-\mathbf{i}''}}{\binom{m}{\mathbf{m}',\mathbf{m}''}} = \frac{\binom{i}{\mathbf{i}',\,b}\binom{m-i}{\mathbf{m}'-\mathbf{i}',a-b}}{\binom{m}{\mathbf{m}',a}}.$$

A direct simplification yields

$$\sum_{\substack{(0,0)\leq(i_{K_m-1},i_{K_m})\leq(m_{K_m-1},m_{K_m})\\ ii_{K_m-1}+i_{K_m}=b}} \frac{m_{K_m-1}!m_{K_m}!}{i_{K_m-1}!i_{K_m}!(m_{K_m-1}-i_{K_m-1})!(m_{K_m}-i_{K_m})!}$$

$$= \frac{(a!}{b!(a-b)!}.$$

Rearranging terms, the last expression can be written

$$\sum_{\substack{(0,0)\leq(i_{K_m-1},i_{K_m})\leq(m_{K_m-1},m_{K_m})\\ ii_{K_m-1}+i_{K_m}=b}} \binom{b}{i_{K_m-1}}\binom{a-b}{m_{K_m-1}-i_{K_m-1}} = \binom{a}{m_{K_m-1}}$$

which holds by Vandermonde's identity.

References

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.

Bernardo, J-M. and Smith, A.F. M. (1994). *Bayesian theory.* John Wiley & Sons Ltd, Chichester.

Cox, J.C., Ingersoll, J.E. and Ross, S.A. (1985). A Theory of the Term Structure of Interest Rates. *Econometrica* **53**, 385–407.

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in hidden Markov models.* Springer.

Chaleyat-Maurel, M. and Genon-Catalot, V. (2006). Computable infinite-dimensional filters with applications to discretized diffusion processes. *Stoch. Proc. Appl.* **116**, 1447–1467.

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

Dawson, D.A. (1993). *Measure-valued Markov processes.* Ecole d'Eté de Probabilités de Saint Flour XXI. Lecture Notes in Mathematics **1541**. Springer, Berlin.

DAWSON, D.A. (2009). *Stochastic population systems*. Lecture Notes of the PIMS-UBC Summer School in Probability.

ETHERIDGE, A.M. (2000). *An introduction to superprocesses*. University Lecture Series, 20. American Mathematical Society, Providence, RI.

ETHIER, S.N. (1981). A class of infinite-dimensional diffusions occurring in population genetics. *Indiana Univ. Math. J.* **30**, 925–935.

ETHIER, S.N. AND GRIFFITHS, R.C. (1993). The transition function of a Fleming-Viot process. *Ann. Probab.* **21**, 1571–1590.

ETHIER, S.N. AND KURTZ, T.G. (1981). The infinitely-many-neutral-alleles diffusion model. *Adv. Appl. Probab.* **13**, 429–452.

ETHIER, S.N. AND KURTZ, T.G. (1986). *Markov processes: characterization and convergence*. Wiley, New York.

ETHIER, S.N. AND KURTZ, T.G. (1993). Fleming-Viot processes in population genetics. *SIAM J. Control Optim.* **31**, 345–386.

FELLER, W. (1951) Diffusion processes in genetics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, 227–246.

FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

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.

GHOSHAL, S. AND VAN DER VAART, A. (2013). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, New York.

GRIFFITHS, R.C. (2006). Coalescent lineage distributions. *Adv. Appl. Prob.* **38**, 405–429.

GHOSH, J.K. AND RAMAMOORTHI, R.V. (2003). *Bayesian nonparametrics*. Springer-Verlag, New York.

HJORT, N.L., HOLMES, C.C., MLLER, P. AND WALKER, S.G., eds. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge Univ. Press.

HUTZENTHALER, M. AND WAKOLBINGER, A. (2007). Ergodic behavior of locally regulated branching populations. *Ann. Appl. Probab.* **17**, 474–501.

JANSEN, S. AND KURT, N. (2013). On the notion(s) of duality for Markov processes. arXiv:1210.7193 [math.PR].

JOHNSON, N.L., KOTZ, S. AND BALAKRISHNAN, N. (1994). *Continuous Univariate Distribu-*

tions. John Wiley & Sons, New York.

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

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

Kingman, J.F.C. (1982). The coalescent. *Stoch. Proc. Appl.* **13**, 235–248.

Korwar, R. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Statist.* **1** , 706–7011.

Lévy, P. (1948). *Processus stochastiques et mouvement Brownien*. Gauthier-Villars, Paris.

Liggett, T.M. (2005). *Interacting particle systems*. Springer-Verlag, Berlin. Reprint of the 1985 original.

MacEachern, S.N. (1999). Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statist. Assoc., Alexandria, VA.

MacEachern, S.N. (2000). Dependent Dirichlet processes. *Tech. Rep.*, Ohio State University.

Sen, A. and Balakrishnan, N. (1999). Convolutions of geometrics and a reliability problem. *Stat. Probab. Lett.* **43**, 421–426.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetic models. *Theor. Popul. Biol.* **26**, 119–164.