Author/s:
Pauley, M;Mclean, C;Manton, JH

Title:
A numerical filtering method for linear state-space models with Markov switching

Date:
2020-07

Citation:
Pauley, M., Mclean, C. & Manton, J. H. (2020). A numerical filtering method for linear state-space models with Markov switching. International Journal of Adaptive Control and Signal Processing, 34 (7), pp.813-838. https://doi.org/10.1002/acs.3109.

Persistent Link:
https://hdl.handle.net/11343/275578

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

A numerical filtering method for linear state-space models with Markov switching†

Michael Pauley* | Christopher Mclean | Jonathan H. Manton

[1]Department of Electrical and Electronic Engineering, The University of Melbourne, Victoria, Australia

Correspondence
*Michael Pauley, Corresponding address.
Email: mpauley85@gmail.com

Abstract

A class of discrete-time random processes arising in engineering and econometrics applications consists of a linear state-space model whose parameters are modulated by the state of a finite-state Markov chain. Typical filtering approaches are *collapsing* methods, which approximate filtered distributions by mixtures of Gaussians, each Gaussian corresponding to one possibility of the recent history of the Markov chain, and particle methods. This paper presents an alternative approach to filtering these processes based on keeping track of the values of the underlying filtered density and its characteristic function on grids. We prove that it has favourable convergence properties under certain assumptions. On the other hand, as a grid method, it suffers from the curse of dimensionality, and so is only suitable for low-dimensional systems. We compare our method to collapsing filters and a particle filter with examples, and find that it can outperform them on 1- and 2-dimensional problems, but loses its speed advantage on 3-dimensional systems. Meanwhile, our method has a proven theoretical convergence rate that is probably not achieved by collapsing and particle methods.

KEYWORDS:
Markov switching, filtering, linear state-space models.

1 | INTRODUCTION

In filtering problems, optimal filters sometimes cannot be computed exactly or are too computationally expensive and it becomes necessary to approximate the filtered distribution in some way. This paper is concerned with a variant of the linear state-space model where the parameters are modulated by the state of a finite-state Markov chain. A common approach is to use a sum of a small number of Gaussians. In recent years, particle filters have also been applied to this class of problems. In this paper we present an alternative approach, and provide theory and examples to demonstrate the benefits.

The following simple example will help explain the goals. Let $S(k), k \in \mathbb{N}$, be a Markov chain on the set $\{0, 1\}$ with transition matrix

$$\begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix} \tag{1}$$

and initial distribution (i.e., distribution of $S(0)$) the discrete uniform distribution on $\{0, 1\}$. Let $Z(k), k \in \mathbb{N}$ be iid $N(0, 1)$ random variables. Let $X(0) \sim N(0, 1)$ and let $X(k) = 0.9X(k-1) + 0.25 + 0.1Z(k)$ when $S(k) = 0$ and $X(k) = 0.8X(k-1) - 0.5 + 0.1Z(k)$ when $S(k) = 1$. How can we approximate the distribution of $X(100)$? One idea is to use a "collapsing" method:

†Research supported by ARC Linkage Project LP140100473. A preliminary report of the work in this paper previously appeared in [1].

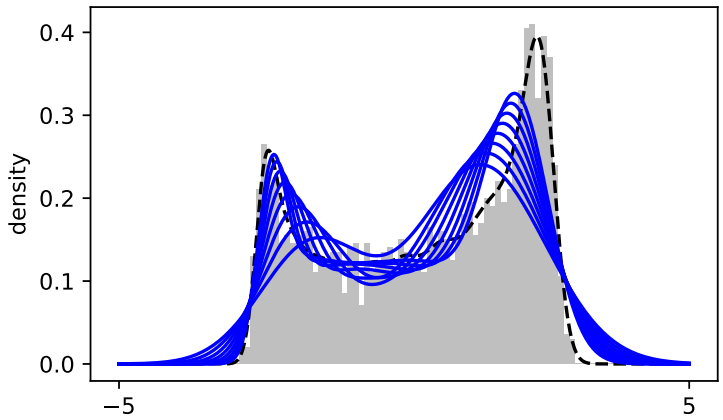


FIGURE 1 Example for Section 1. Solid blue curves: approximation of the density of $X(100)$ using sums of Gaussians, for depth parameter varying from 1 to 8. Dashed black curve: a much closer approximation of the density, using the method of our paper. Grey bars: histogram of the result from 2000 simulations. The exact density has too many terms to realistically compute.

choose a positive integer D and approximate the distribution of $X(k)$ conditional on $S(k-D+1), \cdots, S(k)$ by a Gaussian. The distribution of $X(k)$ is then approximated by a sum of 2^D Gaussians. The prediction step of a Kalman filter can then be used to approximate the distribution of $X(k+1)$ as a sum of 2^{D+1} Gaussians which can then be approximated by a sum of 2^D Gaussians. Repeating this procedure we can approximate the distribution of $X(100)$. The result of this procedure is shown for the example above in the blue curves of Fig. 1 for values of D ranging from 1 to 8. Increasing D results in a better approximation, but at exponentially increasing cost. The exact density has too many terms to compute in a reasonable amount of time, but a much closer approximation is given by the dashed black curve of Fig. 1 which is computed using the prediction step of the filtering method described in this paper. Computation times for the sum of Gaussians vary from 0.006s (D=1) to 0.292s (D=8) while the computation time for our method is 0.052s. (It should be noted that our implementations of both methods were not designed for speed.)

In general, this paper looks at a class of probabilistic models described as follows. Let $\mathbb{N}$ denote the non-negative integers. Let $S(k), k \in \mathbb{N}$, be a Markov chain on a finite set $\mathcal{S}$ of states having transition matrix $M : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. Let X and Y be random processes on $\mathbb{N}$, taking values in $\mathbb{R}^d$ and $\mathbb{R}^n$ respectively, satisfying the equations

$$\begin{aligned} X(k) &= A_{S(k)}X(k-1) + B_{S(k)}U(k) + C^{\text{proc}}_{S(k)}Z_{\text{proc}}(k),\ k \geq 1 \\ Y(k) &= F_{S(k)}X(k) + G_{S(k)}U(k) + C^{\text{obs}}_{S(k)}Z_{\text{obs}}(k),\ k \geq 0 \end{aligned} \tag{2}$$

where $Z_{\text{proc}}(k) \in \mathbb{R}^c$ and $Z_{\text{obs}}(k) \in \mathbb{R}^m$ are iid $N(0, I)$, the $U(k)$ take values in $\mathbb{R}^b$ that are known by time k,[1] and, for each $s \in \mathcal{S}$, A_s is a $d \times d$ matrix, B_s is a $d \times b$ matrix, C^{proc}_s is a $d \times c$ matrix, F_s is an $n \times d$ matrix, G_s is an $n \times b$ matrix and C^{obs}_s is an $n \times m$ matrix. An extreme case of (2) occurs when the C^{obs}_s and C^{proc}_s are all zero; then the distribution of $X(k)$ given $Y(0), \cdots, Y(k)$, is discrete. In this paper we are interested in the opposite extreme, where the parameters are such that the filtered distribution always has a density. In Section 2 we give the details of our assumptions regarding the parameters.

Equation (2) does not describe the distributions of the initial values $S(0)$ and $X(0)$. There are several options. One is to choose an arbitrary distribution for $S(0)$ and an arbitrary nondegenerate Gaussian for $X(0)$. Another option, available under certain conditions on the parameters and U, is to assume that the model is in steady state. Equation (2) is a Hidden Markov Model[3], in which $(S(k), X(k))$ is the hidden part of the state and $Y(k)$ is the observation. This model unifies several notions of linear models modulated by a Markov chain.

Equation (2) includes vanilla finite-state Markov chains and vanilla linear state space equations as degenerate cases. The optimal filters for these models can be written down exactly[3] Chapter 5. There do exist other special cases of (2) that still have

[1] In this paper we take $U(k)$ to be deterministic. For likelihood calculations one could more generally assume that the $U(k)$ are weakly exogenous for the parameters of interest. [2]

exact optimal filters. These include regime-switching models in econometrics[4,5], which have been pointed out to be special cases of (2).[6] However, in many other cases of (2) of practical interest, efficient exact methods are unlikely to be found, so suboptimal filters, i.e., numerical approximations, become useful.

Models of the form (2) have been applied in the following scenarios.

- *Machines with failures.* The parameters for one state describe the state dynamics of a machine in normal operation. One or more other states describe the dynamics with failed components. The probability of failure per time step can be encoded in the transition matrix of S. Filtering can then be used to infer when failure has occurred.[7]
- *Economic regime switching.* In this application the observations are economic data such as GDP, and the model can be used to classify time periods as either recession or boom.[6]
- *Approximation of nonlinear models or models with non-Gaussian observation noise.* A nonlinear model might be approximated by a piecewise linear model which in turn can be approximated by the Markov switching model. The dynamics then have a "coarse" component S and a "fine" component X. Non-Gaussian observation noise can be approximated by a mixture of Gaussian components and the Markov-switching model can be used to choose between the components.[8,7]

A timeline of the suboptimal filtering methods for this model and special cases is given in [9, Section 13.3.5]. A common theme of many filters is *collapsing*:[8] choose a *depth* D and approximate the filtered density of $X(k-1)$ by a mixture of Gaussians indexed by the $|\mathcal{S}|^D$ tuples $(s_1, \cdots, s_D)$ of states. The mixture weighting for a given tuple is interpreted as an approximation of $P(S(k-1) = s_1, \cdots, S(k-D) = s_D | Y(0), \cdots, Y(k))$ and the corresponding Gaussian is interpreted as an approximation of the probability density of $X(k-1)$ conditional on $S(k-1) = s_1, \cdots, S(k-D) = s_D, Y(0), \cdots, Y(k)$. Applying the prediction step exactly to this approximation multiplies the number of terms by $|\mathcal{S}|$. To prevent an explosion in the number of terms, the Gaussians are later merged together so that the filtered density of $X(k)$ is again approximated by a mixture of $|\mathcal{S}|^D$ terms. A collapsing method for (2) in its full generality was described in [7, Section 3.1] in an engineering context (building on older results for special cases), and in [6] in econometrics (also building on older results, and only for a depth of 1). As the depth increases, the estimates produced by collapsing methods converge to the true filtered density, but the computational cost is exponential in the depth. Other filtering algorithms build on the collapsing idea by adaptively both merging and pruning histories.[10] An alternative to these collapsing methods is to use particle filters, which have been proposed for Markov modulated linear and non-linear systems.[11,12,13,14,15]

The main contribution of the present paper is an alternative numerical method for filtering for a large class of the models given by (2). A central question in suboptimal filtering is how to numerically represent a distribution. The aim is to represent the sorts of distributions that can arise in the particular filtering problem, with a high degree of accuracy. The *collapsing* answer to this question is to represent distributions as a mixture of Gaussians. Our representation is a tuple $(\mathcal{G}, \tilde{\mathcal{G}}, h, \tilde{h})$ where $\mathcal{G}$ and $\tilde{\mathcal{G}}$ are grids of points in $\mathbb{R}^d$, $h : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ is a discretisation of the filtered probability density function (PDF), and $\tilde{h} : \mathcal{S} \times \tilde{\mathcal{G}} \to \mathbb{R}$ is a discretisation of the corresponding characteristic function (CF). The advantage of our approach is that the error between this discretisation and the true filtered density converges to zero exponentially as the number of grid points in each dimension increases. This convergence property is stated precisely in Theorem 2. In a quite different way, [16] has previously shown the usefulness of considering the CF in a suboptimal filter.

General-purpose grid-based methods have been developed for other nonlinear filtering problems[17,18,19]. The model (2) is different from the usual nonlinear model. Firstly, the state has a discrete component, which needs to be accounted for by keeping track of several "layers" of the grid. Secondly, conditional on the discrete part of the state, the continuous part behaves according to the linear state-space model; it should be possible to take advantage of this simple behaviour.

Our approach is similar in concept to [20] which also filters this class of models using grid representations of the PDF and CF. The present paper has several significant new aspects: firstly, we provide conditions under which the method works in higher dimensions, and prove asymptotic properties of the approximation errors as the grid size increases to infinity. Next, at one point it is necessary to approximate the CF at non-grid points. The authors of [20] use a Fast Fourier Transform and grid interpolation. We point out how in certain circumstances (including the 1-dimensional case) a Chirp Z-Transform can be used. When this option is not available we use a (non-fast) approximation of the Fourier Transform. Either way we can prove asymptotic properties of the errors as the grid size increases; it is not clear whether these properties can be achieved with a Fast Fourier Transform and grid interpolation.

Outline. Section 2 describes the class of problems we consider. We give an example that shows how a grid-based approximation can fail for some choices of the parameters in (2). After this we place additional constraints on the problem. In Section 3 we

describe our method in detail and give Theorem 2 which describes the asymptotics of the error introduced by this method. In Section 4 we provide demonstrations of our method including a comparison to a collapsing method. We also investigate our method's ability to approximate a steady state distribution, and make some comments on efficiency of implementations.

There are two significant weaknesses to our method. The first is that the assumptions we make in Section 2 restrict the models to those that behave stably. The second is the requirement to do non-fast Fourier Transform calculations. Section 5 discusses a possible direction for alleviating these weaknesses. By keeping track of transformed versions of the random variables (or equivalently: allowing for moving grids), we can weaken the assumptions about the system parameters, and also ensure that a Chirp Z-Transform can be used in all circumstances. However, this comes with additional overheads, and it seems that further innovations are necessary to achieve useful performance with this extension.

In Section 6 we discuss the advantages and disadvantages of our approach and the outlook. The appendices give experimental parameters, and proofs of asymptotic properties of the filtered distribution that are required for the proof of Theorem 2. Appendix A provides the experimental parameters for our numerical results. Appendix B proves the asymptotic properties of the distributions needed for our error bounds.

2 | ASSUMPTIONS AND JUSTIFICATION

The filtering method we describe later in the paper is not designed for all possible parameters in (2). The assumptions (Assumptions 1 to 4 below) on the parameters of (2) are designed so that the distribution of $X(k)$ does not spread out too much over time (Assumption 1);[2] the filtered distribution of $X(k)$ stays absolutely continuous with respect to the Lebesgue measure (Assumption 2); the filtered characteristic function of $X(k)$ does not spread out too much over time (Assumption 3); and the distribution of $X(0)$ is well localised in space as well as frequency (Assumption 4). Appendix B shows how Assumptions 1 to 4 lead to these useful features.

The class of models satisfying the assumptions is quite large, but does not include models where the process noise and observation noise do not "fill all the dimensions" in a certain way. Thus we exclude, for example, models of [4, 5]. Let us see an example where our method is unsuitable. Suppose S can take the states 0 and 1 and its transition matrix is $\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$. Let $\lambda \in (0,1)$ and in (2) with $X(k), Y(k)$ being one-dimensional let $A_s = \lambda, B_s = 2s-1, C_s^{\text{proc}} = 0, F_s = 1, G_s = 0, C_s^{\text{obs}} = 1$ for all s and $U(k) = 1$ for all k, so that $X(k) = \lambda X(k-1) + (2S(k)-1)$.[3] Suppose we take X to have a known (deterministic) initial value $X(0)$, and $Y(k) = X(k) + Z(k)$ where the $Z(k)$ are iid $N(0,1)$. Now the probability distribution of $(S(k), X(k))$ conditional on $Y(0), \cdots, Y(T)$ is a mixture of up to 2^T delta distributions, so that treating it as a continuous distribution and recording the values on a grid will be ineffective. The assumptions below avoid this problem. It is not enough to assume that C_s^{obs} and C_s^{proc} are nonzero: one could construct an example of (2) with $C_s^{\text{obs}} \neq 0 \neq C_s^{\text{proc}}$, that decomposes into an "orthogonal sum" of the above model and some other model; our grid representation would again be ineffective.

On the other hand, Assumptions 1 to 4 below ensure that, for all k, the probability distribution of $X(k)$ conditional on $S(k)$ and $Y(0), \cdots, Y(k)$ has a PDF, and this PDF can be approximated well by a structure with a small amount of data.

2.1 | Model description and assumptions

Model description.

1. S is a Markov chain on $\mathbb{N}$, taking values in a finite set $\mathcal{S}$ and having transition matrix $M : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$.
2. X, Y are random processes on $\mathbb{N}$, taking values in $\mathbb{R}^d, \mathbb{R}^n$ respectively, and satisfying (2). The $Z_{\text{proc}}(k)$ and $Z_{\text{obs}}(k)$ are all independent, distributed as $N(0, I)$ and independent of $(S(0), S(1), \cdots)$. We interpret $(S(k), X(k))$ as the underlying, unknown state at time k and $Y(k)$ as the observation.
3. The matrix parameters $A_s, B_s, C_s^{\text{proc}}, F_s, G_s, C_s^{\text{obs}}$ are specified in advance for each $s \in \mathcal{S}$. The $U(k)$ are taken to be deterministic and known.

[2] Assumption 1 is somewhat severe; in Section 5 we discuss a possible extension to our method that avoids requiring it.

[3] The steady state distribution of $X(k)$ is the *Bernoulli convolution distribution*. For $\lambda \in \left(0, \frac{1}{2}\right)$ the distribution is singular and fractal in nature. For $\lambda \in \left(\frac{1}{2}, 1\right)$ it is difficult to know whether the distribution is absolutely continuous or not; the problem has been investigated by many, yet a complete answer is unknown. [21]

4. The distribution of $X(0)$, conditional on $S(0)$ is specified in advance. $X(0)$ is conditionally independent of $(S(1), S(2), \cdots)$ given $S(0)$.

Additional assumptions.

1. for every $s \in \mathcal{S}$, for all nonzero $x \in \mathbb{R}^d$, $\|A_s x\| < \|x\|$.
2. C_s^{obs} is an invertible (hence square) matrix;
3. There is a positive integer η such that for any sequence $s_1, \cdots, s_\eta$ of states of the Markov chain, the matrix
$$\begin{pmatrix} C_{s_\eta}^{\text{proc}} & A_{s_\eta} C_{s_{\eta-1}}^{\text{proc}} & \cdots & A_{s_\eta} \cdots A_{s_2} C_{s_1}^{\text{proc}} \end{pmatrix} \tag{3}$$
has rank equal to d, the number of rows.
4. the distribution of $X(0)$ is absolutely continuous and there are $\alpha_0, \gamma_0 > 0$ such that for every $s \in \mathcal{S}$ the PDF $f_{X(0)|S(0)=s}$ satisfies
$$f_{X(0)|S(0)=s}(x) = O(\exp(-\alpha_0 \|x\|^2)) \text{ as } \|x\| \to \infty. \tag{4}$$
and the CF $\tilde{f}_{X(0)|S(0)=s}$ satisfies
$$\tilde{f}_{X(0)|S(0)=s}(\tilde{x}) = O(\exp(-\gamma_0 \|\tilde{x}\|^2)) \text{ as } \|\tilde{x}\| \to \infty. \tag{5}$$

2.2 | Comments and consequences

No assumptions are made about the transition matrix M. We mention that two of the above assumptions can be considered *generic*, in that they hold for an open dense subset of the parameters:

- If C_s^{obs} is square then Assumption 2 is generic. This is because det $: \mathbb{R}^{d \times d} \to \mathbb{R}$ is a nonzero polynomial, so $\{C_s^{\text{obs}} : \det(C_s^{\text{obs}}) \neq 0\}$ is a non-empty Zariski open set, which is automatically open and dense in the Euclidean topology.
- Assumption 3 is also generic. If we define a map ζ that maps $(A_{s_1}, \cdots, A_{s_\eta}, C_{s_1}^{\text{proc}}, \cdots, C_{s_\eta}^{\text{proc}})$ to the matrix consisting of the first d columns of (3) then det $\circ\zeta$ is a nonzero polynomial map, so the set of parameters for which det $\circ\zeta$ is non-zero is non-empty and Zariski open, so it is is open and dense in the Euclidean topology. There are $|\mathcal{S}|^d < \infty$ possible choices of $s_1, \cdots, s_d$, and taking the intersection of the corresponding open dense sets gives an open dense set.

However, we should note that the accuracy/computation time tradeoff gets worse as the parameters approach the non-generic points.

Some intuitive explanation for Assumption 3 is necessary. Let us write, for now, $C(s_1, \cdots, s_\eta)$ for the matrix given in (3). By repeated application of the first part of (2), we find that $X(k+\eta)$ contains — among other items — the terms
$$C_{S(k+\eta)}^{\text{proc}} Z_{\text{proc}}(k+\eta) + A_{S(k+\eta)} C_{S(k+\eta-1)}^{\text{proc}} Z_{\text{proc}}(k+\eta-1) + \cdots + A_{S(k+\eta)} \cdots A_{S(k+2)} C_{S(k+1)}^{\text{proc}} Z_{\text{proc}}(k+1). \tag{6}$$
This sum can be rewritten as
$$C(S(k+\eta), \cdots, S(k+1)) Z, \tag{7}$$
where Z is the vector formed by stacking $Z_{\text{proc}}(k+\eta), \cdots, Z_{\text{proc}}(k+1)$ on top of each other. The entries of Z are iid $N(0,1)$ random variables. So the job of Assumption 3 is to ensure that the noise contribution to $X(k+\eta)$ has nonsingular covariance. The effect is to stop the characteristic function of $X(k)$ from spreading out too much over time.

Note that if all C_s^{proc} are invertible then Assumption 3 holds. We also note a consequence of Assumption 3. Write col(L) for the column space of a matrix L, i.e., the range of its corresponding linear transformation.

Proposition 1
If Assumption 3 holds, then for each s, col $\left(\begin{pmatrix} C_s^{\text{proc}} & A_s \end{pmatrix}\right) = \mathbb{R}^d$.

Proof. Choose $s_1 = \cdots = s_\eta = s$ and use the fact that col$(AB) \subseteq$ col(A) for any matrices A, B. □

3 | NUMERICAL METHOD

3.1 | Notation

Throughout this section we write f_W for the PDF of a random variable W and $f_{W|W'=w'}$ for the conditional distribution of W on another random variable W' at $W' = w'$. For each k suppose $y(k) \in \mathbb{R}$ is a realisation of the random variable $Y(k)$. We use the common shorthand $f_{W|\ell}$ for the PDF of W conditional on the observations up to time ℓ, that is, $f_{W|Y(0)=y(0),\cdots,Y(\ell)=y(\ell)}$. Similarly, write $f_{W|W'=w,\ell}$ as shorthand for $f_{W|W'=w,Y(0)=y(0),\cdots,Y(\ell)=y(\ell)}$.

If f is the distribution of an $\mathbb{R}^d$-valued random variable W, then we write $\tilde{f}$ for the characteristic function:

$$\tilde{f}(\tilde{w}) = \mathbb{E}(\exp(\imath W^{\mathrm{T}}\tilde{w})). \tag{8}$$

That is, $\tilde{f}$ is the Inverse Fourier Transform (IFT) of f, if we take the convention that the IFT of f is

$$\tilde{w} \mapsto \int_{\mathbb{R}^d} \exp(\imath w^{\mathrm{T}}\tilde{w}) f(w)\, dw. \tag{9}$$

Under this notation, the Fourier Transform (FT) of $\tilde{f}$ is

$$w \mapsto \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp(-\imath\tilde{w}^{\mathrm{T}} w) \tilde{f}(\tilde{w})\, d\tilde{w}. \tag{10}$$

3.2 | Description of the method

Our method is based on the observation that the distribution of $X(k)$ conditional on $S(k), Y(0), \cdots, Y(\ell)$ (where ℓ is $k-1$ or k) is well localised in space as well as frequency. The Poisson Summation Formula then implies that the function can be accurately described by its values on a finite grid,[4] and that its (continuous) Fourier Transform can be well approximated by a discrete Fourier Transform.

For the model described in Section 2.1 we can define a map $f_{k|\ell} : S \times \mathbb{R}^d \to \mathbb{R}$ that describes the (theoretical) filtered distribution:

$$(s, x) \mapsto P\left(S(k) = s | Y(0) = y(0), \cdots, Y(\ell) = y(\ell)\right) f_{X(k)|S(k)=s,Y(0)=y(0),\cdots,Y(\ell)=y(\ell)}(x). \tag{11}$$

The goal of the filter is to describe these maps accurately. Let us also define a corresponding "characteristic function" $\tilde{f}_{k|\ell} : S \times \mathbb{R}^d \to \mathbb{C}$:

$$(s, \tilde{x}) \mapsto P\left(S(k) = s | Y(0) = y(0), \cdots, Y(\ell)\right) \mathbb{E}\left(\exp(\imath X(k)^{\mathrm{T}}\tilde{x}) | S(k) = s, Y(0) = y(0), \cdots, Y(\ell) = y(\ell)\right). \tag{12}$$

We first explain an ideal version of the filter. The (continuous) Fourier Transform can be used to map between $f_{k|\ell}(s, x)$ and $\tilde{f}_{k|\ell}(s, \tilde{x})$.

Updating takes $f_{k|k-1}(s, x)$ and an observation $y(k)$ and produces $f_{k|k}(s, x)$. Using the nature of an HMM,

$$f_{k|k}(s, x) = \frac{f_{S(k),X(k),Y(k)|k-1}(s, x, y(k))}{f_{Y(k)|k-1}(y(k))}, \tag{13}$$

where

$$f_{S(k),X(k),Y(k)|k-1}(s, x, y) = f_{k|k-1}(s, x) f_{Y(k)|S(k)=s,X(k)=x}(y), \tag{14}$$

$$f_{Y(k)|k-1}(y) = \sum_{s\in S} \int_{\mathbb{R}^d} f_{S(k),X(k),Y(k)|k-1}(s, x, y)\, dx, \tag{15}$$

and using (2),

$$f_{Y(k)|S(k)=s,X(k)=x}(y) = \frac{1}{\det(C_s^{\mathrm{obs}})} g\left((C_s^{\mathrm{obs}})^{-1}\left(y - F_s x - G_s U(k)\right)\right). \tag{16}$$

where g is the PDF of a standard multivariate normal random variable.

[4] One can also relate this fact to the Sampling Theorem (which is itself related to the Poisson Summation Formula): If a PDF is *approximately* band-limited then it can be approximately reconstructed from its samples on an infinite grid; if it is also *approximately* compactly supported then it can be approximately reconstructed from the values on a finite grid.

The prediction step takes $f_{k-1|k-1}$ and produces $f_{k|k-1}$. If C_s^{proc} and the A_s are all invertible we can write this out: taking $M_{s_1,s}$ to be the transition probability from state s_1 to state s, then $f_{k|k-1}(s,x)$ becomes

$$\left(\frac{1}{\det A_s}\sum_{s_1\in\mathcal{S}} M_{s_1,s} f_{k-1|k-1}(s, A_s^{-1}(x - B_s U(t)))\right) * \left(\frac{1}{\det(C_s^{\text{proc}})} g\left((C_s^{\text{proc}}))^{-1}x\right)\right), \tag{17}$$

where the convolution is taken over x. The prediction step can be easily described in the Fourier domain, even if C_s^{proc} and A_s are not invertible:

$$\tilde{f}_{k|k-1}(s,\tilde{x}) = \left(\sum_{s_1\in\mathcal{S}} M_{s_1,s}\tilde{f}_{k-1|k-1}(s, A_s^{\mathrm{T}}\tilde{x})\right)\exp\left(\iota(B_s U(k))^{\mathrm{T}}\tilde{x}\right)\tilde{g}\left(C_s^{\text{proc}\,\mathrm{T}}\tilde{x}\right), \tag{18}$$

where $\tilde{g}$ is the CF of a standard multivariate normal random variable.

We now describe how $f_{k|\ell}$ and $\tilde{f}_{k|\ell}$ are discretised and how (13)–(15) and (18) can be computed from these discretisations. Let $x_0 \in \mathbb{R}^d$ and let $e_1, \cdots, e_d$ denote the standard basis vectors of $\mathbb{R}^d$. Let $q_1, \cdots, q_d$ be positive integers, $\rho_1, \cdots, \rho_d$ be positive real numbers and let

$$\mathcal{G} = \left\{x_0 + r_1\rho_1 e_1 + r_2\rho_2 e_2 + \cdots + r_d\rho_d e_d \,:\, \text{each } r_p, \cdots \in \{1, \cdots, q_p\}\right\}. \tag{19}$$

Let $V = \prod_{p=1}^d \rho_p$. Let

$$\tilde{\mathcal{G}} = \left\{\left(r_1 - \frac{q_1+1}{2}\right)\frac{2\pi}{q_1\rho_1}e_1 + \cdots + \left(r_d - \frac{q_d+1}{2}\right)\frac{2\pi}{q_d\rho_d}e_d \,:\, \text{each } r_p \in \{1, \cdots, q_p\}\right\}.$$

Let $\tilde{V} = \prod_{p=1}^d \frac{2\pi}{q_p\rho_p}$. Define the following boxes:

$$\mathcal{C} = \left\{x_0 + \begin{pmatrix} x_1 & \cdots & x_d\end{pmatrix}^{\mathrm{T}} : -q_p\rho_p < x_p < q_p\rho_p \text{ for } p = 1, \cdots, d\right\}, \tag{20}$$

$$\tilde{\mathcal{C}} = \left\{\begin{pmatrix}\tilde{x}_1 & \cdots & \tilde{x}_d\end{pmatrix}^{\mathrm{T}} : -\frac{\pi}{\rho_p} < \tilde{x}_p < \frac{\pi}{\rho_p} \text{ for } p = 1, \cdots, d\right\}. \tag{21}$$

Now $f_{k|\ell}(s,x)$ can be approximated by a discretisation $h_{k|\ell} : \mathcal{S}\times\mathcal{G} \to \mathbb{R}$ and, if $\tilde{\mathcal{G}}$ is some other grid, $\tilde{f}_{k|\ell}(s,x)$ can be approximated by a corresponding discretisation $\tilde{h}_{k|\ell} : \mathcal{S}\times\tilde{\mathcal{G}} \to \mathbb{R}$. We can convert between a discretisation of $f_{k|\ell}$ using $\mathcal{G}$ and a discretisation of $\tilde{f}_{k|\ell}$ using $\tilde{\mathcal{G}}$ with a discrete Fourier Transform, as follows. If we have $\tilde{h}_{k|\ell}$ and want to know $h_{k|\ell}$ we can set

$$h_{k|\ell}(s,x) = \tilde{V}\sum_{\tilde{x}\in\tilde{\mathcal{G}}} \exp(-\iota\tilde{x}^{\mathrm{T}}x)\tilde{h}_{k|\ell}(s,\tilde{x}). \tag{22}$$

A similar computation applies if we know $h_{k|\ell}$ and want to know $\tilde{h}_{k|\ell}$ (using V instead of $\tilde{V}$). Thanks to the choice of $\mathcal{G}, \tilde{\mathcal{G}}$, these sums can be computed simply by performing a Fast Fourier Transform (or its inverse) in each dimension. These transformations are straightforward approximations of certain Riemann integrals for the (inverse) Fourier Transform. But they are very accurate as we will see in Theorem 2. Given the number q_p of grid points in dimension p, we choose the grid so that ρ_p is of the form $\rho_{p,0}q_p^{-1/2}$ for some constant $\rho_{p,0}$; under this choice, Theorem 2 describes the error asymptotically as a function of the q_p.

The integral of (15) now becomes a summation (scaled by V). Since the summation is equivalent to evaluating the Fourier Transform at zero frequency, the fact that discretising the Fourier Transform is accurate applies also to discretisation of (15). Equations (13), (14) and (18) are now just componentwise calculations, except for the slightly tricky expression

$$\tilde{f}_{k|k}(s, A_s^{\mathrm{T}}\tilde{x}) \tag{23}$$

in (18). This expression requires resampling $\tilde{f}_{k|k}$. To this end, at the end of the preceding update step we have access to an approximation of $f_{k|k}$, so evaluating (23) is simply a matter of computing the Fourier Transform on a different grid; i.e., $A_s\tilde{\mathcal{G}}$ instead of $\tilde{\mathcal{G}}$. Thus, even though we have chosen $\mathcal{G}, \tilde{\mathcal{G}}$ so that an FFT efficiently transforms between f and $\tilde{f}$, we now have to use a more expensive technique to compute (23). If the dimension d of the underlying process X is 1, the Chirp Z-Transform (CZT)[22] can be applied. For higher dimensions, the CZT could be used to evaluate a Discrete Fourier Transform on a grid $A_s\tilde{\mathcal{G}}$, but only when A_s is symmetric.[23] When A_s is not symmetric, slower procedures are necessary.[5]

[5] We can use the Chirp Z-Transform in general, if we first decompose the necessary transformations as a product of symmetric transformations; see Section 5.

A comment on the choice of grids: they should at least be large enough and fine enough to be able to well approximate (i) the PDF and CF of the initial distribution; (ii) the PDF and CF of the process noise; (iii) the multiplying factor corresponding to the observation noise.

We now summarise the steps of the filter. We assume that the initial distribution $f_{0|-1}$ is known exactly. For a given grid $\mathcal{G}$, we let $h_{0|-1}$ be equal to $f_{0|-1}$ on the points of $S \times \mathcal{G}$.

Update step. Given an observation $y(k)$, for $(s, x) \in S \times \mathcal{G}$, compute

$$h_{S(k),X(k),Y(k)|k-1}(\cdot, \cdot, y(k)) : S \times \mathcal{G} \to \mathbb{R} \tag{24}$$

$$(s, x) \mapsto f_{Y(k)|S(k)=s,X(k)=x}(y(k)) h_{k|k-1}(s, x) \tag{25}$$

where $f_{Y(k)|S(k)=s,X(k)=x}(y)$ is computed as in (16). The values of $h_{S(k),X(k),Y(k)|k-1}$ give a discretisation of $f_{S(k),X(k),Y(k)|k-1}(s, x, y)$.

Approximate the integral of $f_{S(k),X(k),Y(k)|k-1}$ by summing the components of $h_{S(k),X(k),Y(k)|k-1}$ and scaling by V. This gives the contribution of the observation $Y(k) = y(k)$ to the likelihood computation. Finally, $h_{k|k}$ is computed by normalising (25), i.e., dividing by the integral just computed.

Prediction step. The prediction step computes $h_{k|k-1}$ from $h_{k-1|k-1}$. It can be broken down into 3 parts: mixing, conversion to resampled CF and applying translation and noise.

- *(Mixing.)* Compute the map $h^* : S \times \mathcal{G} \to \mathbb{R}$ defined by

$$h^*(s, x) = \sum_{s_1} M_{s_1,s} h_{k-1|k-1}(s_1, x). \tag{26}$$

  This is a discretisation of $P(S(k) = s|k-1) f_{X(k)|k-1}(x)$.

- *(Conversion to resampled CF.)* From the result of the mixing step we can compute the map $\tilde{h}^* : S \times \tilde{\mathcal{G}} \to \mathbb{C}$ defined by

$$\tilde{h}^*(s, \tilde{x}) = \begin{cases} V \sum_{x \in \mathcal{G}} \exp(\imath x^{\mathrm{T}} A_s^{\mathrm{T}} \tilde{x}) h^*(s, x) & A_s \tilde{x} \in \tilde{\mathcal{C}} \\ 0 & A_s \tilde{x} \notin \tilde{\mathcal{C}} \end{cases} \tag{27}$$

  This is a discretisation of $P(S(k) = s|k-1) \tilde{f}_{A_s X|k-1}(\tilde{x})$. Note the A_s^{T} in this expression; it essentially means that the Discrete Fourier Transform is being computed on the points of the grid $A_s^{\mathrm{T}} \tilde{\mathcal{G}}$ but the result is considered as a function on $S \times \tilde{\mathcal{G}}$.

- *(Applying translation and noise.)* Finally we evaluate the map $\tilde{h}_{k|k-1} : S \times \tilde{\mathcal{G}} \to \mathbb{C}$ defined by

$$\tilde{h}_{k|k-1}(s, \tilde{x}) = \exp(\imath (B_s U(k))^{\mathrm{T}} \tilde{x}) \tilde{g}(C_{S(k)}^{\mathrm{proc}\,\mathrm{T}} \tilde{x}). \tag{28}$$

  Given $\tilde{h}_{k|k-1}$ we can compute $h_{k|k-1}$ by performing a Fast Fourier Transform. Since $h_{k|k-1}$ is an approximation of a real-valued function, we can replace the imaginary parts of the computed values by 0.

The following lemma will be used to show that if the values of $h_{k|\ell}$ approximate the filtered density well, then the filtered characteristic function can be well approximated at any point of $\tilde{\mathcal{C}}$.

Lemma 1. Let $q = q_1 q_2 \cdots q_d$. Assume that each ρ_p is of the form $\rho_{p,0} q_p^{-1/2}$. Suppose $f : \mathbb{R}^d \to \mathbb{R}$. Let $\tilde{f} : \mathbb{R}^d \to \mathbb{C}$ be the Fourier Transform of f and suppose that there are constants $\alpha, \beta, \gamma, \delta > 0$ such that

$$f(x) \le \beta \exp\left(-\frac{\alpha}{2} \|x\|^2\right) \text{ for all } x \in \mathbb{R}^d, \tag{29}$$

$$|\tilde{f}(\tilde{x})| \le \delta \exp\left(-\frac{\gamma}{2} \|\tilde{x}\|^2\right) \text{ for all } \tilde{x} \in \mathbb{R}^d. \tag{30}$$

Then there are $\zeta, \xi, \nu > 0$, independent of $q_1, \cdots, q_d$, such that whenever $h : \mathcal{G} \to \mathbb{R}$ satisfies

$$|h(x) - f(x)| < \epsilon, \tag{31}$$

then for $\tilde{x} \in \tilde{\mathcal{C}}$

$$\left| V \sum_{x \in \mathcal{G}} \exp(\imath x^{\mathrm{T}} \tilde{x}) h(x) - \tilde{f}(\tilde{x}) \right| \le \nu q^{1/2} \epsilon + \zeta \exp\left(-\xi \min_{p=1}^{d} q_p\right), \tag{32}$$

for all sufficiently large $q_1, \cdots, q_d$, while for $\tilde{x} \notin \tilde{C}$

$$|\tilde{f}(\tilde{x})| \leq \zeta \exp\left(-\xi \min_{p=1}^{d} q_p\right). \tag{33}$$

An analogous result applies for the reverse procedure, i.e., approximating $f(x)$ from $\tilde{h}$.

Proof. If $\tilde{x} \notin \tilde{C}$ then

$$|\tilde{f}(\tilde{x})| \leq \delta \exp(-\gamma \|\tilde{x}\|^2/2) \leq \delta \exp\left(-\gamma \min_{p=1}^{d}(\pi/\rho_p)^2/2\right) \leq \delta \exp\left(-\gamma\pi^2 \min_{p=1}^{d}(\rho_{p,0}^{-2}) \min_{p=1}^{d}(q_p)/2\right).$$

Otherwise, define the infinite grids

$$\mathcal{G}' = \{x_0 + r_1 \frac{\rho_{p,0}}{\sqrt{q_1}} e_1 + \cdots + r_d \frac{\rho_{p,0}}{\sqrt{q_d}} e_d, \quad r_1, \cdots, r_d \in \mathbb{Z}\}, \tag{34}$$

$$\tilde{\mathcal{G}}' = \{\tilde{x} + 2\pi r_1 \frac{\sqrt{q_1}}{\rho_{1,0}} e_1 + \cdots + 2\pi r_d \frac{\sqrt{q_d}}{\rho_{d,0}} e_d, \quad r_1, \cdots r_d \in \mathbb{Z}\}. \tag{35}$$

By the Poisson Summation Formula and standard properties of the Fourier Transform,

$$\sum_{\tilde{z} \in \tilde{\mathcal{G}}'} \exp(\imath x_0^{\mathrm{T}}(\tilde{z} - \tilde{x}))\tilde{f}(\tilde{z}) = V \sum_{z \in \mathcal{G}'} \exp(\imath \tilde{x}^{\mathrm{T}} z) f(z). \tag{36}$$

All points in $\tilde{\mathcal{G}}' \backslash \{\tilde{z}\}$ are outside $\tilde{C}$ and so there are some $\zeta', \xi' > 0$ such that

$$\left| \sum_{\tilde{z} \in \tilde{\mathcal{G}}'} \exp(\imath x_0^{\mathrm{T}}(\tilde{z} - \tilde{x}))\tilde{f}(\tilde{z}) - \tilde{f}(\tilde{x}) \right| \leq \zeta' \exp(-\xi' \min_{p=1}^{d} q_p) \tag{37}$$

for all sufficiently large $q_1, \cdots, q_d$, because all the terms of the sum except $\tilde{f}(\tilde{x})$ are so small.[6] Meanwhile there are ζ'', ξ'' such that

$$\left| \sum_{z \in \mathcal{G}'} \exp(\imath \tilde{x}^{\mathrm{T}} z) f(z) - \sum_{z \in \mathcal{G}} \exp(\imath \tilde{x}^{\mathrm{T}} z) f(z) \right| \leq \zeta'' \exp(-\xi'' \min_{p=1}^{d} q_p) \tag{38}$$

for all sufficiently large $q_1, \cdots, q_d$. Combining (31) and (36)–(38), and since $V = \left(\prod_{p=1}^{d} \rho_{p,0}\right) q^{-1/2}$, we get (32). A similar proof applies for the reverse procedure. □

Theorem 2. There exists $\alpha' > 0$ such that for any sequence $y_0, \cdots, y_k$ of observations, if the above procedure is used — with the above mentioned method to choose the grid spacing based on the number of grid points — to estimate the likelihood $f_{Y(k)|k-1}$, then the resulting estimated likelihood $h_{Y(k)|k-1}$ satisfies

$$h_{Y(k)|k-1}(y_k|y_0, \cdots, y_{k-1}) - f_{Y(k)|k-1}(y_k|y_0, \cdots, y_{k-1}) = O\left(\exp(-\alpha' \min_{j=1}^{d} q_j)\right)$$

as $q_1, \cdots, q_d \to \infty$.

The proof relies on Theorem 3, which basically states that the filtered and predicted PDF and CF always have Gaussian tails. Theorem 2 describes the error between a density function and its approximation. It implies that, for a fixed sequence of observations, to achieve a specified error E in the approximate calculation of $f_{Y(k)|k-1}$, the number of grid points per dimension should be proportional to $\log(1/E)$. Thus, the total number of grid points should be proportional to $\log(1/E)^d$. The weakness of Theorem 2 is that it does not provide the constant factor in the asymptotic description, and does not preclude the possibility that this factor could grow as k increases. It seems plausible to us that the constant factor might be described in a way that does not grow as k increases (but could still depend on the observations themselves).

Proof. First we prove that there is $\alpha'' > 0$ such that

$$\epsilon_k := \max_{s \in S, x \in \mathcal{G}} \left(h_{k|k-1}(s, x) - f_{k|k-1}(s, x)\right) = O\left(\exp(-\alpha'' \min_{j=1}^{d} q_j)\right) \tag{39}$$

[6] The left hand side of (37) is bounded above by $\delta \sum_{\tilde{z} \in \tilde{\mathcal{G}}' - \{\tilde{x}\}} \exp\left(-\frac{\gamma}{2}\|\tilde{z}\|^2\right)$. Write $\tilde{\mathcal{G}}'$ as the union of sets of the form $\{x \in \tilde{\mathcal{G}} : |x_p| \neq (x_0)_p\}$ for $p = 1, \cdots, d$. Then show that for sufficiently large q_p, the sum on each of these sets is bounded above by $\zeta_p' \exp(-\xi_p' q_p)$ for some ζ_p', ξ_p'.

as $q_1, \cdots, q_d \to \infty$. Certainly this is true for $k = 0$ since we know $h_{0|-1} = f_{0|-1}$. We now proceed by induction: suppose (39) holds for ϵ_k. Since the update step simply involves multiplication by a bounded function, we see immediately that there is some $a_1 > 0$ such that

$$\max_{s\in S, x\in\mathcal{G}} \left(h_{k|k}(s,x) - f_{k|k}(s,x)\right) \le a_1\epsilon_k. \tag{40}$$

For fixed $s \in S$ let

$$f(x) = P(S(k) = s|k) f_{X(k)|S(k)=s,k}(x)$$
$$\tilde{f}(x) = P(S(k) = s|k) \tilde{f}_{X(k)|S(k)=s,k}(x).$$

Then, after the mixing step (26), we see that

$$|h^*(s,x) - f(x)| \le a_1\epsilon_k \text{ for } x \in \mathcal{G}. \tag{41}$$

Theorem 3 implies that there are $a, \alpha > 0$ such that

$$f(x) \le a\exp\left(-\frac{1}{2}\alpha\|x\|^2\right),\ \tilde{f}(\tilde{x}) \le a\exp\left(-\frac{1}{2}\alpha\|\tilde{x}\|^2\right). \tag{42}$$

But $\tilde{f}$ is the Fourier Transform of f. Take ρ_p to be $\rho_{p,0}/\sqrt{q_p}$. We now compute an upper bound on the error in the approximation of $\tilde{f}_{k|k-1}$ at a point $\tilde{x} \in \mathbb{R}^d$.

- When $A_s^{\mathrm{T}}\tilde{x} \in \tilde{C}$, use (27) and (32) to see

$$|\tilde{h}^*(s,\tilde{x}) - \tilde{f}(A_s^{\mathrm{T}}\tilde{x})| \le Vqa_1\epsilon_k + \zeta\exp\left(-\xi\min_{j=1}^{d} q_j\right). \tag{43}$$

- When $A_s^{\mathrm{T}}\tilde{x} \notin \tilde{C}$, Equation (27) gives $\tilde{h}^*(s,\tilde{x}) = 0$ while from (33) we get $|\tilde{f}(A_s^{\mathrm{T}}\tilde{x})| \le \zeta\exp\left(-\xi\min_{j=1}^{d} q_j\right)$. So (43) holds in this case also.

Applying (28), we see that there are some $\zeta, \xi > 0$ such that

$$\left|\tilde{h}_{k+1|k}(s,\tilde{x}) - \tilde{f}_{k+1|s,k}(\tilde{x})\right| \le Vqa_1\epsilon_k + \zeta\exp\left(-\xi\min_{j=1}^{d} q_j\right).$$

Using Lemma 1 in the other direction, we then see that there are some $\zeta', \xi' > 0$ (depending only on the parameters of the model) such that

$$\left|h_{k+1|k}(s,x) - f_{k+1|k}(s,x)\right| \le \tilde{V}Vq^2a_1\epsilon_k + \zeta'\exp\left(-\xi'\min_{j=1}^{d} q_j\right). \tag{44}$$

Since ξ, ξ' do not depend on k, we can assume without loss of generality that $\xi = \xi' = \alpha''$ and now (39) holds for ϵ_{k+1}.

A final application of the Poisson Summation Formula then shows that the approximation of $f_{Y(k)|k-1}$ by a summation as explained in the Update step above also only has an error of $O\left(\exp\left(-\alpha'\min_{j=1}^{d} q_j\right)\right)$. □

4 | DEMONSTRATIONS AND DISCUSSION

4.1 | Comparison to a collapsing method for computing likelihood

In this section we numerically compare our grid-based method against several other methods in the literature:

4.1.1 | Generalised Pseudo-Bayes Algorithm (GPBA)

A class of collapsing methods is described in [7, Section 3.1]. For a given *depth* D (a parameter to the filter), we approximate

$$f_{k-1|k-1,S(k-1)=s_1,S(k-2)=s_2,\cdots,s(k-D)=s_D}(x) \tag{45}$$

by a Gaussian. We keep track of the parameters of this Gaussian as well as our approximation of $P(S(k-1) = s_1, S(k-2) = s_2, \cdots, S(k-D) = s_D|Y(0), \cdots, Y(k-1))$. Under this approximation we can use Kalman filters to compute

$$f_{k|k,S(k)=s_1,S(k-2)=s_2,\cdots,S(k-D)=s_{D+1}} \tag{46}$$

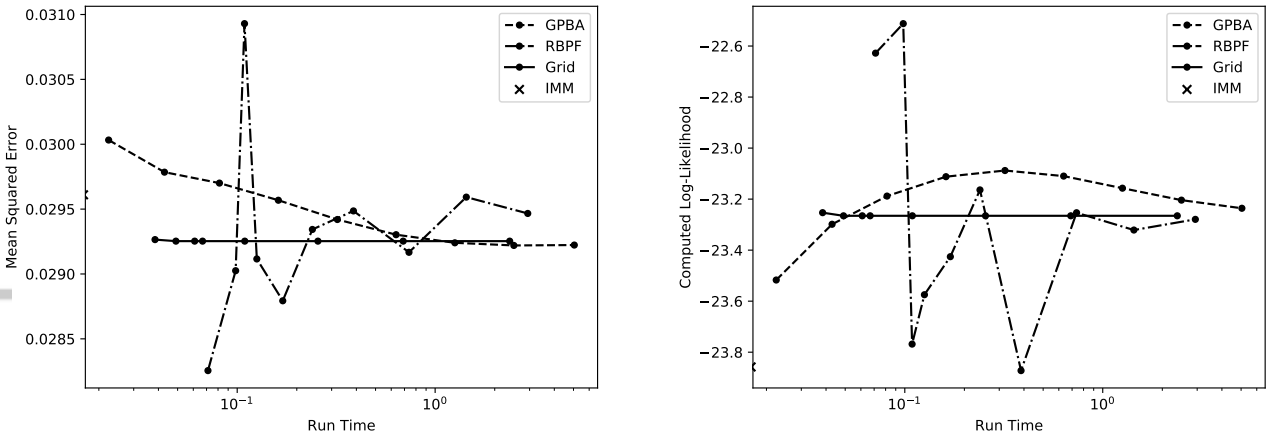


FIGURE 2 Results for 1-dimensional example. Left: mean squared error of the state estimate, plotted against the run time, for several different filters including our own (Grid). Right: log-likelihood computed from the observation sequence, plotted against the run time.

(and the corresponding probability) for each tuple $(s_1, \cdots, s_{D+1})$. After each update step we approximate the distribution $f_{k|k,S(k)=s_1,S(k-2)=s_2,\cdots,S(k-D+1)=s_D}$ by performing a collapsing step, which merges all the Gaussians with this common history into one, with mean and variance matching the mean and variance of the mixture.

4.1.2 | Interacting Multiple Models (IMM)

The Interacting Multiple Models algorithm can be interpreted as a variant of the depth 1 GPBA where the collapsing step occurs after the "mixing" part of the predicton step but before the prediction of the continuous part of the state.[24]

4.1.3 | Rao-Blackwellised Particle Filter (RBPF)

In the implementation of a particle filter it is possible to marginalise the discrete part of the state.[25] Section G A filter of this sort is described in [15] for nonlinear Markov-modulated systems; we apply it to our linear case for comparison to our method. For the proposal kernel, as suggested in [15, Remark 1], we use the state transition kernel.

4.1.4 | Results and discussion.

We conducted numerical experiments for 2-state examples of 1–3 dimensions. The experimental parameters for these examples are given in Appendix A. The results for these examples are in Figs. 2 to 4. Run times include the time to perform all burn steps and filtering steps, but not the time to initialise the data structures.

In terms of the mean squared error of the state estimate, the grid-based filter does not significantly outperform the other filters in these examples. This is because all the filters come within a few percent of the ideal filter without much effort. The real advantage of the grid-based filter, at least in 1–2 dimensions, is that it has an accurate numerical representation of the filtered density, which is helpful for computing any function of this density. For example, it converges more quickly to the log-likelihood computed from the observation sequence. For the 1-dimensional example, the filter requires very little time to get an accurate approximation of this value. The benefit is less pronounced in 2 dimensions, and by 3 dimensions the favourable asymptotics are neutralised by the large constant factor. Only the collapsing filters (GPBA and IMM) get close to the optimal MSE state estimate and the correct log-likelihood calculation in a reasonable amount of time. (It should be noted that the RBPF, being a randomised method, would not produce the same performance curve from run to run with different random number seeds.)

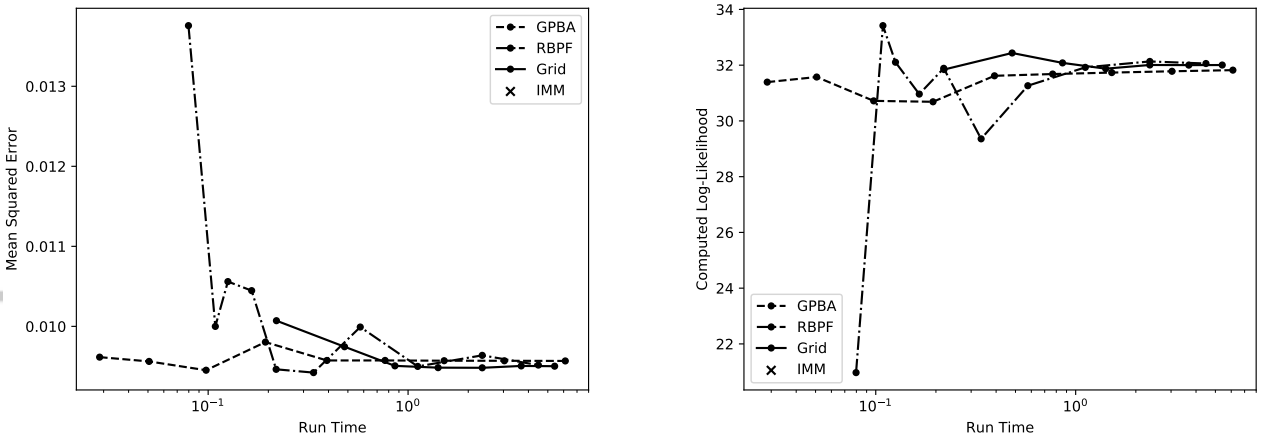


FIGURE 3 Results for 2-dimensional example. Left: mean squared error of the state estimate. Right: log-likelihood.

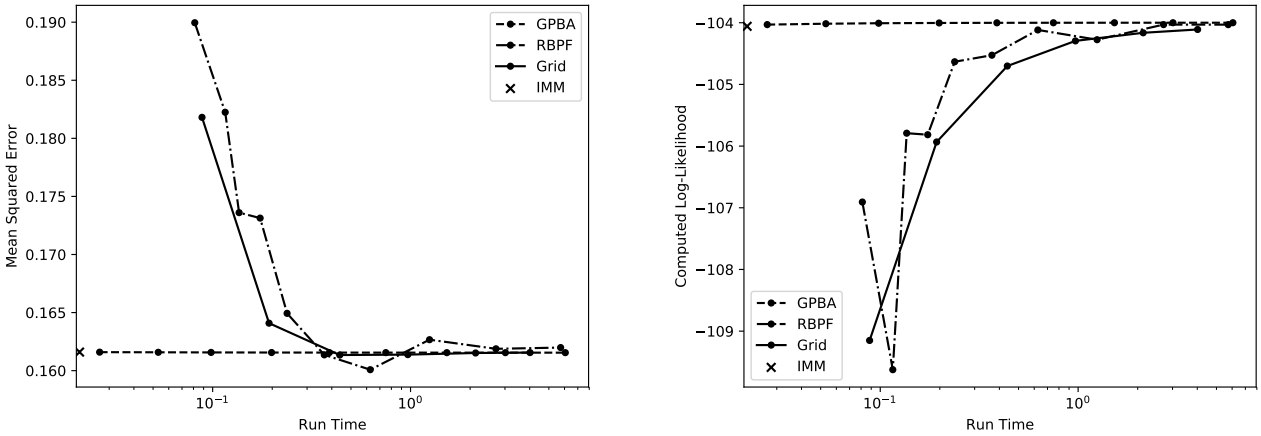


FIGURE 4 Results for 3-dimensional example. Left: mean squared error of the state estimate. Right: log-likelihood.

4.2 | Comparison to a method with piecewise linear interpolation

The filtering method described in [20] is similar to ours although an FFT is conducted to convert from the PDF values to the CF values. Following this, it is necessary to resample the CF to implement (23). It is not made explicit in [20] how to compute the resampled values. In a sense, the slow Fourier Transform calculation (or Chirp Z-Transform when it is available) is one option for resampling. A tempting alternative is to perform piecewise linear interpolation. This is faster for the same number of grid points, but we lose the ability to prove a fast convergence rate. In fact, in Fig. 5 we show that for the example from Section 1, linear interpolation leads to larger approximation errors. We use $q \in [64, 8192]$ points in the grid; the grid size is set to $\sqrt{q}$. A moderate improvement might be made by tuning the grid size; nevertheless it appears that the fast convergence is lost.

4.3 | Approximation of the computed steady state

If the steady state PDF of $X(t)$ exists, it can be approximated by repeating the prediction step without any observation steps, for long enough for the convergence to the fixed point to take effect. After each prediction step, we normalise to ensure that the

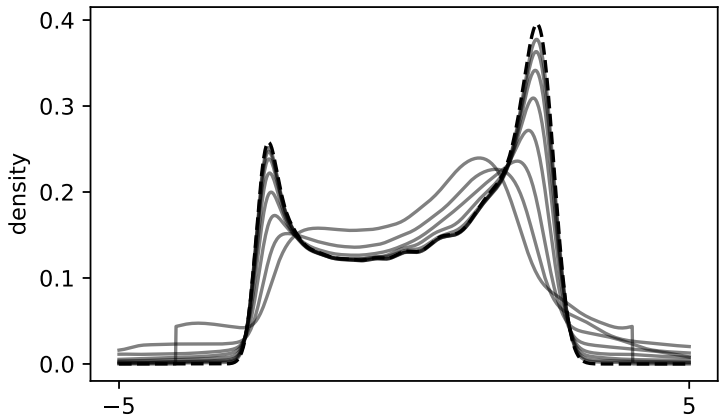


FIGURE 5 What happens if the resampling Fourier Transform step is replaced by a FFT followed by linear interpolation? The model parameters are the same as in the example in Section 1. Solid grey curves: approximation of the density of $X(100)$ using FFT and linear interpolation (Section 4.2). The number of grid points ranges from 64 to 8192 in powers of 2. Dashed black curve: approximation using the method of our paper, using 64 grid points.

integral is 1 (which is not automatically guaranteed due to the approximation errors). The initial choice of $h, \tilde{h}$ is not particularly important; we assume in our implementation that the state of the Markov chain has a discrete uniform distribution and the conditional PDF is an arbitrary nondegenerate Gaussian independent of the state. Since our prediction step is an approximation of the true prediction step, the limit we will actually arrive at will not in general be the best possible approximation of the steady state PDF. We do not have a theory for how big the difference is but we will numerically demonstrate that it can be very small if the discretisation is reasonably fine.

By using a Markov chain with only 1 state, we end up with a system that can be filtered exactly with a Kalman filter. The true steady state can be computed symbolically, which we can use for comparison to our result. We used the following parameters: $A_0 = 0.5, B_0 = 0, C_0^{\text{proc}} = 1$. Since we are only concerned with the steady state, the other parameters do not matter. For $q = 20$ and $q = 200$, we set both $\mathcal{G}$ and $\tilde{\mathcal{G}}$ to be symmetric around 0 with spacing $\sqrt{2\pi/q}$. We show the computed steady state after 100 iterations for these choices of q in Fig. 6 (top) against a plot of the symbolically computed steady state.

We next conducted an experiment aiming to get an idea of how the error depends on the number of iterations used. For $m = 1$ through to $m = 40$ iterations we compute the steady state using $q = 20$ and $q = 200$ grid points. We then evaluate the error between the approximate steady state and the true steady state at the grid points and find the maximum such error. Figure 6 (bottom) shows the result. For $q = 20$ the maximum error does not go below around 10^{-6} regardless of how many iterations are used. But for $q = 200$ the error swiftly decreases until it is comparable to the errors expected as a result of the machine precision.

4.4 | Notes on efficiency

We mention various options that are available for implementing the algorithm efficiently. (We did not aim for efficiency in our experimental implementation.)

- The grid operations of the algorithm is well-suited to modern computational architectures such as GPGPU, being highly parallelisable. Consider the steps of the filter, described in Section 3.2. The update step consists of pointwise calculations, and a summation. The prediction step consists of pointwise calculations, a (slow) Discrete Fourier Transform, and a Fast Fourier Transform. Summation and Fast Fourier Transform are well known to have efficient parallel implementations. The pointwise calculations, and the slow Discrete Fourier Transform, can be parallelised by computing each grid value in a separate thread.

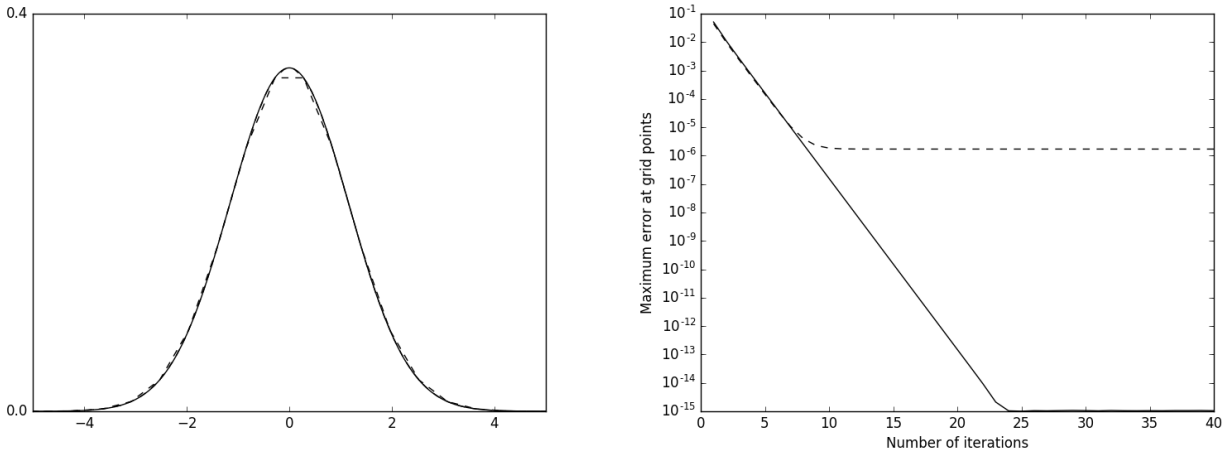


FIGURE 6 Left: a dashed curve shows the steady state PDF, as computed with 100 iterations of our prediction step, of a model whose steady state PDF is exactly the Gaussian shown by the solid curve. We used 20 points in the PDF and CF grids. Also shown (with a second dotted curve) is the PDF computed with 100 iterations on a grid with 200 points, but it is barely any different from the solid curve at this scale. Right: convergence of the error in the steady state PDF, as the number of iterations of the update step increases. We use the model described in Section 4.3. We use our method with 20 grid points (dashed curve) and 200 grid points (solid curve), and measure the maximum error — over all the grid points — between our approximation and the true (Gaussian) steady state. Note that with 200 grid points, the error reduces to a few machine epsilons after a small number of iterations.

- If the number of states of the Markov chain is very large, it may also be useful to use a parallel and/or cache-optimised matrix multiplication in the mixing part of the prediction step. (In our examples, the number of states does not warrant this).
- Since PDFs are real valued, it is not necessary to store or compute the imaginary part of the PDF at each step. Also, if the grid points for the CF are chosen to be symmetric around 0, then the values of the CF only need to be computed at half of the points.

5 | EXTENDED METHOD

Our main method has two significant limitations. Firstly, it requires all the matrices to have operator norm less than 1 (Assumption 1), which implies that it generally does not work for processes that do not have a steady state. Secondly, due to the possibility that A_s is not symmetric, it is unable to use a Fast Fourier Transform or Chirp Z-Transform to compute the resampled characteristic function (23). In this section we briefly explore a possibility to extend the method in two ways to alleviate the above issues.

- We associate with each state s and time step k an affine transformation $x \mapsto L_s(k)x + c_s(k)$, and write $X(k)$ as

$$X(k) = L_s(k)W_s(k) + c_s(k), \tag{47}$$

  and store the conditional distribution of $W_{S(k)}(k)$ instead of $X(k)$. This is equivalent to allowing the grids to "float". By choosing $L_s(k)$ and $c_s(k)$ so that $W_{S(k)}(k)$ is concentrated near zero, we remove the need for Assumption 1. The cost is that we can no longer mix the PDFs together at the start of the prediction step, because they are represented on different grids. As a consequence, a larger number of Fourier transforms have to be computed in each prediction step.
- The linear transformations of the grids required to perform the prediction steps are not generally symmetric. This precludes using fast Fourier transforms in both directions. However, it is possible to decompose a linear transformation as a product

of two symmetric transformations.[23,26] Thus, if we choose an appropriate grid in Fourier space, the Chirp Z-Transform can be used in both directions.

For $k \geq 1$ and for any states s, s', using (2) together with (47), we see that

$$\begin{aligned} W_{s'}(k) &= L_{s'}(k)^{-1}(X(k) - c_{s'}(k)) \\ &= L_{s'}(k)^{-1}\left(A_{S(k)}X(k-1) + B_{S(k)}U(k) + C^{\text{proc}}_{S(k)}Z_{\text{proc}}(k) - c_{s'}(k)\right) \\ &= L_{s'}(k)^{-1}\left(A_{S(k)}(L_s(k-1)W_s(k-1) + c_s(k-1)) + B_{S(k)}U(k) + C^{\text{proc}}_{S(k)}Z_{\text{proc}}(k) - c_{s'}(k)\right) \\ &= L_{s'}(k)^{-1}A_{S(k)}L_s(k-1)W_s(k-1) + L_{s'}(k)^{-1}\left(A_{S(k)}c_s(k-1) + B_{S(k)}U(k) - c_{s'}(k)\right) + L_{s'}(k)^{-1}C^{\text{proc}}_{S(k)}Z_{\text{proc}}(k). \end{aligned} \quad (48)$$

Suppose that for each s, we know the conditional distribution of $W_s(k-1)$ given $S(k-1) = s$. Based on the above calculation, we can design a method to compute the conditional distribution of $W_{s'}(k)$ given $S(k) = s'$. Let P, Q be symmetric matrices such that $PQ = L_{s'}(k)^{-1}A_sL_s(k-1)$. Let

$$V = QW_s(k-1) + P^{-1}L_{s'}(k)^{-1}\left(A_{S(k)}c_s(k-1) + B_{S(k)U(k)} - c_{s'}(k)\right) + P^{-1}L_{s'}(k)^{-1}C^{\text{proc}}_{S(k)}Z_{\text{proc}}(k). \quad (49)$$

If $S(k-1) = s$ and $S(k) = s'$ then (48) and (49) give $W_{s'}(k) = PV$. This means that

$$f_{W_{s'}(k)|S(k-1)=s,S(k)=s'} = f_{PV|S(k-1)=s,S(k)=s'}.$$

Therefore, we can replace the old prediction step with the following:

- For each s, s', compute the values of $\tilde{f}_{V|S(k-1)=s,S(k)=s'}$ on a grid;
- For each s, s', compute the values of $f_{PV|S(k-1)=s,S(k)=s'}$ on a grid;
- Apply mixing to compute the values of $f_{PV|S(k)=s'}$ on a grid.

The first two steps require conversion between ordinary space and Fourier space and back, but since P and Q are symmetric, both steps can be done with Chirp Z-Transforms.

We are free to choose $L_s(k)$ and $c_s(k)$. For example, the IMM can be used (Section 4.1.2) to get initial estimates μ, V of the conditional mean and variance of $X(k)$ given $S(k) = s$ and the observations so far. We can set $c_s(k) = X(k)$ and choose $L_s(k)$ such that $L_s(k)L_s(k)^T = V$.

In our experience, the inability to mix the PDFs at the start of the prediction step makes this method too slow. As such, without further innovations, it seems to us that this approach is not competitive.

6 | CONCLUSION AND OUTLOOK

Linear state-space models with Markov switching have been found useful for a variety of problems in engineering and econometrics. The general set of equations stated in Section 1 includes a multitude of well-known special cases. Traditional filtering methods in the literature are based on the notion of collapsing. More recent solutions for this problem and its generalisations are based on particle methods.

In contrast, our method is based on the idea that the underlying distributions are always well localised in space as well as frequency, so that they can be approximated with high accuracy by recording the values of the PDF and CF on a grid. The error in these representations decreases exponentially with the number of grid points in each dimension, while the cost only increases polynomially with the total number of grid points. As such, for low dimensions, this method has favourable convergence properties. The method is supported with a theory proving that the true filtered density has the required asymptotic conditions for the method to be efficient.

For the examples we considered, all the filters we tried arrived close (within a few percent) to the optimal mean squared error without much computation time. The grid-based filter, however, achieves an accurate approximation of the filtered PDF with lower computational cost, and this makes it capable of computing functions of the filtered PDF such as the log-probability of the observation sequence.

The cost, however, increases exponentially with the dimension of the system. We mention one possible direction for alleviating this. In our algorithm we have covered two cubes (one in PDF-space and one in CF-space) by grids. If the dimension is large, we

expect that the fraction of these cubes where the PDF and CF have large enough values to matter is small. For example, suppose the filtered PDF is approximately supported on a ball of radius 1. As the dimension d increases, the d-dimensional volume of a ball of radius 1 tends to zero, while the volume of its bounding cube grows exponentially.[7] Therefore,it is reasonable to predict that in high dimensions a much smaller number of grid points is needed, if we only use the grid points that are contained in the inscribed ball.[8] (This probably precludes using fast Fourier Transforms for both the PDF to CF and CF to PDF conversions.) Since we have not investigated the dependence of the required grid sizes on the dimension, it remains to be seen to what extent this resolves the curse of dimensionality.

Another limitation of the present method is that the error achieved with a given grid depends on the observations themselves – to the extent that it can possibly return non-positive likelihoods if the grid is not large enough to handle the observations. It would be ideal to have a way to temporarily upgrade the grid whenever rare enough observations are received to demand it. Failing that, it would be helpful to automatically select a grid based on the observations. We have not investigated these ideas.

We lastly remark on the potential for generalisation to parameters that we excluded in Assumptions 1 to 3. If Assumption 1 fails then the bulk of the filtered distribution might not stay close to the origin. We suggested in Section 5 a possible direction for alleviating this assumption, as well as allowing the use of Chirp Z-Transforms in both directions of the calculations. However, this comes with other computational costs. As for the other assumptions, the analysis in Appendix B hints that, at least in some situations, a filter could represent the filtered density as a mixture of absolutely continuous distributions on subspaces. At each filter step, $|S|$ copies of each of these subspaces are made and deformed by affine transformations. We suspect that the efficiency of such a method depends on whether the number of resulting subspaces grows quickly, slowly or stays finite over time.

□

APPENDIX

A EXPERIMENTAL PARAMETERS

A.1 Common parameters

Burn time for the simulation and all filters: 50 samples. Number of observations: 50.

Common filter parameters are as follows:

- GPBA: depths: $1, 2, 3, \cdots, 9$. We assume that at time 0 all "histories" are equally likely.
- RBPF: number of particles: $100, 200, 400, \cdots, 51200$. Resampling criterion: $N_{\text{eff}} < \frac{1}{3} \times$ (number of particles).
- Grid: PDF grid width: $\sqrt{q}$ where q is the grid points per dimension.

A.2 1-dimensional example

Model parameters for the 2-state, 1-dimensional example are as follows, with $U(t) \equiv 1$.

$$\text{Distribution vector of } S(0): \ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix},$$
$$\text{Distribution of } X(0): \ N(0, I),$$
$$\text{Transition matrix of } S: \ M = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix},$$
$$A_0 = 0.9, \qquad A_1 = 0.8,$$
$$B_0 = 0.25, \qquad B_1 = -0.5,$$
$$C_0^{\text{proc}} = C_1^{\text{proc}} = 0.1, \qquad F_0 = F_1 = 1,$$
$$G_0 = G_1 = 0, \qquad C_0^{\text{obs}} = C_1^{\text{obs}} = 0.3,$$

[7] The volume of a ball of radius 1 is [27] Section 5.19 $\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$ while the volume of its bounding cube is 2^d.

[8] This mirrors an argument for why particle filters may be capable of avoiding the curse of dimensionality. [28]

Grid filter points per dimension: $32, 64, 128, \cdots, 4096$.

A.3 2-dimensional example

Parameters for the 2-state, 2-dimensional example are as follows, with $U(t) \equiv 1$.

$$\text{Distribution vector of } S(0): \ \begin{pmatrix}0.5 & 0.5\end{pmatrix},$$
$$\text{Distribution of } X(0): \ N(0, I),$$
$$\text{Transition matrix of } S: \ M = \begin{pmatrix}0.98 & 0.02\\ 0.06 & 0.94\end{pmatrix},$$
$$A_0 = \begin{pmatrix}0.9 & 0\\ 0 & 0.9\end{pmatrix}, \qquad A_1 = \begin{pmatrix}0.8 & 0.2\\ 0.2 & 0.8\end{pmatrix},$$
$$B_0 = \begin{pmatrix}0.25\\ -0.25\end{pmatrix}, \qquad B_1 = \begin{pmatrix}-0.5\\ 0\end{pmatrix},$$
$$C_0^{\text{proc}} = C_1^{\text{proc}} = 0.1I, \qquad F_0 = F_1 = I,$$
$$G_0 = G_1 = 0, \qquad C_0^{\text{obs}} = C_1^{\text{obs}} = 0.1I,$$

where I is the 2×2 identity matrix.

Grid filter points per dimension: $32, 40, 48, \cdots, 80$.

A.4 3-dimensional example

Parameters for the 2-state, 3-dimensional example are as follows, with $U(t) \equiv 1$.

$$\text{Distribution vector of } S(0): \ \begin{pmatrix}0.5 & 0.5\end{pmatrix},$$
$$\text{Distribution of } X(0): \ N(0, I),$$
$$\text{Transition matrix of } S: \ M = \begin{pmatrix}0.9 & 0.1\\ 0.2 & 0.8\end{pmatrix},$$
$$A_0 = \begin{pmatrix}0.9 & 0 & 0\\ 0 & 0.8 & 0.1\\ 0 & -0.1 & 0.8\end{pmatrix}, \qquad A_1 = \begin{pmatrix}0.8 & 0.2 & 0\\ -0.2 & 0.8 & 0\\ 0 & 0.1 & 0.8\end{pmatrix},$$
$$B_0 = \begin{pmatrix}0\\ 0.2\\ 0\end{pmatrix}, \qquad B_1 = \begin{pmatrix}0\\ -0.2\\ 0\end{pmatrix},$$
$$C_0^{\text{proc}} = C_1^{\text{proc}} = 0.3I, \qquad F_0 = F_1 = I,$$
$$G_0 = G_1 = 0, \qquad C_0^{\text{obs}} = C_1^{\text{obs}} = 0.3I,$$

where I is the 3×3 identity matrix.

Grid filter points per dimension: $8, 10, 12, \cdots, 18$.

B ASYMPTOTICS OF THE DISTRIBUTIONS

Here we prove that the (theoretical) filtered distribution satisfies at every point in time the asymptotic conditions required for our method to be efficient.

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space underlying the model. We could, for example, let $(\Omega, \mathcal{F}, \mathcal{P})$ be the product of the probability spaces for (i) the finite-state Markov chain S, (ii) the distribution of $X(0)$, (iii) the iid process Z^{proc} and (iv) the iid process Z^{obs}. It can then be seen that all of the quantities referred to in (2) are random variables in this probability space. However, we will actually assume that $(\Omega, \mathcal{F}, \mathcal{P})$ is larger than this so that there is room to construct auxiliary random variables, independent from $S(k)$, $Z^{\text{proc}}(k)$, $Z^{\text{obs}}(k)$, as we need them.

In Appendix B.1 we prove lemmas related to the comparison of conditional densities of random variables. If V is an arbitrary random variable, c is a $\sigma(V)$-measurable random vector in $\mathbb{R}^d$ and K is a $\sigma(V)$-measurable random positive semidefinite matrix, we define $\mathcal{E}(V, c, K)$ which is the set of random variables that, conditional on V, have bounded density relative to the probability measure of an $N(c, K)$ random variable. This definition, made precise in Appendix B.1, allows us to compare the random variables of interest with the random variables in a (non-Markov) linear state-space model. The equations for a Kalman Filter then become equations describing the asymptotics of the conditional densities of our random variables. Appendix B.2 proves analogous lemmas for comparing the characteristic functions. Appendix B.3 puts these lemmas to use, giving for all k an asymptotic description of the conditional density and characteristic function of $X(k)$ given the observations up to time k.

For some measurable spaces we save notation by not stating the σ-algebra. Specifically for $U \subseteq \mathbb{R}^d$ or $U \subseteq \mathbb{R}^{m\times n}$ (the space of $m \times n$ matrices) we implicitly use the Borel σ-algebra $\mathcal{B}(U)$. We use $\mathcal{X} \times \mathcal{Y}$ for the Cartesian product of sets $\mathcal{X}, \mathcal{Y}$; when $\mathcal{X}$ and $\mathcal{Y}$ are measurable spaces (resp. measure spaces), the expression $\mathcal{X} \times \mathcal{Y}$ refers to the product measurables spaces (resp. measure spaces).

Write $\mathbb{R}^{d\times d}_{\geq 0}$ (resp. $\mathbb{R}^{d\times d}_{>0}$) for the set of positive semidefinite symmetric $d \times d$ matrices (resp. positive definite symmetric $d \times d$ matrices). These sets inherit the Lebesgue measure from $\mathbb{R}^{d\times d}$. For matrices A, B, write $A \geq B$ if $A - B \in \mathbb{R}^{d\times d}_{\geq 0}$; write $A > B$ if $A - B \in \mathbb{R}^{d\times d}_{>0}$. We assume that the reader is familiar with well-known properties of positive semidefinite matrices. In particular, if $K \in \mathbb{R}^{d\times d}_{\geq 0}$ then there is a nonempty set $\mathcal{U}(K)$ of $d \times r$ matrices L satisfying $\operatorname{rank}(L) = r$ and $LL^{\mathrm{T}} = K$. For all such L the column spaces $\operatorname{col}(L)$ and $\operatorname{col}(K)$ are equal. If $K : \Omega \to \mathbb{R}^{d\times d}_{>0}$ is a random variable then it is possible to choose a random variable $L : \Omega \to \mathbb{R}^{d\times d}$ such that $LL^{\mathrm{T}} = K$: for example, $L(\omega)$ can be computed by applying the Gram-Schmidt procedure to $\mathbb{R}^d$ with the inner product defined by $\langle x, y\rangle := x^{\mathrm{T}} K(\omega)^{-1} y$.

B.1 Asymptotic classes for density

Our approach to proving asymptotics of conditional probability densities is to compare the random variables to simpler random variables whose densities provide asymptotic bounds. It is frequently easier to work with the language of conditional expectations rather than conditional densities. The following lemma shows how two random variables, conditioned on a third, can be compared in equivalent ways: in terms of (1) their probabilities, (2, 3) their effects on expectations and (4) their densities.

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. Let Θ be the set of bounded measurable maps $\mathbb{R}^d \to [0, \infty)$. Let Θ_V be the set of bounded measurable maps $\mathcal{V} \times \mathbb{R}^d \to [0, \infty)$.

Lemma 2. Let $\mathcal{V}$ be an arbitrary measurable space and let $V : \Omega \to \mathcal{V}$ and $X, W : \Omega \to \mathbb{R}^d$ be random variables. Suppose $a : \mathcal{V} \to [0, \infty)$ is measurable. The following are equivalent:

1. For each each $E \in \mathcal{B}(\mathbb{R}^d)$,
$$\mathcal{P}(X \in E|V) \leq a(V)\mathcal{P}(W \in E|V),\ \mathcal{P} - \text{a.s.}; \tag{B1}$$
2. For each $\theta \in \Theta$,
$$\mathbb{E}(\theta(X)|V) \leq a(V)\,\mathbb{E}(\theta(W)|V),\ \mathcal{P} - \text{a.s.}; \tag{B2}$$
3. For each $\theta' \in \Theta_V$,
$$\mathbb{E}(\theta'(V, X)|V) \leq a(V)\,\mathbb{E}(\theta'(V, W)|V),\ \mathcal{P} - \text{a.s.}; \tag{B3}$$
4. For any version of $\mu_{W|V}$ it is possible to choose a version of $\mu_{X|V}$ such that for each $y \in \mathcal{V}$, $\mu_{X|V}(\cdot|v)$ is absolutely continuous with respect to $\mu_{W|V}(\cdot|v)$ and a version of the density $f_{X|V}$ of $\mu_{X|V}$ relative to $\mu_{W|V}$ can be chosen so that $f_{X|V}(x|v) \leq a(v)$, for all x, v.

Proof. (2) $\Rightarrow$ (1). In (B2), set $\theta(x) = \chi_E(x)$, the indicator function for E.
(1) $\Rightarrow$ (4). Let E be a measurable subset of $\mathbb{R}^d \times \mathcal{V}$. Write $d\mu_V$ for the probability measure of V on $\mathcal{V}$. If $\mathcal{P}((W, V) \in E) = 0$ then $\mathcal{P}((X, V) \in E) = \int_{\mathcal{V}} \mathbb{E}(\chi_E(X, v)|V = v)\, d\mu_V(v) \leq \int_{\mathcal{V}} a(v)\,\mathbb{E}(\chi_E(W, v)|V = v)\, d\mu_V(v) = 0$, showing that the joint density $\mu_{X,V}$ is absolutely continuous relative to the joint density $\mu_{W,V}$. Let $f_{X,V} : \mathbb{R}^d \times \mathcal{V} \to \mathbb{R}$ be the density. A standard argument shows that $f_{X,V}(x, v) \leq a(v)$, $\mu_{W,V}$ – a.s.[9] In fact we may take $f_{X,V}(x, v) \leq a(v)$ everywhere, since changing a density

[9] Prove that $\mu_{W,V}(\{(x, v) : f_{X,V}(x, v) > a(v) + 1/n\}) = 0$ for each integer $n > 0$ and take a countable union.

on a set of probability 0 does not stop it from being a density. We can choose the density f_V of μ_V relative to μ_V to be 1. The conditional density $f_{X|V}$ is related to the joint density via the following formula[29] Theorem B.52:

$$f_{X|V}(x|v) = \frac{f_{X,V}(x,v)}{f_V(v)}, \tag{B4}$$

So $f_{X|V}(x|v) = f_{X,V}(x,v) \le a(v)$.
(4) $\Rightarrow$ (3). With $f_{X|V}$ as in Item 4,

$$\mathbb{E}(\theta'(V,X)|V=v) = \int_{\mathbb{R}^d} \theta'(v,x)\, d\mu_{X|V}(x|v) = \int_{\mathbb{R}^d} \theta'(v,x) f_{X|V}(x|v)\, d\mu_{W|V}(x)$$
$$\le \int_{\mathbb{R}^d} \theta'(v,x) a(v)\, d\mu_{W|V}(x) = a(v)\, \mathbb{E}(\theta'(V,W)|V=v).$$

(3) $\Rightarrow$ (2). If $\theta \in \Theta$, let $\theta'(v,x) := \theta(x)$; then $\theta' \in \Theta_V$ and (B3) implies (B2). □

Let $\mathcal{F}_1$ be the σ-algebra generated by the random variables $X(0)$ as well as $S(k)$, $Z^{\text{proc}}(k)$, $Z^{\text{obs}}(k)$ for $k \in \mathbb{N}$. That is, $\mathcal{F}_1$ is the smallest σ-algebra with respect to which all of the random variables in (2) are measurable. Suppose $V : \Omega \to \mathcal{V}$ is a random variable measurable with respect to $\mathcal{F}_1$ (the measurable space $\mathcal{V}$ is arbitrary). Let $c : \Omega \to \mathbb{R}^d$ and $K : \Omega \to \mathbb{R}^{d\times d}_{\ge 0}$ be $\sigma(V)$-measurable maps. Let $W_{c,K} : \Omega \to \mathbb{R}^d$ be such that the conditional distribution of $W_{c,K}$ on $\mathcal{F}_1$ is $N(c,K)$. (The existence of such $W_{c,K}$ can be ensured by replacing Ω by an appropriate product probability space). Define $\mathcal{E}(V,c,K)$ to be the set of random variables $X : \Omega \to \mathbb{R}^d$ such that there is measurable $a : \mathcal{V} \to [0,\infty)$ such that, for any $E \in \mathcal{B}(\mathbb{R}^d)$,

$$\mathcal{P}(X \in E|V) \le a(V)\mathcal{P}(W_{c,K} \in E|V). \tag{B5}$$

This is the condition of Lemma 2 Item 1 applied to X and $W_{c,K}$; any of the other conditions of Lemma 2 can be used equivalently.

Lemma 3. Let $K, K_1, K_2 : \Omega \to \mathbb{R}^{d\times d}_{\ge 0}$ and $c, c_1 : \Omega \to \mathbb{R}$ be $\sigma(V)$-measurable.

1. If $V = (V_1, V_2)$ where the codomain of V_1 is finite and c, K are $\sigma(V_2)$-measurable, then $\mathcal{E}(V,c,K) \subseteq \mathcal{E}(V_2,c,K)$.
2. If $0 < K_1(\omega) \le K_2(\omega)$ for all ω then $\mathcal{E}(V,c,K_1) \subseteq \mathcal{E}(V,c,K_2)$.
3. If $0 < K(\omega)$ for all ω then for any $\epsilon > 0$, we have $\mathcal{E}(V,c,K) \subseteq \mathcal{E}(V,c_1,(1+\epsilon)K)$.

Proof. 1. Let $\mathcal{V}_1$ be the codomain of V_1. Suppose $X \in \mathcal{E}(V,c,K)$. Let a be the coefficient in (B5). For $E \in \mathcal{B}(\mathbb{R}^d)$,

$$\mathcal{P}(X \in E|V_2 = y_2) = \sum_{y_1 \in \mathcal{V}_1} \mathcal{P}(V_1 = y_1|V_2 = y_2)\mathcal{P}(X \in E|V_1 = y_1, V_2 = y_2)$$
$$\le \left(\sum_{y_1 \in \mathcal{V}} \mathcal{P}(V_1 = y_1|V_2 = y_2) a(y_1,y_2) \right) \mathcal{P}(W_{c,K} \in E|V_2 = y_2).$$

2. Let $L_1, L_2 : \Omega \to \mathbb{R}^{d\times d}$ be $\sigma(V)$-measurable such that $L_i L_i^{\mathrm{T}} = K_i, i = 1,2$. Let $J = L_1^{-1} L_2$. Since $K_1 \le K_2$, for any $z \in \mathbb{R}^d$, $\|z\| \le \|Jz\|$.[10] Let $Z : \Omega \to \mathbb{R}^d$ be an auxiliary random variable, independent of V, such that $Z \sim N(0,I)$. Then, for any $E \in \mathcal{B}(\mathbb{R}^d)$,

$$\mathcal{P}(Z \in E|V = y) = \frac{1}{\sqrt{2\pi}^d} \int_E \exp\left(-\frac{1}{2}\|z\|^2\right) dz$$
$$\le \frac{\det(J)}{\sqrt{2\pi}^d} \int_{J^{-1}E} \exp\left(-\frac{1}{2}\|u\|^2\right) du = \det(J)\mathcal{P}(JZ \in E|V = y). \tag{B6}$$

Suppose $X \in \mathcal{E}(V,c,K_1)$; let a be the corresponding coefficient in (B5). By (B6), for $E \in \mathcal{B}(\mathbb{R}^d)$,

$$\mathcal{P}(X \in E|V) \le a(V)\mathcal{P}(W_{c,K_1} \in E|V) = a(V)\mathcal{P}(Z \in L_1^{-1}(E-c)|V)$$
$$\le \det(J)a(V)\mathcal{P}(JZ \in L_1^{-1}(E-c)|V) = \det(J)a(V)\mathcal{P}(W_{c,K_2} \in E|V).$$

Therefore $X \in \mathcal{E}(V,c,K)$.

[10] Since $\|L_1^{\mathrm{T}} z\| \le \|J^{\mathrm{T}} L_1^{\mathrm{T}} z\|$ and L_1 is invertible, for any $z \in \mathbb{R}^d$ we have $\|z\| \le \|J^{\mathrm{T}} z\|$. So the operator norm of $J^{-\mathrm{T}}$, and hence J^{-1}, is less than 1.

3. Let $Z : \Omega \to \mathbb{R}^d$ be independent of V such that $Z \sim N(0, I)$. For $\epsilon > 0$ there is Borel-measurable $\zeta : (\mathbb{R}^d)^2 \to \mathbb{R}$ such that for $b, b_1, z \in \mathbb{R}^d$,

$$\frac{1}{(1+\epsilon)}\|z - b_1\|^2 - \|z - b\|^2 \le \zeta(b, b_1), \tag{B7}$$

implying

$$\exp\left(-\frac{1}{2}\|z - b\|^2\right) \le \exp\left(\frac{\zeta(b, b_1)}{2}\right) \exp\left(-\frac{1}{2(1+\epsilon)}\|z - b_1\|^2\right), \tag{B8}$$

so that scaling and integrating over any $E \in \mathcal{B}(\mathbb{R}^d)$ gives

$$\mathcal{P}(Z \in E - b | V = y) \le \eta(b, b_1)\mathcal{P}(\sqrt{1+\epsilon}Z \in E - b_1 | V = y), \tag{B9}$$

where $\eta(b, b_1) = \exp\left(\frac{\zeta(b,b_1)}{2}\right)(1+\epsilon)^{d/2}$.

Suppose $X \in \mathcal{E}(V, c, K)$; let a be the coefficient in (B5). Let $L : \Omega \to \mathbb{R}^{d\times d}$ be $\sigma(V)$-measurable such that $LL^{\mathrm{T}} = K$. Let $b = L^{-1}c$ and $b_1 = L^{-1}c_1$. Then for $E \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}
\mathcal{P}(X \in E|V) &\le a(V)\mathcal{P}(W_{c,K} \in E|V) = a(V)\mathcal{P}(Z \in L^{-1}(E - c)|V) \\
&\le a(V)\eta(L^{-1}c, L^{-1}c_1)\mathcal{P}(\sqrt{1+\epsilon}Z \in L^{-1}(E - c_1)) \\
&= a(V)\eta(L^{-1}c, L^{-1}c_1)\mathcal{P}(W_{c_1,(1+\epsilon)K} \in E),
\end{aligned} \tag{B10}$$

showing that $x \in \mathcal{E}(V, c_1, (1+\epsilon)K_1)$. □

Suppose Γ is a finite subset of $\mathbb{N}$. Then the elements of Γ can be indexed so that $\Gamma_1 < \Gamma_2 < \cdots < \Gamma_{|\Gamma|}$. For $\ell_0, \ell_1 \in \mathbb{N}$ let $V_\Gamma^{\ell_0:\ell_1}$ denote the tuple

$$\left(S(\ell_0), S(\ell_0 + 1), \cdots, S(\ell_1), Y(\Gamma_1), Y(\Gamma_2), \cdots, Y(\Gamma_{|\Gamma|})\right). \tag{B11}$$

That is, $V_\Gamma^{\ell_0:\ell_1}$ is a random variable which consists of all the observations at times $\Gamma_1, \Gamma_2, \cdots, \Gamma_{|\Gamma|}$, together with the states of the Markov chain between times ℓ_0, ℓ_1. Suppose $K : \Omega \to \mathbb{R}^{d\times d}_{\ge 0}$ is $\sigma(V_\Gamma^{\ell_0:\ell_1})$-measurable. For $s \in S$ let

$$R_s(K) = A_s K A_s^{\mathrm{T}} + C_s^{\mathrm{proc}} C_s^{\mathrm{proc}\,\mathrm{T}}, \tag{B12}$$

$$H_s(K) = K F_s^{\mathrm{T}}(F_s K F_s^{\mathrm{T}} + C_s^{\mathrm{obs}} C_s^{\mathrm{obs}\,\mathrm{T}})^{-1}, \tag{B13}$$

$$Q_s(K) = (I - H_s(K)F_s)K. \tag{B14}$$

These definitions come from the equations for a Kalman Filter[30]: $R_s(K)$ computes the predicted covariance from the previous a posteriori covariance, $H_s(K)$ gives the Kalman gain and $Q_s(K)$ gives the updated a posteriori covariance from the predicted covariance. For each ω choose some $L(\omega) \in \mathcal{U}(K(\omega))$. Then[11]

$$Q_s(K) = L(\omega)\left(I + L(\omega)^{\mathrm{T}} F_s^{\mathrm{T}}(C_s^{\mathrm{obs}} C_s^{\mathrm{obs}\,\mathrm{T}})^{-1} F_s L(\omega)\right)^{-1} L(\omega)^{\mathrm{T}}. \tag{B15}$$

If $K(\omega)$ is invertible, then so is $L(\omega)$ and (B15) simplifies to

$$Q_s(K) = \left(K^{-1} + F_s^{\mathrm{T}}(C_s^{\mathrm{obs}} C_s^{\mathrm{obs}\,\mathrm{T}})^{-1} F_s\right)^{-1}. \tag{B16}$$

Lemma 4. Let $k, \ell \in \mathbb{N}$. Suppose Γ is a subset of $\{0, 1, \cdots, k-1\}$. Let c, K be $\sigma(V_\Gamma^{\ell:k-1})$-measurable and suppose $X(k-1) \in \mathcal{E}\left(V_\Gamma^{\ell:k-1}, c, K\right)$. Then:

1. $X(k-1) \in \mathcal{E}\left(V_\Gamma^{\ell:k}, c, K\right)$.
2. $A_{S(k)}X(k-1) \in \mathcal{E}\left(V_\Gamma^{\ell:k}, A_{S(k)}c, A_{S(k)}KA_{S(k)}^{\mathrm{T}}\right)$.
3. $X(k) \in \mathcal{E}\left(V_\Gamma^{\ell:k}, c_1, R_{S(k)}(K)\right)$ where

$$c_1 = A_{S(k)}c + B_{S(k)}U(k) \tag{B17}$$

with R_s as defined in (B12).

[11] To see that (B13) and (B14) give the same value for $Q_s(K)$ as (B15), show that $I - L^{\mathrm{T}}F_s^{\mathrm{T}}(F_s K F_s^{\mathrm{T}} + C_s^{\mathrm{obs}} C_s^{\mathrm{obs}\,\mathrm{T}})^{-1} F_s L$ is the inverse of $I + L^{\mathrm{T}}F_s^{\mathrm{T}}(C_s^{\mathrm{obs}} C_s^{\mathrm{obs}\,\mathrm{T}})^{-1} F_s L$.

Proof. Let a be the coefficient in (B5), i.e., for measurable $E \subseteq \mathbb{R}^d$,

$$\mathcal{P}(X(k-1) \in E | V_\Gamma^{\ell:k-1}) \leq a(V_\Gamma^{\ell:k-1}) \mathcal{P}(W_{c,K} \in E | V_\Gamma^{\ell:k-1}). \tag{B18}$$

1. The random variables $S(k)$ and $X(k-1)$ are conditionally independent given $V_\Gamma^{\ell:k-1}$. Therefore $\mathcal{P}(X(k-1) \in E | V_\Gamma^{\ell:k}) = \mathcal{P}(X(k-1) \in E | V_\gamma^{\ell:k-1})$. By the definition of $W_{c,K}$ we have $\mathcal{P}(W_{c,K} \in E | V_\Gamma^{\ell:k-1} = \mathcal{P}(W_{c,K} \in E | V_\Gamma^{\ell:k})$. Let $a_1(V_\Gamma^{\ell:k}) = a(V_\Gamma^{\ell:k-1})$. For measurable $E \subseteq \mathbb{R}^d$,

$$\mathcal{P}(X(k-1) \in E | V_\Gamma^{\ell:k}) \leq a_1(V_\Gamma^{\ell:k}) \mathcal{P}(W_{c,K} \in E | V_\Gamma^{\ell:k}). \tag{B19}$$

2. First note that $A_{S(k)}c$ and $A_{S(k)}KA_{S(k)}^{\mathrm{T}}$ are both $\sigma(V_\Gamma^{\ell:k})$-measurable. Now, it follows from (B19) that if $E \subseteq \mathbb{R}^d$ is measurable,

$$\mathcal{P}(A_{S(k)}X(k-1) \in E | V_\Gamma^{\ell:k}) \leq a_1(V_\Gamma^{\ell:k}) \mathcal{P}(A_{S(k)}W_{c,K} \in E | V_\Gamma^{\ell:k}) \tag{B20}$$

and the conditional distribution of $A_{S(k)}W_{c,K}$ given Y is $N(A_{S(k)}c, A_{s(k)}KA_{S(k)}^{\mathrm{T}})$.
3. Note that $Z_{\mathrm{proc}}(k)$ is independent of $V_\Gamma^{\ell:k}$. Let $\theta'(v,x) = \mathbb{E}(\theta(x + B_{S(K)}U(k) + C_{S(k)}^{\mathrm{proc}}Z(k)) | V_\Gamma^{\ell:k} = v)$. Then $\theta' \in \Theta_{V_\Gamma^{\ell:k}}$, and by Lemma 2 (1 $\implies$ 3) and (B20),

$$\begin{aligned}
&\mathbb{E}(\theta(A_{S(k)}X(k-1) + B_{S(k)}U(k) + C_{S(k)}^{\mathrm{proc}}Z_{\mathrm{proc}}(k)) | V_\Gamma^{\ell:k} = v) \\
&\quad = \mathbb{E}(\theta'(v, A_{S(k)}X(k-1)) | V_\Gamma^{\ell:k} = v) \\
&\quad \leq a(v)\, \mathbb{E}(\theta'(v, W_{A_{S(k)}c, A_{S(k)}KA_{S(k)}^{\mathrm{T}}}) | V_\Gamma^{\ell:k} = v) \\
&\quad = a(v)\, \mathbb{E}(\theta(W_{A_{S(k)}c, A_{S(k)}KA_{S(k)}^{\mathrm{T}}} + B_{S(k)}U(k) + C_{S(k)}^{\mathrm{proc}}Z(k)) | V_\Gamma^{\ell:k} = v) \\
&\quad = a(v)\, \mathbb{E}(\theta(W_{A_{S(k)}c + B_{S(k)}U(k), R_{S(k)}(K)}) | V_\Gamma^{\ell:k} = v),
\end{aligned} \tag{B21}$$

since $W_{c,K}$ and $Z^{\mathrm{proc}}(k)$ are conditionally independent given $V_\Gamma^{\ell:k}$. So Lemma 2 (2 $\implies$ 1) gives the result. □

Lemma 5. Let $k, \ell \in \mathbb{N}$. Suppose $\Gamma \subseteq \{0, 1, \cdots, k-1\}$. Let c, K be $\sigma(V_\Gamma^{\ell:k})$-measurable. Suppose $X(k) \in \mathcal{E}\left(V_\Gamma^{\ell:k}, c, K\right)$. Then $X(k) \in \mathcal{E}\left(V_{\Gamma\cup\{k\}}^{\ell:k}, c_1, Q_{S(k)}(K)\right)$ where

$$c_1 = c + H_{S(k)}(Y(k) - F_{S(k)}c - G_{S(k)}U_{S(k)}) \tag{B22}$$

with H_s, Q_s as defined in (B13) and (B14).

Proof. Write V as a shorthand for $V_\Gamma^{\ell:k}$. Note that, from its definition, $V_{\Gamma\cup\{k\}}^{\ell:k}$ can be identified with the pair $(V, Y(k))$. Let $Y' = F_{S(k)}W_{c,K} + G_{S(k)}U(k) + C_{S(k)}^{\mathrm{obs}}Z_{\mathrm{obs}}(k)$, so that the distribution of Y' conditional on $(W_{c,K}, V)$ is the same as the distribution of $Y(k)$ conditional on $(X(k), V)$. Choose a measure μ with respect to which the probability measure of $W_{c,K}$ conditional on V is absolutely continuous. Let μ_1 be the product measure of μ with Lebesgue measure on $\mathbb{R}^n$. Writing all densities relative to these measures,

$$\begin{aligned}
f_{X(k)|V} &= \frac{f_{X(k)|V}(x|v) f_{Y(k)|X(k),V}(y|x,v)}{f_{Y(k)|V}(y|v)} \\
&\leq \frac{a(v) f_{W_{c,K}|V}(x|v) f_{Y'|W_{c,K},V}(x|v)}{f_{Y'|V}(y|v)} \frac{f_{Y'|V}(y|v)}{f_{Y(k)|V}(y|v)}.
\end{aligned} \tag{B23}$$

Choosing $a_1(v,y) = \frac{a(v) f_{Y'|V}(y|v)}{f_{Y(k)|V}(y|v)}$, we get

$$f_{X(k)|V,Y(k)}(x|v,y) \leq a_1(v,y) f_{W_{c,K}|V,Y'}(x|v,y). \tag{B24}$$

Finally, the operation of the update step of a Kalman filter describes the distribution of $W_{c,K}$ conditional on (V, Y'); it is the distribution of $W_{c_1, Q_{S(k)}(K)}$. □

B.2 Asymptotic classes for the characteristic function

Suppose $Y : \Omega \to \mathcal{V}$ is a random variable and $K : \Omega \to \mathbb{R}_{\geq 0}^{d\times d}$ is $\sigma(Y)$-measurable. Define $\tilde{\mathcal{E}}(Y, K)$ to be the set of $\sigma(Y)$-measurable random variables X such that there is $\sigma(Y)$-measurable $a : \Omega \to \mathbb{R}$ satisfying

$$\left|\mathbb{E}(\exp(\imath X^{\mathrm{T}}\tilde{x}) | Y)\right| \leq a \exp\left(-\frac{1}{2}\tilde{x}^{\mathrm{T}}K\tilde{x}\right). \tag{B25}$$

Lemma 6. Let $K, K_1, K_2 : \Omega \to \mathbb{R}^{d\times d}_{\geq 0}$ be $\sigma(Y)$-measurable.

1. If $Y = (Y_1, Y_2)$ where Y_1 has finitely many states and K is $\sigma(Y_2)$-measurable, then $\tilde{\mathcal{E}}(Y, K) \subseteq \tilde{\mathcal{E}}(Y_2, K)$.
2. [12] If $K_1(\omega) \leq K_2(\omega)$ for all ω then $\tilde{\mathcal{E}}(Y, K_2) \subseteq \tilde{\mathcal{E}}(Y, K_1)$.

Proof. 1. Suppose $\mathcal{V}_i$ is the codomain of Y_i. Suppose $X \in \tilde{\mathcal{E}}(Y, K)$. Let a be the coefficient of (B25);

$$\begin{aligned}
&\left|\mathbb{E}(\exp(\imath X^{\mathrm{T}}\tilde{x})|Y_2 = y_2)\right| \\
&\quad\leq \sum_{y_1\in\mathcal{V}} \mathcal{P}(Y_1 = y_1|Y_2 = y_2)\left|\mathbb{E}(\exp(\imath X^{\mathrm{T}}\tilde{x})|Y_1 = y_1, Y_2 = y_2)\right| \\
&\quad\leq \left(\sum_{y_1\in\mathcal{V}} \mathcal{P}(Y_1 = y_1|Y_2 = y_2)a(y_1, y_2)\right)\exp\left(-\frac{1}{2}\tilde{x}^{\mathrm{T}}K\tilde{x}\right).
\end{aligned}$$

2. This follows from (B25) and the fact that $\tilde{x}^{\mathrm{T}}K_1\tilde{x} \leq \tilde{x}^{\mathrm{T}}K_2\tilde{x}$. □

Lemma 7. Let $\Gamma \subseteq \{0, 1, \cdots, k-1\}$. Suppose $X(k-1) \in \tilde{\mathcal{E}}\left(V_\Gamma^{\ell:k-1}, K\right)$. Then:

1. $X(k-1) \in \tilde{\mathcal{E}}(V_\Gamma^{\ell:k}, K)$.
2. $A_{S(k)}(X(k-1)) \in \tilde{\mathcal{E}}(V_\Gamma^{\ell:k}, A_{S(k)}KA_{S(k)}^{\mathrm{T}})$.
3. $X(k) \in \tilde{\mathcal{E}}(V_\Gamma^{\ell:k}, R_{S(k)}(K))$.

Proof. 1. This follows from the fact that $S(k)$ and $X(k-1)$ are conditionally independent given $S(k-1)$.
2. This follows from the fact that

$$\mathbb{E}(\exp(\imath(A_{S(k)}X(k-1))^{\mathrm{T}}\tilde{x})|V_\Gamma^{\ell:k}) = \mathbb{E}(\exp(\imath X(k-1)^{\mathrm{T}}(A_{S(k)}^{\mathrm{T}}\tilde{x}))|V_\Gamma^{\ell:k}). \tag{B26}$$

3.

$$\begin{aligned}
&\mathbb{E}(\exp(\imath X(k)^{\mathrm{T}}\tilde{x})|V_{0:k-1}^{\ell:k}) \\
&= \mathbb{E}\left(\exp\left(\imath\left(A_{S(k)}X(k-1) + B_{S(k)}U(k) + C_{S(k)}^{\mathrm{proc}}Z(k)\right)\tilde{x}\right)\middle|V_{0:k-1}^{\ell:k}\right) \\
&= \mathbb{E}\left(\exp\left(\imath\left(A_{S(k)}X(k-1)\right)^{\mathrm{T}}\tilde{x}\right)\middle|V_{0:k-1}^{\ell:k}\right)\exp\left(\imath B_{S(k)}U(k)\tilde{x} - \frac{1}{2}\tilde{x}^{\mathrm{T}}C_{S(k)}^{\mathrm{proc}}C_{S(k)}^{\mathrm{proc}}\tilde{x}\right).
\end{aligned}$$

So

$$|\,\mathbb{E}(\exp(\imath X(k)^{\mathrm{T}}\tilde{x})|V_{0:k-1}^{\ell:k})| \leq \exp\left(-\frac{1}{2}\left(\tilde{x}^{\mathrm{T}}A_{S(k)}KA_{S(k)}^{\mathrm{T}}\tilde{x} + \tilde{x}^{\mathrm{T}}C_{S(k)}^{\mathrm{proc}}C_{S(k)}^{\mathrm{proc}}\tilde{x}\right)\right) \tag{B27}$$

leading to the result. □

Lemma 8. Let $\Gamma \subseteq \{0, 1, \cdots, k-1\}$. Suppose $X(k) \in \tilde{\mathcal{E}}(V_\Gamma^{\ell:k}, K)$. Then

$$X(k) \in \tilde{\mathcal{E}}(V_{\Gamma\cup\{k\}}^{\ell:k}, Q_{S(k)}(K)). \tag{B28}$$

Proof. 4. Because C^{obs} is invertible, it follows from (2) that $\mu_{Y(k)|V_\Gamma^{\ell:k}}$ is absolutely continuous with respect to Lebesgue measure, and that a version of the density $q(\cdot|v) : \mathbb{R}^n \to \mathbb{R}$ is nonzero for all y, v.

For fixed $\tilde{x}, v$ define a map $h : \mathbb{R}^n \to \mathbb{R}$ by

$$h(y) = \mathbb{E}\left(\exp(\imath X(k)^{\mathrm{T}}\tilde{x}|V_\Gamma^{\ell:k} = v, Y(k) = y)\right)q(y). \tag{B29}$$

[12] Note that the inclusion is reversed compared to the corresponding item in Lemma 3, and the condition that $K_1 > 0$ is not required here.

Then, performing an inverse Fourier Transform,

$$\begin{aligned}
&\int_{y\in\mathbb{R}^n} \exp(\iota y^{\mathrm{T}}\tilde{y})h(y)\,dy \\
&\quad= \int_{y\in\mathbb{R}^n} \mathbb{E}\left(\exp(\iota(X(k)^{\mathrm{T}}\tilde{x}+y^{\mathrm{T}}\tilde{y})|V_\Gamma^{\ell:k}=v, Y=y)\right) q(y)\,dy \\
&\quad= \mathbb{E}\left(\exp\left(\iota\begin{pmatrix}X(k)\\Y(k)\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\tilde{x}\\\tilde{y}\end{pmatrix}\right)\middle| V_\Gamma^{\ell:k}=v\right).
\end{aligned} \tag{B30}$$

By the Fourier Inversion Theorem,

$$h(y)=\frac{1}{2\pi}\int_{\mathbb{R}^d}\exp\left(\iota\tilde{y}^{\mathrm{T}}y\right)\mathbb{E}\left(\exp\left(-\iota\begin{pmatrix}X(k)\\Y(k)\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\tilde{x}\\\tilde{y}\end{pmatrix}\right)\middle| V_\Gamma^{\ell:k}=v\right)d\tilde{y}. \tag{B31}$$

Let a be the coefficient in (B25) for $X(k)$. Using (2),

$$\begin{aligned}
&\left|\mathbb{E}\left(\exp\left(\iota\begin{pmatrix}X(k)\\Y(k)\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\tilde{x}\\\tilde{y}\end{pmatrix}\right)\middle| V_\Gamma^{\ell:k}\right)\right| \\
&\quad= \left|\mathbb{E}\left(\exp\left(\iota\left(X(k)^{\mathrm{T}}\tilde{x}+(F_{S(k)}X(k)+G_{S(k)}U(k)+C^{\mathrm{obs}}_{S(k)}Z^{\mathrm{obs}}(k))^{\mathrm{T}}\tilde{y}\right)\right)\middle| V_\Gamma^{\ell:k}\right)\right| \\
&\quad\le a\exp\left(-\frac{1}{2}\left((\tilde{x}+F_{S(k)}^{\mathrm{T}}\tilde{y})^{\mathrm{T}}K(\tilde{x}+F_{S(k)}^{\mathrm{T}}\tilde{y})+\tilde{y}^{\mathrm{T}}C^{\mathrm{obs}}_{S(k)}C^{\mathrm{obs}\ \mathrm{T}}_{S(k)}\tilde{y}\right)\right).
\end{aligned} \tag{B32}$$

Using (B29), (B31) and (B32), letting $E=(F_{S(k)}KF_{S(k)}^{\mathrm{T}}+C^{\mathrm{obs}}_{S(k)}C^{\mathrm{obs}\ \mathrm{T}}_{S(k)})$ and making the subsitution $\tilde{y}=\tilde{u}-^{\mathrm{T}}E^{-1}F_{S(k)}K\tilde{x}$,

$$\left|\mathbb{E}\left(\exp(\iota X(k)^{\mathrm{T}}\tilde{x})\,|V_\Gamma^{\ell:k},Y(k))\right)\right|\le a_1\exp\left(-\frac{1}{2}\left(\tilde{x}^{\mathrm{T}}K_1\tilde{x}\right)\right),$$

where $K_1=Q_{S(k)}(K)$ and

$$a_1(v,y)=\frac{a(v)}{(2\pi)^d q(v,y)}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{2}\tilde{u}^{\mathrm{T}}E\tilde{u}\right)d\tilde{u}. \tag{B33}$$

□

B.3 Asymptotic result

Lemma 9. Let $K_1, K_2\in\mathbb{R}^{d\times d}_{\ge 0}$.

1. If $K_1\le K_2$ then $R_s(K_1)\le R_s(K_2)$.
2. If $K_1\le K_2$ then $Q_s(K_1)\le Q_s(K_2)$.

Proof. 1. Assume $K_1\le K_2$. Then $R_s(K_2)-R_s(K_1)=A_s(K_2-K_1)A_s^{\mathrm{T}}\ge 0$. So $R_s(K_1)\le R_s(K_2)$.
2. If $0<K_1\le K_2$ then (B16) gives $Q_s(K_1)\le Q_s(K_2)$. But Q_s is continuous on $\mathbb{R}^{d\times d}_{\ge 0}$, so the inequality holds for all $K_1\le K_2$. □

For $K\in\mathbb{R}^{d\times d}_{\ge 0}$, any tuple of states $\mathbf{s}=(s_1,\cdots,s_{|\mathbf{s}|})\in S^\eta$, and any set of numbers $\Gamma\subseteq\{1,\cdots,s_{|\mathbf{s}|}\}$, define a matrix $\kappa(K,\mathbf{s},\Gamma)$ as follows. Let $\kappa_0=K$ and for $1\le\ell\le|\mathbf{s}|$ let

$$\kappa_\ell=\begin{cases}Q_{s_\ell}(R_{s_\ell}(\kappa_\ell)) & \ell\in\Gamma\\ R_{s_\ell}(\kappa_\ell) & \text{otherwise.}\end{cases} \tag{B34}$$

Then let $\kappa(K,\mathbf{s},\Gamma)=\kappa_{|\mathbf{s}|}$. The value $\kappa(K,\mathbf{s},\Gamma)$ is the computed covariance matrix after $|\mathbf{s}|$ steps of the Kalman filter, assuming the states are $s_1,\cdots,s_{|\mathbf{s}|}$. We can similarly define $c(x,\mathbf{s},\Gamma)$ to be the corresponding mean vector after $|\mathbf{s}|$ steps. Let $\mathbf{s}'=(s_1,\cdots,s_{|\mathbf{s}|-1})$ and let $\Gamma'=\Gamma-\{|\mathbf{s}|\}$.

Lemma 10. 1. If $K_1\le K_2$ then $\kappa(K_1,\mathbf{s},\Gamma)\le\kappa(K_2,\mathbf{s},\Gamma)$.

2. For any $K>0$, there is $\alpha'>0$ such that for any $\mathbf{s},\Gamma$, $\kappa(K,\mathbf{s},\Gamma)\le\alpha' I$.

3. There is real γ' such that for any $\mathbf{s}$ with $|\mathbf{s}| = \eta$, $\kappa(0, \mathbf{s}, \Gamma) \geq \gamma' I$.

4. There is real γ such that for any $K \geq 0$ and any $\mathbf{s}$, $\kappa(K, \mathbf{s}, \Gamma) \geq \gamma I$.

Proof. 1. Apply Lemma 9 inductively.
2. Let $p \in (0,1), r \in (0,\infty)$ be such that for all $s \in \mathcal{S}$ and for all $x \in \mathbb{R}^d$, $\|A_s^{\mathrm{T}} x\| \leq p\|x\|$ and $\|C_s^{\mathrm{proc}\,\mathrm{T}} x\| \leq r\|x\|$. The existence of r is immediate; p can be chosen in $(0,1)$ because of Assumption 1 and the fact that A_s and A_s^{T} have equal operator norms.

Let $\alpha_1 > 0$ be such that $K \leq \alpha_1 I$. Choose some $\epsilon' > 0$ and then choose any $\alpha' \geq \alpha_1$ satisfying $\alpha' p^2 + r^2 < \alpha' - \epsilon'$.[13] We will prove the following two steps, which applied inductively give the result: (i) if $\kappa(K, \mathbf{s}', \Gamma') \leq \alpha' I$ then $\kappa(K, \mathbf{s}, \Gamma') \leq \alpha' I$; (ii) if $\kappa(K, \mathbf{s}, \Gamma') \leq \alpha' I$ then $\kappa(K, \mathbf{s}, \Gamma' \cup \{|\mathbf{s}|\}) \leq \alpha' I$.

For Step (i), if $\kappa(K, \mathbf{s}', \Gamma') \leq \alpha I$, then using Lemma 9 Item 1, $\kappa(K, \mathbf{s}, \Gamma') = R_{s_{|\mathbf{s}|}}(\kappa(K, \mathbf{s}, \Gamma')) \leq R_{s_{|\mathbf{s}|}}(\alpha I) = \alpha A_{s_{|\mathbf{s}|}} A_{s_{|\mathbf{s}|}}^{\mathrm{T}} + C_{s_{|\mathbf{s}|}}^{\mathrm{proc}} C_{s_{|\mathbf{s}|}}^{\mathrm{proc}\,\mathrm{T}} \leq (\alpha p^2 + r^2) I \leq \alpha I$.

For Step (ii), suppose $\kappa(K, \mathbf{s}, \Gamma') \leq \alpha I$. Then by Lemma 9 Item 2, $\kappa(K, \mathbf{s}, \Gamma) = Q_{|\mathbf{s}|}(\kappa(K, \mathbf{s}, \Gamma')) \leq Q_{|\mathbf{s}|}(\alpha I)$. From (B16) we see that $Q_{|\mathbf{s}|}(\alpha I) \leq \alpha I$.
3. From (B15) we see that $\mathrm{col}(Q_s(R_s(\kappa_{\ell-1}))) = \mathrm{col}(R_s(\kappa_{\ell-1}))$ and from (B12) we see that $\mathrm{col}(R_s(\kappa_{\ell-1})) = A_{s_\ell} \mathrm{col}(\kappa_{\ell-1}) + \mathrm{col}(C_{s_\ell}^{\mathrm{proc}})$. Therefore, $\mathrm{col}(\kappa_\ell) = A_{s_\ell} \mathrm{col}(\kappa_{\ell-1}) + \mathrm{col}(C_{s_\ell}^{\mathrm{proc}})$. So all of

$$\mathrm{col}(C_{s_\eta}^{\mathrm{proc}}), \mathrm{col}(A_{s_\eta} C_{s_\eta - 1}^{\mathrm{proc}}), \cdots, \mathrm{col}(A_{s_\eta} A_{s_{\eta-1}} \cdots A_{s_2} C_{s_1}^{\mathrm{proc}}) \tag{B35}$$

are in $\mathrm{col}(\kappa(0, \mathbf{s}, \Gamma))$. By Assumption 3, $\kappa(0, \mathbf{s}, \Gamma)$ is invertible. Since there are only finitely many pairs of the form $(\mathbf{s}, \Gamma)$, there is some $\epsilon > 0$ such that for all $\mathbf{s}, \Gamma$, we have $\kappa(0, \mathbf{s}, \Gamma) \geq \epsilon I$.
4. We split this into 3 cases; since a choice of γ can be given for each of the 3 cases, the problem is solved by taking the smallest of them:

- $|\mathbf{s}| = \eta$: this is answered by Items 1 and 3.
- $|\mathbf{s}| > \eta$: decompose the tuple of states $\mathbf{s}$ into the tuples $\mathbf{s}_1 = (s_1, \cdots, s_{|\mathbf{s}|-\eta})$ and $\mathbf{s}_2 = (s_{|\mathbf{s}|-\eta+1}, \cdots, s_\eta)$. Also decompose Γ into $\Gamma_1 = \Gamma \cap \{1, 2, \cdots, |\mathbf{s}| - \eta\}$ and $\Gamma_2 = \Gamma \cap \{|\mathbf{s}| - \eta + 1, |\mathbf{s}| - \eta + 2, \cdots, |\mathbf{s}|\}$. Then
$$\kappa(K, \mathbf{s}, \Gamma) = \kappa(\kappa(K, \mathbf{s}_1, \Gamma_1), \mathbf{s}_2, \Gamma_2) \geq \gamma' I, \tag{B36}$$
by Items 1 and 3.
- $|\mathbf{s}| < \eta$: From (B16) we see that Q_s preserves invertibility. From (B12) and Proposition 1, we see that R_s also preserves invertibility. Since $K > 0$, we therefore have $\kappa(K, \mathbf{s}, \Gamma) > 0$ for any $\mathbf{s}, \Gamma$. Since there are only finitely many possibilities for $\mathbf{s}, \Gamma$ with $|\mathbf{s}| < \eta$ and $\Gamma \subseteq \{1, 2, \cdots, |\mathbf{s}|\}$, there is some $\gamma > 0$ such that $\kappa(K, \mathbf{s}, \Gamma) > \gamma I$.

□

Theorem 3. There exist constants α, γ such that for each k, and for $\Gamma \subseteq \{0, 1, \cdots, k-1\}$,

$$X(k) \in \mathcal{E}(V_\Gamma^{k:k}, 0, \alpha I) \cap \tilde{\mathcal{E}}(V_\Gamma^{k:k}, \gamma I). \tag{B37}$$

Proof. From Assumption 4 we know that there are β_0', δ_0', r_0 such that the initial PDF $f = f_{X(0)|S(0)=s}$ and the initial CF $\tilde{f} = \tilde{f}_{X(0)|S(0)=s}$ satisfy

$$\forall x, \|x\| > r_0 \implies f(x) \leq \beta_0' \exp\left(-\frac{1}{2}\alpha_0 \|x\|^2\right) \tag{B38}$$

$$\forall \tilde{x}, \|\tilde{x}\| > r_0 \implies |\tilde{f}(\tilde{x})| \leq \delta_0' \exp\left(-\frac{1}{2}\gamma_0 \|\tilde{x}\|^2\right). \tag{B39}$$

By the definition of the CF, we also know that $\tilde{f}$ is bounded. Combining this fact with (B39), we see that there is δ_0 such that $|\tilde{f}(\tilde{x})| \leq \delta_0 \exp(-\gamma_0 \|\tilde{x}\|^2)$ for all $\tilde{x}$. It follows that there is $\tilde{K}$ such that

$$X(0) \in \tilde{\mathcal{E}}(V_\emptyset^{0:0}, \tilde{K}). \tag{B40}$$

[13] Since $0 < p < 1$ this is true for all sufficiently large α'.

We also see for any x that $|f(x)| \leq \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\tilde{f}(\tilde{x})|\, dx$. So f is bounded and a similar procedure, using Lemma 2 Item 4 implies that there is $K > 0$ such that

$$X(0) \in \mathcal{E}(V_{\emptyset}^{0:0}, 0, K). \tag{B41}$$

Proof that there is α such that $X(k) \in \mathcal{E}(V_{\Gamma}^{k:k}, 0, \alpha I)$: Let $\mathbf{s} = (s_1, \cdots, s_k)$ and let $\Gamma \subseteq \{0, \cdots, k\}$. First we prove that

$$X(k) \in \mathcal{E}(V_{\Gamma}^{0:k}, c(0, \mathbf{s}, \Gamma), \kappa(K, \mathbf{s}, \Gamma)). \tag{B42}$$

The base case is given by (B41). Suppose

$$X(k-1) \in \mathcal{E}(V_{\Gamma'}^{0:k-1}, c(0, \mathbf{s}', \Gamma'), \kappa(K, \mathbf{s}', \Gamma')). \tag{B43}$$

Lemma 4 Item 3 and Lemma 5 give $X(k) \in \mathcal{E}(V_{\Gamma}^{0:k}, c(0, \mathbf{s}, \Gamma), \kappa(K, \mathbf{s}, \Gamma)$. This proves the inductive step, so (B42) holds. Choose any $\epsilon > 0$. Let α' be as in Lemma 10 Item 2 for K. From Lemma 3 Items 1 and 3, $X(k) \in \mathcal{E}(V_{\Gamma}^{0:k}, 0, (1+\epsilon)\alpha' I)$. Choosing $\alpha = (1+\epsilon)\alpha'$ gives $X(k) \in \mathcal{E}(V_{\Gamma}^{k:k}, 0, \alpha I)$.

Proof that there is γ such that $X(k) \in \tilde{\mathcal{E}}(V_{\Gamma}^{k:k}, \gamma I)$: First we prove that

$$X(k) \in \tilde{\mathcal{E}}(V_{\Gamma}^{0:k}, \kappa(\tilde{K}, \mathbf{s}, \Gamma). \tag{B44}$$

Equation (B40) gives the base case. Suppose $X(k-1) \in \tilde{\mathcal{E}}(V_{\Gamma'}^{0:k-1}, \kappa(\tilde{K}, \mathbf{s}', \Gamma')$. Lemma 7 Item 3 and Lemma 8 give $X(k) \in \tilde{\mathcal{E}}(V_{\Gamma}^{0:k}, \kappa(\tilde{K}, \mathbf{s}, \Gamma))$. This proves the inductive step. Let γ be as in Lemma 10 Item 4. From Lemma 6, $X(k) \in \tilde{\mathcal{E}}(V_{\Gamma}^{k:k}, \gamma I)$. □

References

1. Pauley M, McLean C, Manton J. Numerical Filtering of Linear State-Space Models with Markov Switching. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)* 2017. doi: 10.1109/ICASSP.2017.7952942

2. Engle RF, Hendry DF, Richard JF. Exogeneity. *Econometrica: Journal of the Econometric Society* 1983: 277–304. doi: 10.2307/1911990

3. Cappé O, Moulines E, Rydén T. *Inference in Hidden Markov Models*. Springer Series in StatisticsSpringer-Verlag . 2005.

4. Hamilton JD. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 1989; 57(2): 357–384. doi: 10.2307/1912559

5. Lam Ps. The Hamilton model with a general autoregressive component: estimation and comparison with other models of economic time series. *Journal of Monetary Economics* 1990; 26(3): 409–432. doi: 10.1016/0304-3932(90)90005-O

6. Kim CJ. Dynamic linear models with Markov-switching. *Journal of Econometrics* 1994; 60(1): 1–22. doi: 10.1016/0304-4076(94)90036-1

7. Tugnait JK. Detection and estimation for abruptly changing systems. *Automatica* 1982; 18(5): 607–615. doi: 10.1016/0005-1098(82)90012-7

8. Murphy K. Switching Kalman Filters. tech. rep., Compaq Cambridge Research Lab; : 1998.

9. Frühwirth-Schnatter, Sylvia . *Finite Mixture and Markov Switching Models*. Springer Series in StatisticsSpringer New York . 2006

10. Boers Y, Driessen H. A multiple model multiple hypothesis filter for Markovian switching systems. *Automatica* 2005; 41(4): 709–716.

11. McGinnity S, Irwin GW. Multiple model bootstrap filter for maneuvering target tracking. *IEEE Transactions on Aerospace and Electronic systems* 2000; 36(3): 1006–1012.

12. Doucet A, Gordon NJ, Krishnamurthy V. Particle filters for state estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing* 2001; 49(3): 613–624.

13. Andrieu C, Davy M, Doucet A. Efficient particle filtering for jump Markov systems. Application to time-varying autoregressions. *IEEE Transactions on Signal Processing* 2003; 51(7): 1762–1770.

14. Driessen H, Boers Y. Efficient particle filter for jump Markov nonlinear systems. *IEE Proceedings - Radar, Sonar and Navigation* 2005; 152(5): 323-326. doi: 10.1049/ip-rsn:20045075

15. Saha S, Hendeby G. Rao-Blackwellized particle filter for Markov modulated nonlinear dynamic systems. *2014 IEEE Workshop on Statistical Signal Processing (SSP)* 2014: 272–275.

16. Manton JH, Krishnamurthy V, Elliott RJ. Discrete time filters for doubly stochastic poisson processes and other exponential noise models. *International Journal of Adaptive Control and Signal Processing* 1999; 13(5): 393–416. doi: 10.1002/(SICI)1099-1115(199908)13:5<393::AID-ACS561>3.0.CO;2-J

17. Bucy R, Senne K. Digital synthesis of non-linear filters. *Automatica* 1971; 7(3): 287 - 298. doi: https://doi.org/10.1016/0005-1098(71)90121-X

18. Simandl M, Kralovec J, Soderstrom T. Anticipative grid design in point-mass approach to nonlinear state estimation. *IEEE Transactions on Automatic Control* 2002; 47(4): 699-702. doi: 10.1109/9.995053

19. Šimandl M, Královec J, Söderström T. Advanced point-mass method for nonlinear state estimation. *Automatica* 2006; 42(7): 1133 - 1145. doi: https://doi.org/10.1016/j.automatica.2006.03.010

20. Allam S, Dufour F, Bertrand P. Finite fast fourier transform filter for discrete linear systems with Markov jump parameters. *1999 European Control Conference (ECC)* 1999: 1654–1659.

21. Peres Y, Schlag W, Solomyak B. Sixty Years of Bernoulli Convolutions. In: No. 46 in Progress in Probability. Birkhäuser Basel. 2000 (pp. 39–65)

22. Rabiner L, Schafer R, Rader C. The chirp z-transform algorithm. *IEEE Transactions on Audio and Electroacoustics* 1969; 17(2): 86–92. doi: 10.1109/TAU.1969.1162034

23. Lawton WM. Multidimensional chirp algorithms for computing Fourier transforms. *IEEE transactions on image processing* 1992; 1(3): 429–431. doi: 10.1109/83.148616

24. Blom HAP, Bar-Shalom Y. The interacting multiple model algorithm for systems with Markovian switching coefficients. ; 33(8): 780–783. doi: 10.1109/9.1299

25. Cappé O, Godsill SJ, Moulines E. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE* 2007; 95(5): 899–924.

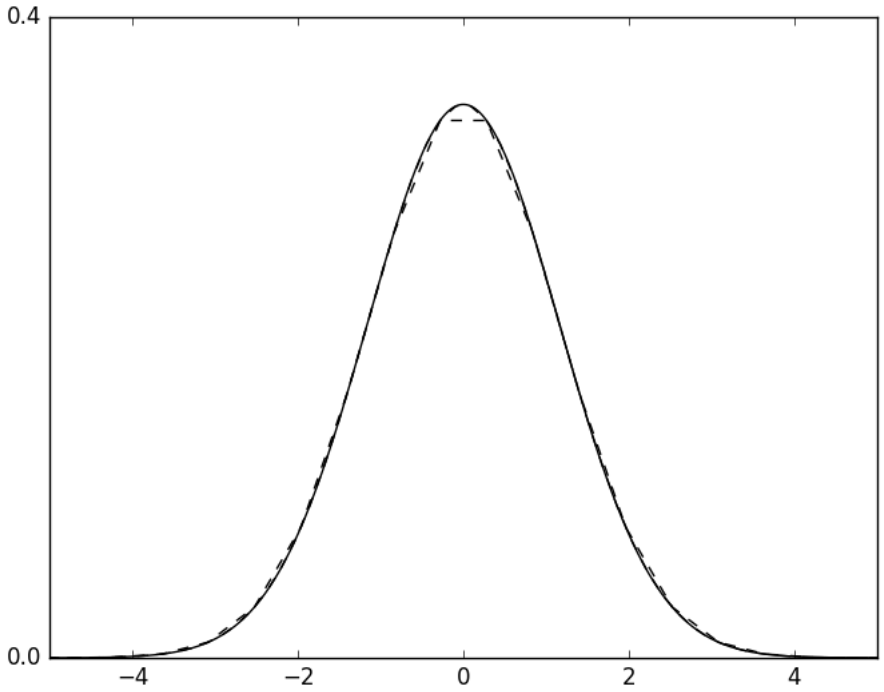
26. Carlson D. On real eigenvalues of complex matrices. *Pacific Journal of Mathematics* 1965; 15(4): 1119–1129.

27. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.19 of 2018-06-22; . F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

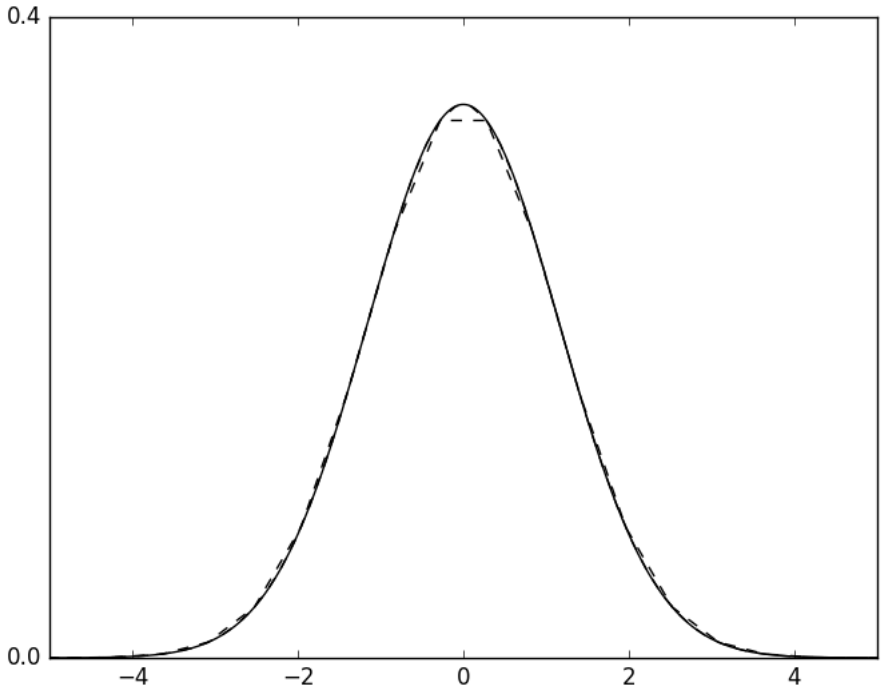
28. Daum F, Huang J. Curse of dimensionality and particle filters. *2003 IEEE Aerospace Conference Proceedings.* 2003; 4: 4-1979–4-1993.

29. Schervish MJ. *Theory of statistics*. Springer Science & Business Media . 2012

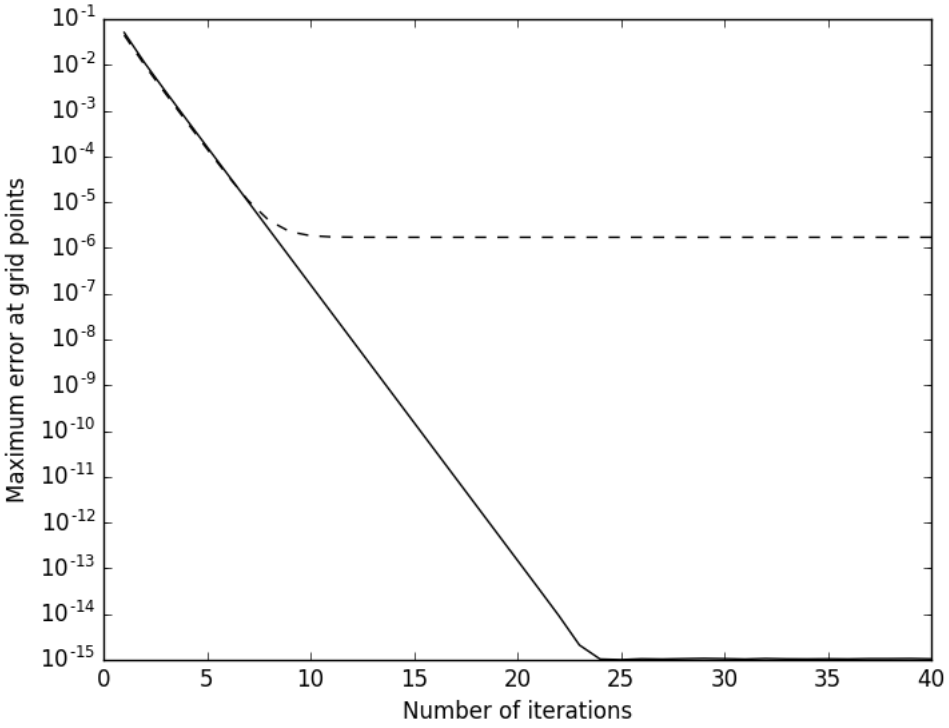
30. Bishop G, Welch G. An introduction to the Kalman filter. Online at http://www.cs.unc.edu/~welch/kalman/kalmanIntro.html; .



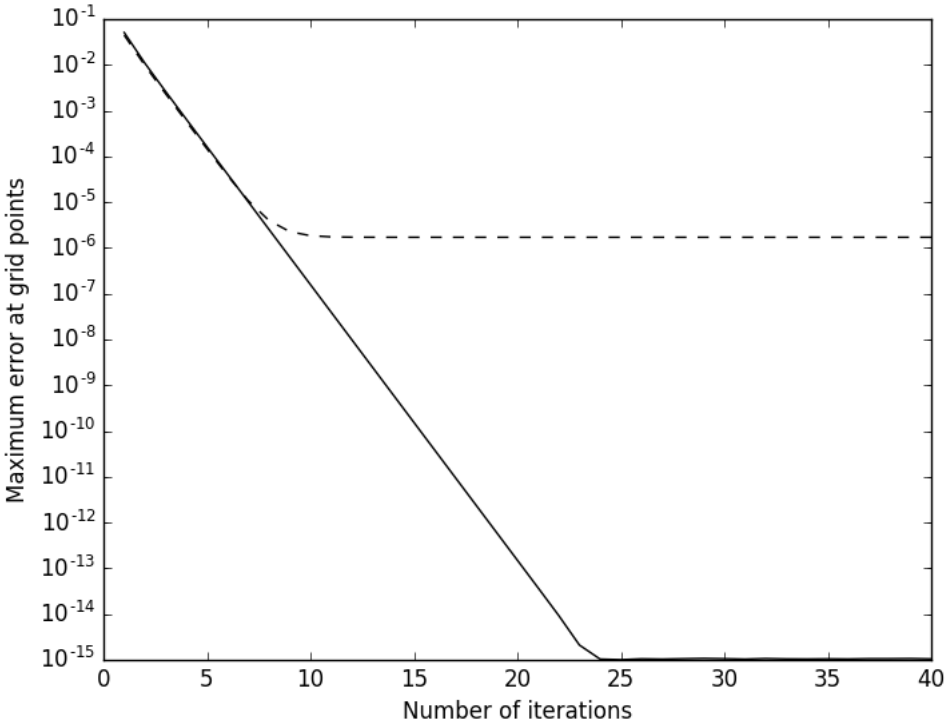
burn_comparison.tif



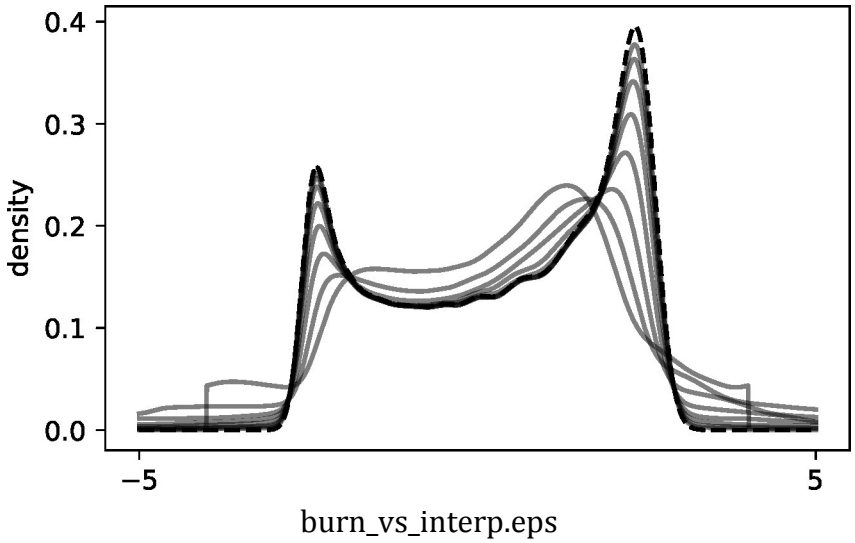
burn_comparison.tif

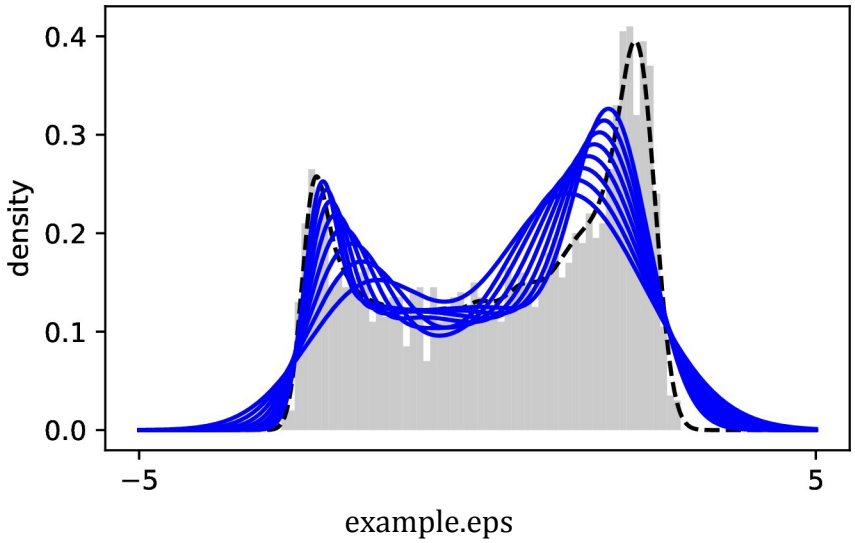


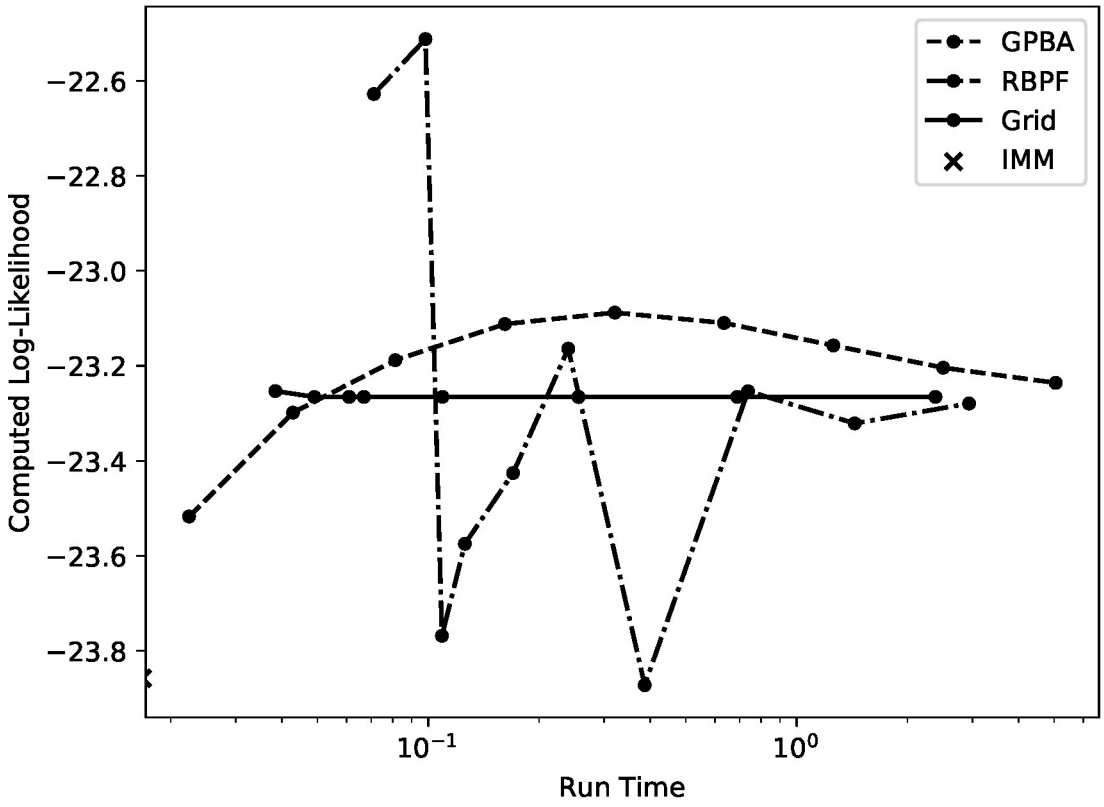
burn_convergence.tif



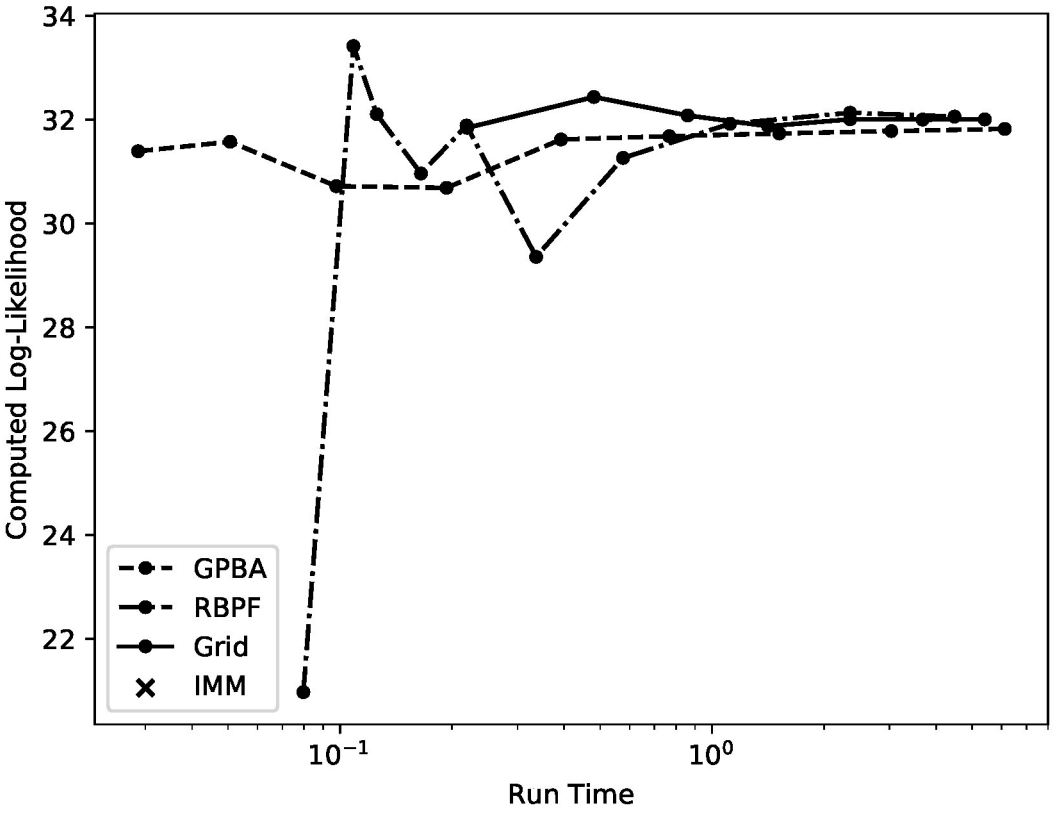
burn_convergence.tif



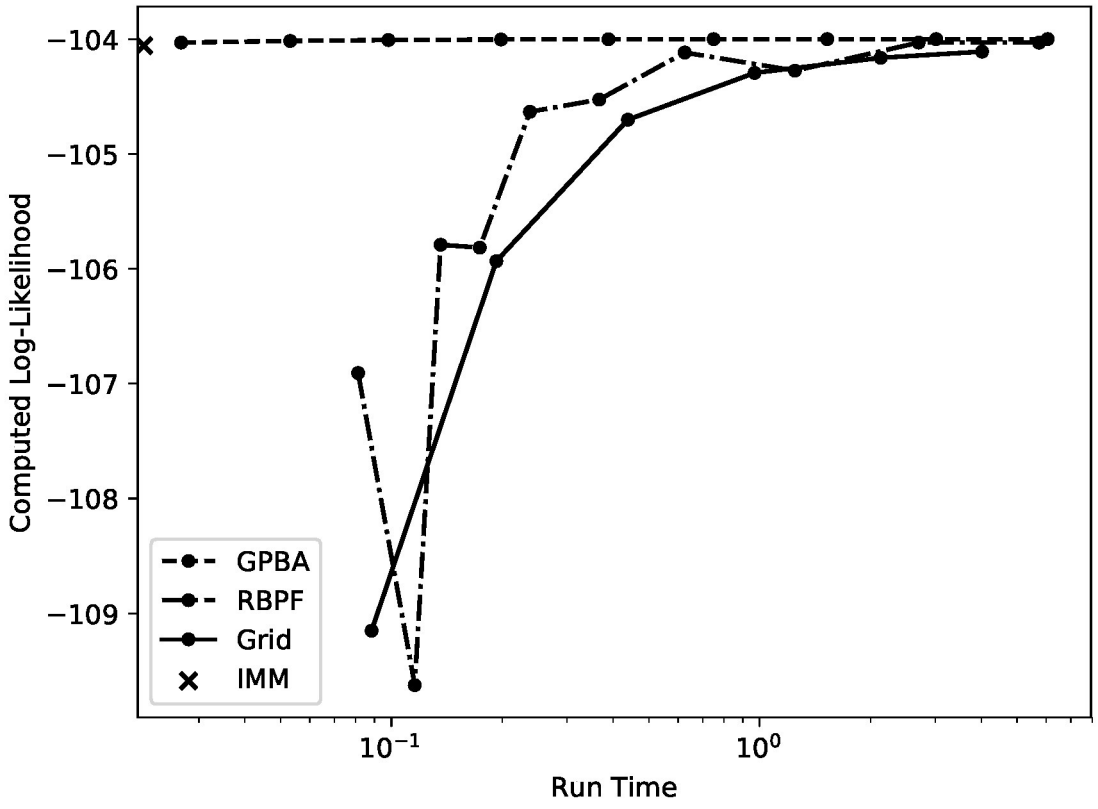




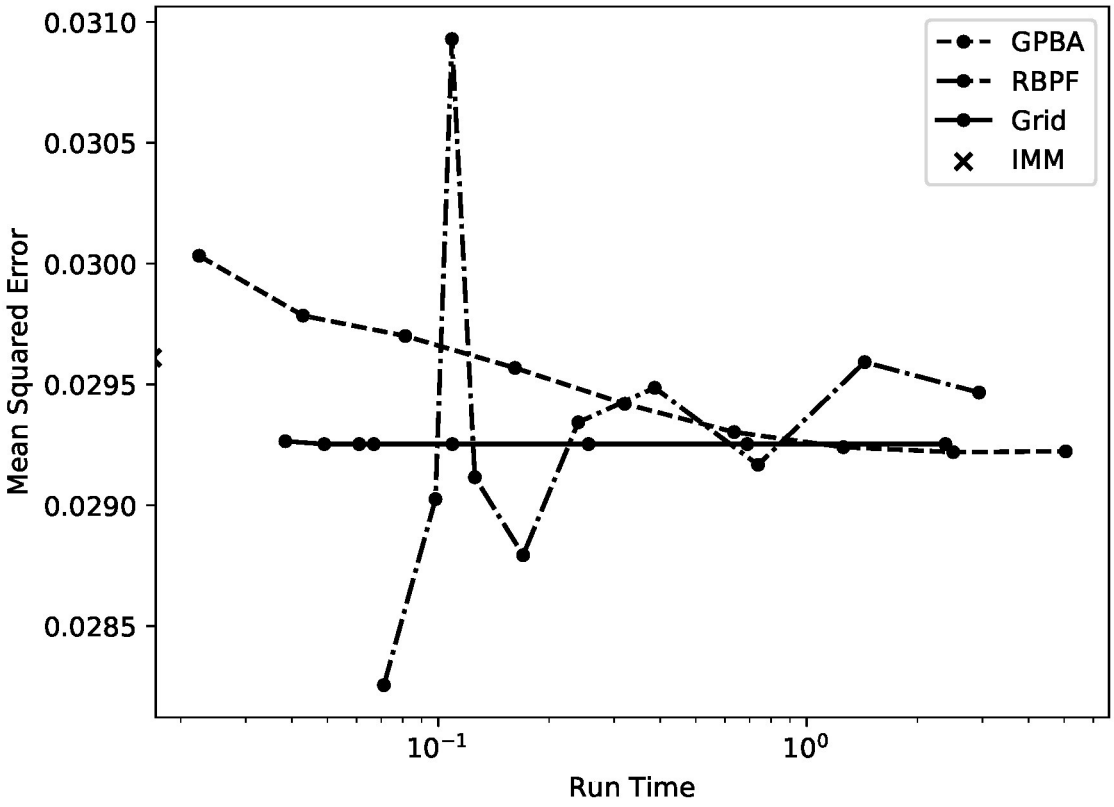
logprob_1d.eps



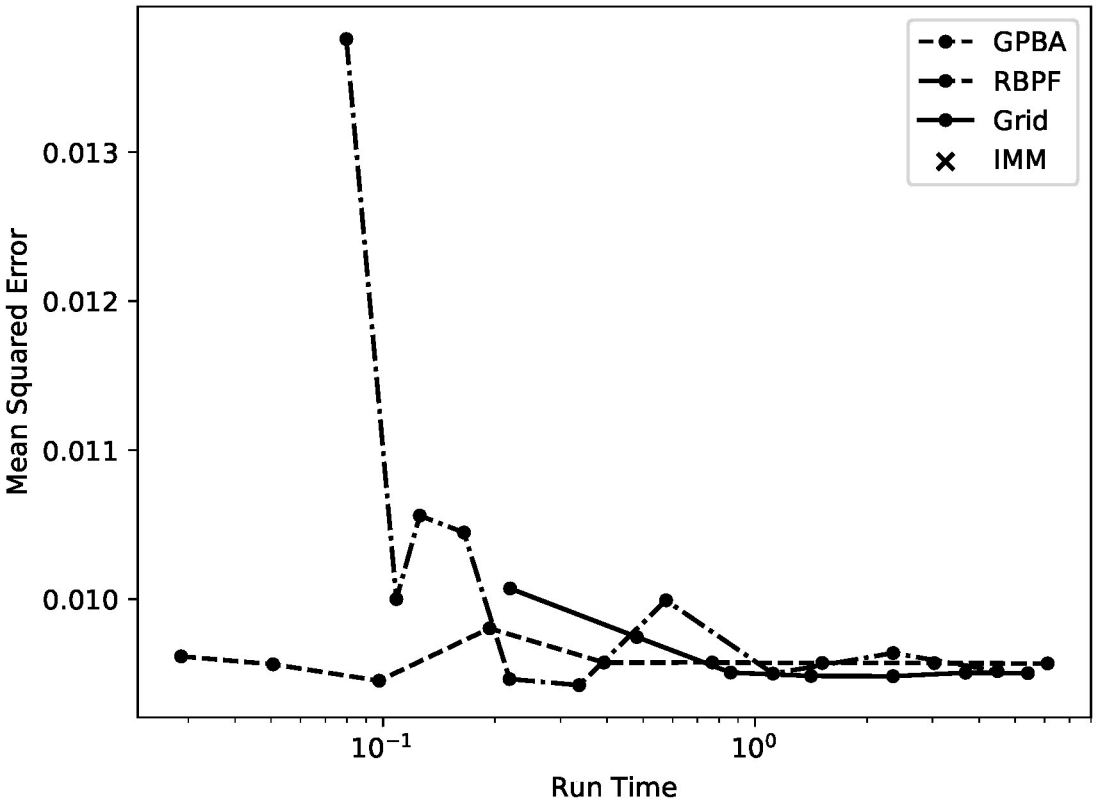
logprob_2d.eps



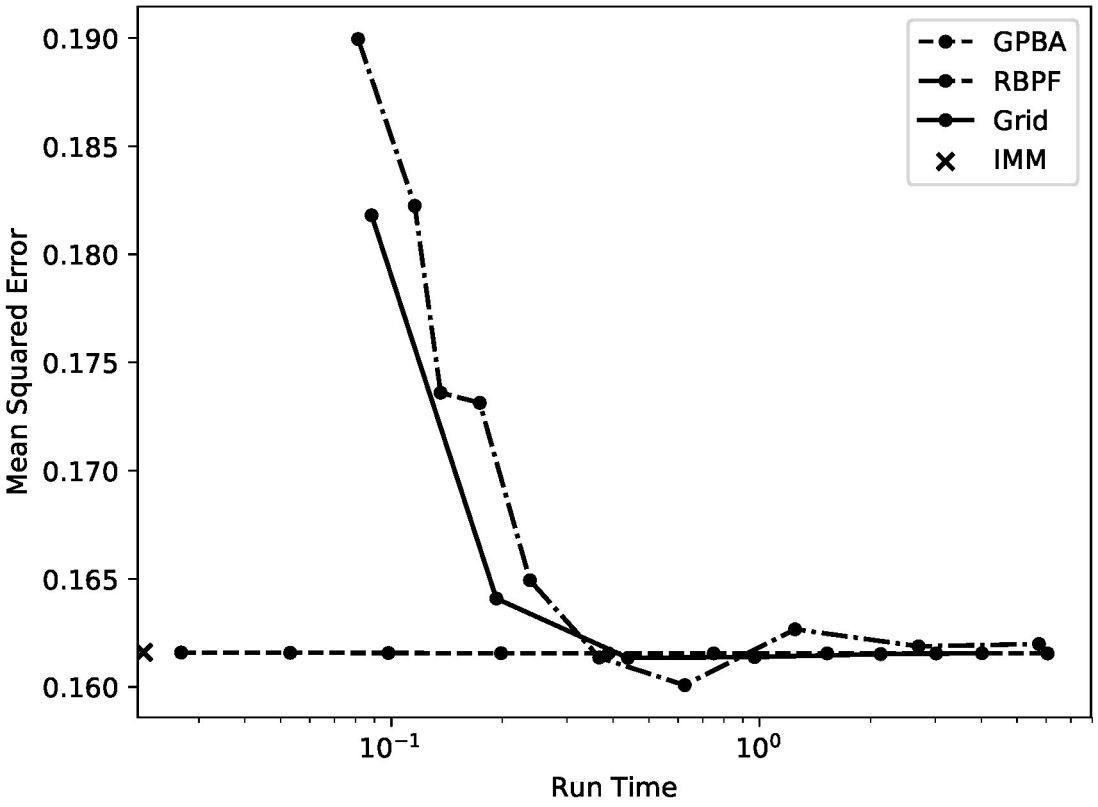
logprob_3d.eps



mse_1d.eps



mse_2d.eps



mse_3d.eps