Bayesian inference for hidden Markov models via duality and approximate filtering distributions

Inferenza bayesiana per modelli di Markov nascosti via dualità e filtraggio approssimato

Guillaume Kon Kam King, Omiros Papaspiliopoulos and Matteo Ruggiero

Abstract Filtering hidden Markov models is analytically tractable only for a handful of models (e.g. Baum-Welch and Kalman filters). Recently, Papaspiliopoulos & Ruggiero (2014) proposed another analytical approach exploiting a duality relation between the hidden process and an auxiliary process, called dual and related to the time reversal of the former. With this approach, the filtering distributions are obtained as a recursive series of finite mixtures. Here, we study the computational effort required to implement this strategy in the case of two hidden Markov models, the Cox-Ingersoll-Ross process and the K-dimensional Wright-Fisher process, and examine several natural and very efficient approximation strategies.

Abstract *Il filtraggio dei modelli di Markov nascosti è un problema trattabile analiticamente solo per un numero limitato di modelli (p.es. i filtri di Baum-Welch e Kalman). Papaspiliopoulos & Ruggiero (2014) hanno proposto un nuovo approccio che sfrutta una relazione di dualità tra il processo nascosto ed un processo ausiliario, detto duale e legato alla reversibilità del segnale. Con questo approccio la soluzione del problema di filtraggio assume la forma di una mistura finita valutata mediante una ricorsione. Qui studieremo il costo computazionale richiesto per implementare tale strategia nel caso di due modelli di Markov nascosti, i cui segnali evolvono come un processo di Cox-Ingersoll-Ross e come una diffusione K-dimensionale di Wright-Fisher; analizzeremo diverse strategie di approssimazione.*

Key words: Optimal filtering, duality, Wright-Fisher, hidden Markov models

---

Guillaume Kon Kam King
University of Torino and Collegio Carlo Alberto, e-mail: guillaume.konkamking@unito.it

Omiros Papaspiliopoulos
ICREA - Pompeu Fabra University, Barcelona e-mail: omiros.papaspiliopoulos@upf.edu

Matteo Ruggiero,
University of Torino and Collegio Carlo Alberto, e-mail: matteo.ruggiero@unito.it

1 Introduction to optimal filtering using a dual process

Consider a hidden stochastic process and some noisy observations of this process. As new data arrives, obtaining the distribution for the last hidden state given all the values observed previously is called filtering the hidden process. Let the series $\{Y_k, 0 \geq k \geq n\}$ be the sequence of observations, denoted $Y_{0:n}$ for $Y \in \mathscr{Y}$, and let the Markov chain $\{X_k, 0 \geq k \geq n\}$, similarly denoted $X_{0:n}$, be the unobserved stochastic process. We assume $X_{0:n}$ to be the discrete-time sampling of a homogeneous continuous-time Markov process X_t. We also assume that X_t has state-space $\mathscr{X}$, transition kernel $P_t(x, dx')$ and initial distribution $\nu(\mathrm{d}x)$. The observations relate to the hidden signal by means of conditional distributions assumed to be given by the kernel $F(x, \mathrm{d}y)$ and we let $F(x, \mathrm{d}y) = f_x(y)\mu(\mathrm{d}y)$ for some measure $\mu(\mathrm{d}y)$. The filtering distributions, which are the target of inference, are $\mathscr{L}(X_n|Y_{0:n})$, denoted $\nu_n(\mathrm{d}x)$. Define now an update and prediction operator acting on probability measures ν:

$$\text{update:} \qquad \phi_y(\nu)(\mathrm{d}x) = \frac{f_x(y)\nu(\mathrm{d}x)}{p_\nu(y)}, \qquad \text{with } p_\nu(y) = \int_\chi f_x(y)\nu(\mathrm{d}x) \quad (1)$$

$$\text{prediction:} \qquad \psi_t(\nu)(\mathrm{d}x') = \int_\chi \nu(\mathrm{d}x) P_t(x, \mathrm{d}x') \quad (2)$$

Then, the filtering distributions can be obtained by repeated applications of the update and prediction operators, as the recursion: $\nu_0 = \phi_{Y_0}(\nu)$ and $\forall n > 0, \nu_n = \phi_{Y_n}(\psi_{t_n - t_{t-1}}(\nu_{n-1}))$ (see for instance [1]). An explicit solution to the filtering problem is seldom available, except in two notorious cases: unobserved Markov chains with a discrete state-space, and Gaussian unobserved Markov chains with Gaussian conditional distribution. [2] extended the class of models for which an explicit solution is available by exploiting a duality relation between the unobserved Markov chain and a pure death stochastic process. In order to describe this, assume that Θ_t is a deterministic process and that $r : \Theta \to \Theta$ is such that the differential equation: $\mathrm{d}\Theta_t/\mathrm{d}t = r(\Theta_t)$ with $\Theta_0 = \theta_0$ 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 $\mathscr{M} \times \Theta$, where Θ_t evolves autonomously according to the previous differential equation, 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 $\lambda(|\mathbf{m}|)\rho(\theta)m_j$. We say that (M_t, Θ_t) is *dual* to X_t with respect to a family of functions h, e.g.

$$\mathbb{E}^x\left[h(X_t, \mathbf{m}, \theta)\right] = \mathbb{E}^{\mathbf{m},\theta}\left[h(x, M_t, \Theta_t)\right], \quad \forall x \in \chi, \mathbf{m} \in \mathscr{M}, \theta \in \Theta, t \geq 0.$$

where $\mathbb{E}^x[f(X_t)] = \mathbb{E}[f(X_t)|X_0 = x] = \int_\chi f(x')P_t(x, \mathrm{d}x')$ and the duality functions are such that $h : \chi \times \mathscr{M} \times \Theta \to \mathbb{R}_+$, $\Theta \subseteq \mathbb{R}^l$. The dual process (M_t, Θ_t) is separated into a deterministic component Θ_t and a pure death process M_t, whose rates are subordinated to the deterministic process. The transition probabilities of the dual

process are denoted $p_{\mathbf{m},\mathbf{n}}(t,\theta) = \mathbb{P}\left[M_t = \mathbf{n} | M_0 = \mathbf{m}, \Theta_0 = \theta\right], \forall\, \mathbf{n},\mathbf{m} \in \mathscr{M}^2, \mathbf{n} \leq \mathbf{m}$. The duality property is key to the computability of the filters, as it allows to replace the expectation with respect to realisations of the original stochastic process in the prediction operation (Eq. (2)) by an expectation over realisations of the pure death component of the dual process, which involves finite sums.

The transition probabilities can be found by exploiting the duality relation ([2]):

$$p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t,\theta) = \gamma_{|\mathbf{m}|,|\mathbf{i}|} C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t) p(\mathbf{i};\mathbf{m},|\mathbf{i}|) \tag{3}$$

with:

$$\gamma_{|\mathbf{m}|,|\mathbf{i}|} = \left(\prod_{h=0}^{|\mathbf{i}|-1} \lambda_{|\mathbf{m}|-h}\right), \text{ and } C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t) = (-1)^{|\mathbf{i}|} \sum_{k=0}^{|\mathbf{i}|} \frac{e^{-\lambda_{|\mathbf{m}|-k} t}}{\prod_{0\leq h\leq |\mathbf{i}|, h\neq k} (\lambda_{|\mathbf{m}|-k} - \lambda_{|\mathbf{m}|-h})} \tag{4}$$

and $p(\mathbf{i};\mathbf{m},|\mathbf{i}|)$ is the hypergeometric probability mass function.

We also define the following notion of *conjugacy*, by assuming that $\mathscr{F}_0 = \{h(x,\mathbf{m},\theta)\pi(\mathrm{d}x), \mathbf{m} \in \mathscr{M}, \theta \in \Theta\}$ is a family of probability measures such that there exist functions $t : \mathscr{Y} \times \mathscr{M} \to \mathscr{M}$ and $T : \mathscr{Y} \times \Theta \to \Theta$ with $\mathbf{m} \to t(y,\mathbf{m})$ increasing and such that $\phi_y(h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)) = h(x,t(y,\mathbf{m}),T(y,\theta))\pi(\mathrm{d}x)$. The filtering algorithm proposed in [2] can be summarised by the two following relations. For the family of finite mixtures $\bar{\mathscr{F}}\{\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x) : \Lambda \subset \mathscr{M}, |\Lambda| < \infty, \sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} = 1\}$, the update operation acts as:

$$\phi_y\left(\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)\right) = \sum_{\mathbf{n}\in t(y,\Lambda)} \hat{w}_{\mathbf{m}} h\left(x,\mathbf{n},T(y,\theta)\right)\pi(\mathrm{d}x) \tag{5}$$

with $t(y,\Lambda) = \{\mathbf{n} : \mathbf{n} = t(y,\mathbf{m}), \mathbf{m} \in \Lambda\}$, and $\hat{w}_{\mathbf{m}} \propto w_{\mathbf{m}}$ and for $\mathbf{n} = t(y,\mathbf{m})$, $\sum_{\mathbf{n}\in t(y,\Lambda)} \hat{w}_{\mathbf{n}} = 1$. This updates the signal given the new data by means of the Bayes theorem. The prediction operation acts as:

$$\psi_t\left(\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)\right) = \sum_{\mathbf{n}\in G(\Lambda)} \left(\sum_{\mathbf{m}\in\Lambda,\mathbf{m}\geq\mathbf{n}} w_{\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t,\theta)\right) h\left(x,\mathbf{n},\theta_t\right)\pi(\mathrm{d}x) \tag{6}$$

where $G(\Lambda) = \{\mathbf{n} \in \mathscr{M} : \mathbf{n} \leq \mathbf{m}, \mathbf{m} \in \Lambda\}$, propagating the current filtering distribution by means of the signal transition kernel. As such, filtering a hidden Markov model using the duality relation consists in recursive operations on finite mixtures of distributions, where the number of components remains finite and the components remain within the same family of distributions. At each new observation, the mixture distribution is shifted towards the data, then until the next observation, the mixture progressively forgets the past information and drifts back towards the prior distribution.

2 Implementation of the dual filtering algorithm

The filtering algorithm resulting from the method presented above is similar to the Baum-Welch filter and it alternates update and prediction steps. The update step shifts each component and modifies its weight, while the prediction step lets all the components propagate some of their mass towards the components close to the prior. We illustrate this dual filtering algorithm (Eq. (1)) for two stochastic processes: the Cox-Ingersoll-Ross (CIR) process and the Wright-Fisher (WF), presented in full details later. For these two models, the number of mixture components in the filtering distributions evolves as $|\Lambda_n| = \prod_{i=1}^{K}(m_{0,i} + 1 + \sum_{i=1}^{n} Y_i)$, where K is the dimension of the latent space and $\mathbf{m}_0 = m_{0,1:K}$ is the initial state.

The prediction step is much costlier than the update step, as at each iteration it involves computing the transitions from all elements of Λ_i to those reachable by a pure death process in $G(\Lambda_i)$. It is possible to contain the cost of the prediction operation by storing the transition terms $p_{\mathbf{m},\mathbf{n}}$, which will be used multiple times during the successive iterations. However, the rapid growth in the number of those terms (proportional to $|G(\Lambda_n)|^2$) does not permit saving all of them in memory. Yet, the $p_{\mathbf{m},\mathbf{n}}$ are themselves a product of a number of terms which grows only quadratically with the sum of all observations and can be saved (Eq. (4) and the hypergeometric coefficients expressed as a product of binomials coefficients). Another technical difficulty is that the sum with terms of alternated sign in (4) is susceptible to both over and underflow. We compute it using the `Nemo` library for arbitrary precision computation ([3]).

Although considerable efficiency gains are achieved by storing the transition terms, further improvements may be obtained by a natural approximation of the filtering distributions. Indeed, the filtering distributions contain a number of components that grows quickly as new observations arrive, but the complexity of the hidden signal does not necessarily increase accordingly. Hence, if the prior is reasonable and the posteriors appropriately concentrated, there is no reason for the number of components with non negligible weight to explode. Indeed, simulation studies show that the number of components representing 99% of the weight of the mixture saturates as new observations arrive (Eq. (2)). This suggests that some components may be deleted from the mixtures, speeding the computations, without loosing much in terms of precision. We envision three strategies for pruning the mixtures:

- prune all the components who have a weight below a certain threshold, which is an attempt at controlling the approximation error at a given step. This approach will be referred to as the *fixed threshold strategy*.
- retain only a given number of components, hopefully chosen above the saturation number (see Eq. (2)). This is an attempt at controlling the computation budget at each time step. This approach will be referred to as the *fixed number strategy*.
- retain all the largest components needed to reach a certain amount of mass, for instance 99%. This is an adaptive strategy to keep the smallest possible

number of components under a certain error tolerance level. This approach will be referred to as the *fixed fraction strategy*.

In Algorithm 1, the pruning is performed just after the update step. This choice is dictated by two reasons: first, after the update step the mixture should be more concentrated because information from the new observation was just incorporated, leading to a smaller number of components with non negligible weight. Then, as the prediction step is the most computationally expensive, reducing the number of components before predicting entails the maximum computational gain. After pruning, we renormalise all the remaining weights so that they sum to 1. As the pruning operation occurs at each time step, the level of approximation on a given filtering distribution results from several successive approximations.

Algorithm 1: Optimal filtering algorithm using the dual process, with the option of pruning.

Data: $Y_{0:n}$, $t_{0:n}$ and $\nu = h(x, \mathbf{m}_0, \theta_0) \in \mathscr{F}$ for some $\mathbf{m}_0 \in \mathscr{M}$, $\theta_0 \in \Theta$
Result: $\Theta_{0:n}$, $\Lambda_{0:n}$ and $W_{0:n}$ with $W_i = \{w^i_{\mathbf{m}}, \mathbf{m} \in \Lambda_i\}$
Initialise
  Set $\Theta_0 = \theta_0$
  Set $\Lambda_0 = \{t(Y_0, \mathbf{m}_0)\} = \{m^*\}$ and $W_0 = \{1\}$ with t as in (5)
  Let Θ_0 evolve during $t_1 - t_0$ and set θ^* equal to the new value
  Set $\Lambda^* = G(\Lambda_0)$ and $W^* = \{p_{m^*,\mathbf{n}}(t_1 - t_0, \theta_0), \mathbf{n} \in \Lambda^*\}$ with G as in (6) and $p_{\mathbf{m},\mathbf{n}}$ as in (3)
for *i from* 1 *to n* **do**
  Update
    Set $\Theta_i = \theta^*$
    Set $\Lambda_i = \{t(Y_i, \mathbf{m}), \mathbf{m} \in \Lambda^*\}$
    Set $W_i = \{\frac{w^*_{\mathbf{m}} p_{h(x,\mathbf{m},\Theta_i)}}{\sum_{\mathbf{n}\in\Lambda^*} w^*_{\mathbf{n}} p_{h(x,\mathbf{n},\Theta_i)}}, \mathbf{m} \in \Lambda^*\}$ with $p_{h(x,\mathbf{m},\theta)}$ defined as in (1)
  if *pruning* **then**
    Prune(Λ_i) and remove the corresponding weights in W_i
    Normalise the weights in W_i
  Predict
    Let Θ_i evolve during $t_{i+1} - t_i$ and set θ^* equal to the new value
    Set $\Lambda^* = G(\Lambda_i)$ and $W^* = \left\{ \sum_{\mathbf{m}\in\Lambda_i, \mathbf{m}\geq\mathbf{n}} w^i_{\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t_{i+1} - t_i, \Theta_i), \mathbf{n} \in \Lambda^* \right\}$
end

3 Filtering two stochastic processes

For illustration we consider two stochastic processes, a 1-dimensional Cox-Ingersoll-Ross process and a 3-dimensional Wright-Fisher process, which we filter using the strategy outlined above. The dimension of the state space of the pure death process is dependent on the dimension of the signal, therefore the number of components in the

filtering distributions for the WF process is much greater than for the CIR process, rendering the inference computationally more challenging. The one-dimensional CIR process has the following generator: $\mathscr{A} = (\delta\sigma^2 - 2\gamma x)\frac{d}{dx} + 2\sigma^2 x\frac{d^2}{dx^2}$. with $\delta, \gamma, \sigma > 0$ and stationary distribution $\mathrm{Ga}(\delta/2, \gamma/\sigma^2)$. A conjugate emission density is: $Y_t|X_t \sim \mathrm{Po}(X_t)$. The duality function can be found in [2]. We simulate a CIR process starting from $X = 3$ with $\delta = 3.6$, $\gamma = 2.08$, $\sigma = 2.8$. which corresponds to a stationary distribution $\mathrm{Gamma}(1.8, 0.38)$. Furthermore, we simulate 10 observations at each time, with 200 time steps separated by 0.011 seconds. For the inference, we use as a prior for the stationary distribution a $\mathrm{Gamma}(1.5, 0.15625)$ which corresponds to $\gamma = 2.5$, $\delta = 3.$, $\sigma = 4.$ and $m_0 = 0$.

The Wright-Fisher model is a K-dimensional diffusion, whose generator is $\mathscr{A} = \frac{1}{2}\sum_{i=1}^{K}(\alpha_i - |\alpha|x_j)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{K} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j}$. and its stationary distribution is a Dirichlet(α). A conjugate emission density is: $f_x(Y) = \prod_{i=1}^{J}\left(|\mathbf{n}_i|!\prod_{k=1}^{K}\frac{x_k^{n_{ki}}}{n_{ki}!}\right)$. The duality function can also be found in [2]. We simulate two datasets using a discrete time and finite population Wright-Fisher model of dimension $K = 3$ initialised at random from a Dirichlet(0.3, 0.3, 0.3) with $\alpha = (0.75, 0.75, 0.75)$ and a population size of 50000. 15 observations are collected at each observation time. There are 10 observation times with a time step of 0.1 second for the first dataset and 20 observation times with a time step of 0.004 second for the second dataset. As a prior, we use a uniform distribution Dirichlet(1,1,1). The two different time steps for the WF model are intended to explore two regimes, one for which the time between observations is large, such that information from previous data is almost forgotten (the predictive distribution has almost moved back to the prior) and one for which that time is very short. In these two regimes, the number of components with non negligible weights is expected to be very different. Notably, in the second regime the impact of the successive approximations is expected to be stronger. Fig. 1 shows that in all the studied cases, the filtering distributions are centred around the signal. For the WF model with the short time step, the filtering distributions do not evolve fast enough to follow the signal exactly, but this is to be expected given the rapid rate at which new observations arrive. Considering how the weights are distributed among the components of the filtering distribution, we observe that the mass is mostly concentrated on a number of components several orders of magnitude smaller than the total number of components. This observation suggests that many components may be deleted with a minimal loss of precision.

To quantify this loss of precision by pruning, we compute the Hellinger distance between the exact and the approximate filtering distributions obtained by pruning: $d_H(f_1, f_2) = \frac{1}{2}\int_\chi(\sqrt{f_1} - \sqrt{f_2})^2$. As there is one filtering distribution per observation time, to compare two sets of filtering distributions we consider the maximum over time of the distance between the distributions at each time, i.e. $\sup_n(d_H(\nu_{n,\mathrm{exact}}, \nu_{n,\mathrm{approx}}))$. The numerical evaluation of the distances is done using standard quadrature and simplicial cubature rules from the R package `SimplicialCubature`. Parallel to the loss of precision due to the approximation, we consider the gain in efficiency by measuring the computing time needed

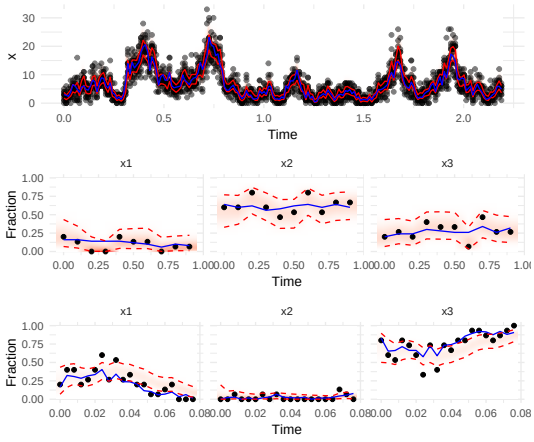

Fig. 1: Hidden signal, data and 95% credible intervals of the filtering distribution for the three datasets. The hidden signal is denoted by the blue line, the data by the black dots and the credible bands by the red dashed lines. Top: CIR, centre: WF, bottom: WF with short time step. For the WF model, each panel corresponds to one marginal, and the data plotted is the proportion of the 15 multinomial observations which are from the corresponding type.

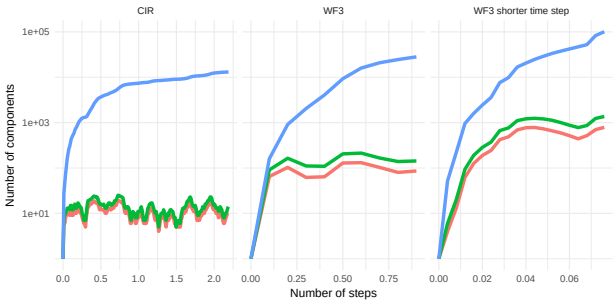

Fig. 2: Number of components (in log scale) in the filtering distributions as a function of the iteration number. Left: CIR, centre: WF, right: WF with short time step. The blue line denotes the total number of components in the filtering distributions, the green line denotes the number of components carrying 99% of the mass and the red 95%.

to filter the whole dataset. Fig. 3 shows that the approximation strategies afford a reduction in computing time by 5 orders of magnitude for the CIR process, or by 2 to 3 orders of magnitude for the three-dimensional WF process. The fixed fraction strategy is noticeably slower in the case of the WF process with the shorter time step because the mass is spread over more components, as was also apparent on Fig. 2. For all strategies and all processes, it seems possible to find a compromise between accuracy and computing time where increasing the computational effort starts yielding diminishing returns. Except in the case of the CIR model where the fixed threshold strategy seems to slightly outperform the others, no strategy seems to offer a fundamentally better precision/cost ratio than the others.

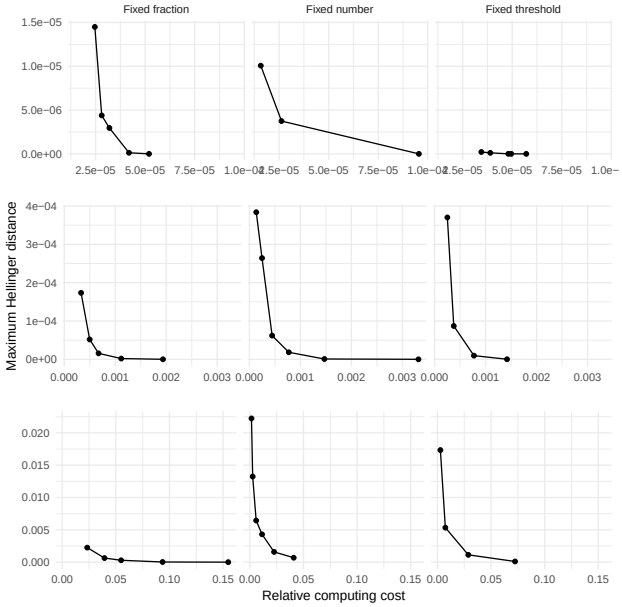

Fig. 3: Approximation error versus computational effort. The computation time is given relative to the time needed for obtaining the exact filtering distributions. The top level represents the CIR process, the middle represents the WF process and the bottom represents the WF process with the shorter time step. Fixed fractions tested are 0.8, 0.9, 0.95, 0.99, 0.999. The fixed numbers tested are 5, 10, 25 for the CIR process, 10, 25, 50, 100, 200, 400 for the WF processes. The fixed thresholds are 0.01, 0.005, 0.001, 0.0005, 0.0001 for the CIR orocess and 0.01, 0.005, 0.001, 0.0001 for the WF processes.

The results presented here are a preliminary study on the computational costs of filtering strategies based on duality. A more thorough investigation of these and other aspects involved in this type of filtering are currently ongoing work.

References

1. Cappé, O., Moulines, E., Ryden, T.: Inference in Hidden Markov Model. Springer, New York, USA (2005)
2. Papaspiliopoulos, O., Ruggiero, M.:Optimal filtering and the dual process. Bernoulli **20**, 1999–2019 (2014)
3. Fieker, C., Hart, W., Hofmann, T., Johansson, F.: Nemo/Hecke: Computer Algebra and Number Theory Packages for the Julia Programming Language. In: Proc. 2017 ACM Int. Symp. Symb. Algebr. Comput., pp. 157–164. ACM, New York, USA (2017)