

Sequential Bayesian Inference for Dynamic State Space Model Parameters

Arnab Bhattacharya and Simon Wilson

1. INTRODUCTION

Dynamic state-space models (Durbin and Koopman, 2001), consisting of an unknown state Markov process X_0, X_1, \dots and noisy observations of that process Y_1, Y_2, \dots that are conditionally independent, are used in a wide variety of applications e.g. wireless networks (Haykin et al., 2004), object tracking (Ristic et al., 2004) etc. The model is specified by an initial distribution $p(x_0|\theta)$, a transition kernel $p(x_t|x_{t-1}, \theta)$ and an observation distribution $p(y_t|x_t, \theta)$. These distributions are defined in terms of a collection of K static (e.g. non-time varying) parameters $\theta = (\theta_1, \dots, \theta_K)$. The joint model to time t is:

$$p(\mathbf{y}_{1:t}, \mathbf{x}_{0:t}, \theta) = \left(\prod_{j=1}^t p(y_j|x_j, \theta)p(x_j|x_{j-1}, \theta) \right) p(x_0|\theta)p(\theta), \quad (1.1)$$

where $\mathbf{y}_{1:t} = (y_1, \dots, y_t)$.

In this paper, we propose a new approach for approximating $p(\theta|\mathbf{y}_{1:t})$. The principal idea is to approximate the posterior on a carefully constructed grid to which points are added or deleted where necessary. Our method is used alongside a large range of state process estimation algorithms making it highly flexible, even for non-linear or non-Gaussian models. We provide evidence corresponding to our claims through a series of examples.

The literature has tended to focus on computation of the predictive distributions $p(x_t|\mathbf{y}_{1:t-1}, \theta)$ and $p(y_t|\mathbf{y}_{1:t-1}, \theta)$, and the filtering distributions $p(x_t|\mathbf{y}_{1:t}, \theta)$. Updating of these distributions is by the well-known forward equations (Arulampalam et al., 2002); for example for the prediction of state process we need to compute

$$p(x_t|\mathbf{y}_{1:t-1}, \theta) = \int p(x_t|x_{t-1}, \theta)p(x_{t-1}|\mathbf{y}_{1:t-1}, \theta) dx_{t-1}. \quad (1.2)$$

For the linear Gaussian case and assuming known parameters, these computations reduce to the closed form of the Kalman filter (Kalman, 1960). Generally, Gaussian forms of the model allows filtering and prediction to be done quickly, exactly and sequentially, without the need to store the data sequence (West and Harrison, 1997). Exact inference has also been achieved for models where the state process is discrete (Cappé et al., 2005).

Extending inference to non-linear and/or non-Gaussian models has proved to be challenging since analytical solutions to the above integrals do not exist. Functional approximation approaches, derived from the Kalman filter, such as the extended Kalman filter (Haykin, 2001), unscented Kalman filter (Julier and Uhlmann, 1997) etc have been proposed. Monte Carlo approaches such as the bootstrap filter (Gordon et al., 1993) and auxiliary particle filter (Pitt and Shephard, 1999) have also been widely used. Attempts have also been made to combine sampling based methods with functional approximations (Merwe et al., 2001).

As regards the static parameter estimation problem, few closed form solutions are available. Different approaches are employed, often involving joint inference of static parameters and the state process (Kantas et al., 2015). We will discuss some of the existing methodologies in the next section.

Section 2. is a review of some of the online static parameter estimation methods, some of which are used later for comparison. Section 3. outlines the principle of the method. Sections 4. and 5. describe the two main issues to be resolved in order to implement the method: approximations to one-step ahead filtering and prediction densities, and updating the grid defining posterior density. A discussion on the theoretical aspects of error accumulation is provided in section 6. Section 7. illustrates the method and assesses its performance against alternative approaches. Section 8. contains some concluding remarks.

2. ONLINE PARAMETER INFERENCE REVIEW

In this paper, interest is in online inference of the the posterior distribution $p(\theta|\mathbf{y}_{1:t})$, along with the filtering and prediction densities associated with the state process (and not on the joint density); a recent review is Kantas et al. (2015). Several approaches have been proposed in the literature:

Joint KF/EKF/UKF: A common practice in the engineering literature is to add dynamics to static parameters, such as assuming $\theta_t \sim N(\theta_{t-1}, V_t)$ with variance V_t decreasing with t , and make inference on $p(x_t, \theta_t|\mathbf{y}_{1:t-1})$ at each time point using a

single Kalman filter or extensions thereof. One advantage of this is that the state and parameters are typically correlated *a posteriori*, even in linear systems (Haykin, 2001); however this is known to suffer from numerical instability issues.

Dual KF/EKF/UKF: Dual filters assume that the state and the parameter have separate state space representations, and thus two filters can be run concurrently (Wan and Nelson, 1997). The prediction $p(x_t|\mathbf{y}_{1:t-1}, \hat{\theta}_{\text{old}})$ is derived using the parameter mean and $p(\theta_t|\mathbf{y}_{1:t-1}, \hat{x}_{\text{old}})$ is updated using the filtered state mean. Because each posterior only uses the first moment of the other posterior and ignores the variance, these methods are known to produce low variance posteriors.

Online gradient method: Sequential optimization of $\log(p(\mathbf{y}_{1:t}|\theta))$ is also possible. If $\hat{\theta}_{t-1}$ is the estimate after the first $t-1$ observations, it is updated to $\hat{\theta}_t$ after receiving a new data y_t (Poyiadjis et al., 2005). The problem of evaluating the gradient for the whole data $\mathbf{y}_{1:t}$ has been bypassed in the case of hidden Markov models (LeGland and Mever, 1997). A problem associated with any gradient method is that it is extremely sensitive to initialization and may converge to a local maximum.

Online EM: Online versions of the EM algorithm (Dempster et al., 1977), suitable for a dynamic state space model with unknown model parameters, have been proposed in Cappé (2011). A major advantage is that it always attempts to maximize the likelihood, allowing methods such as variational inference to be used to estimate the parameters. However this method can also converge to a local maximum.

Liu and West filter: This is the most generic among particle filter that performs dual estimation of the state and parameters. Artificial dynamics are introduced for the parameter and subsequently a kernel density estimate of $p(\theta|\mathbf{y}_{1:T})$ is proposed from which θ can be sampled. Shrinkage is introduced to control for over-dispersion in the kernel density function. A major drawback though is that it requires a significant amount of tuning for quantities such as the kernel bandwidth.

Storvik's filter: Storvik (2002) generates particles from the parameter's posterior distribution without assuming any associated random drift. It is further assumed that the posterior distribution of θ depends on a low-dimensional set of sufficient statistics that can be efficiently updated for each t . The choice of this set is the biggest stumbling block for this algorithm.

Particle learning: Carvalho et al. (2010) have provided a modified version of Storvik's filter which has proved to be more efficient. Sufficient statistics are derived for the class of conditional dynamic linear models (CDLM), thus providing additional structure to the algorithm. Further, the position of the resampling step in Storvik's algorithm is now interchanged with the propagation step, that allows particle deficiency of the posterior of θ to be reduced.

For large t , particle degeneracy for sampling filters is difficult to avoid, even where

artificial drift has been added (Andrieu et al., 2005; Kantas et al., 2015). Online EM and gradient ascent methods do not have this problem. The method in this paper has the same advantage as that of the latter approaches, making it suitable for long sequences of data. It is to be kept in mind that ours is an on-line sequential algorithm and the computational load does not increase over time, unlike other generic methods like SMC² (Chopin et al., 2013).

3. PRINCIPLE

The principle of the proposed method is based on two observations.

The first observation is that many dynamic state space models have a relatively small number of static parameters, so that in principle $p(\theta | \mathbf{y}_{1:t})$ can be computed and stored on a discrete grid of practical size. This has been noted as a property of many latent models (Rue et al., 2009).

The second observation is that there are useful identities for parameter estimation in latent models. The one of interest here is a sequential version of the *basic marginal likelihood identity* or BMI (Chib, 1995),

$$p(\theta | \mathbf{y}_{1:t}) \propto p(\mathbf{y}_{1:t}, \mathbf{x}_{0:t}, \theta) / p(\mathbf{x}_{0:t} | \mathbf{y}_{1:t}, \theta) |_{\mathbf{x}_{0:t} = \mathbf{x}^*(\theta)}, \quad (3.1)$$

that is valid for any $\mathbf{x}_{0:t} = \mathbf{x}^*(\theta)$ for which $p(\mathbf{x}_{0:t} | \mathbf{y}_{1:t}, \theta) > 0$ and forms the basis for the integrated nested Laplace approximation (INLA); see Rue et al. (2009).

The sequential version of the BMI, which lies at the heart of this approach, is:

$$p(\theta | \mathbf{y}_{1:t}) \propto p(\theta | \mathbf{y}_{1:t-1}) p(y_t | x_t, \theta) p(x_t | \mathbf{y}_{1:t-1}, \theta) / p(x_t | \mathbf{y}_{1:t}, \theta) |_{x_t = x^*(\theta)}. \quad (3.2)$$

Typically $x^*(\theta) = \arg \max_{x_t} p(x_t | \mathbf{y}_{1:t}, \theta)$ although the identity is valid for any x_t where $p(x_t | \mathbf{y}_{1:t}, \theta) > 0$. Equation 3.2 is useful for sequential estimation because its computation does not grow with t , as is the case with Equation 3.1. Taking Equation 3.2, then if prediction and filtering approximations $\tilde{p}(x_t | \mathbf{y}_{1:t-1}, \theta)$ and $\tilde{p}(x_t | \mathbf{y}_{1:t}, \theta)$ are available, any approximation $\tilde{p}(\theta | \mathbf{y}_{1:t-1})$ can be updated:

$$\tilde{p}(\theta | \mathbf{y}_{1:t}) \propto \tilde{p}(\theta | \mathbf{y}_{1:t-1}) p(y_t | x_t, \theta) \tilde{p}(x_t | \mathbf{y}_{1:t-1}, \theta) / \tilde{p}(x_t | \mathbf{y}_{1:t}, \theta) |_{x_t = x^*(\theta)}. \quad (3.3)$$

Assuming that θ of low dimension, computing Equation 3.3 on a discrete grid offers the potential for fast sequential estimation.

Alternatively, the sequential version of Bayes' theorem,

$$p(\theta | \mathbf{y}_{1:t}) \propto p(\theta | \mathbf{y}_{1:t-1}) p(y_t | \mathbf{y}_{1:t-1}, \theta). \quad (3.4)$$

and the availability of an approximation $\tilde{p}(y_t|\mathbf{y}_{1:t-1}, \theta)$ from any of the filters mentioned, also allows for updating. Furthermore, using $p(y_t|\mathbf{y}_{1:t-1}, \theta)$, one can compute the marginal log likelihood $\log(p(\mathbf{y}_{1:t})) = \sum_{j=1}^t \log(p(y_j | \mathbf{y}_{1:j-1}))$, where

$$p(y_j|\mathbf{y}_{1:j-1}) = \int p(y_j|\mathbf{y}_{1:j-1}, \theta) p(\theta|\mathbf{y}_{1:j}) d\theta \approx \int \tilde{p}(y_j|\mathbf{y}_{1:j-1}, \theta) \tilde{p}(\theta|\mathbf{y}_{1:j}) d\theta. \quad (3.5)$$

The integration in Equation 3.5 is approximated quickly by a sum over the grid of values of θ .

This suggests the following sequential estimation algorithm when an approximate posterior predictive distribution, or just state prediction and filtering distributions, are available. The process is initiated by choosing a prior $p(\theta)$ and then deterministically selecting a discrete set of points, say Θ_0 , to approximately cover the region of high density of the prior, and compute $p(\theta)$ at each of the selected points. The posterior of θ is updated at each point $\theta_{ti} \in \Theta_t$ by Equation 3.3 or 3.4. We would like to reiterate that the objective of this study is not to track the complete trajectory of the state process, but rather to update the posterior parameter and state-filtering density over time.

In this method, $p(\theta | \mathbf{y}_{1:t})$ may also be approximated by INLA up to some specified time $t = T_{\text{INLA}}$, and from then on the sequential update of Equation 3.3 or 3.4 is used. This can be particularly helpful because INLA produces a grid Θ_t that usually provides good support. The value of T_{INLA} is determined by how quickly the update is required, given that the INLA computation slows down with t . Note that T_{INLA} can be equal to 0, i.e. INLA is not used at all; this means the starting grid is constructed on the prior $p(\theta)$.

A few issues remain to be addressed in order to implement this algorithm: the form of the approximations $\tilde{p}(y_t|\mathbf{y}_{1:t-1}, \theta)$ or that of $\tilde{p}(x_t|\mathbf{y}_{1:t-1}, \theta)$ and $\tilde{p}(x_t|\mathbf{y}_{1:t}, \theta)$ and how to adapt the grid Θ_t so that it tracks the support of the posterior over time. These issues are addressed in the next sections.

4. POSTERIOR PREDICTIVE AND FILTERING DENSITY APPROXIMATIONS

For the Kalman filter, all the distributions required in Equations 3.2 and 3.4 are Gaussian, producing exact updates. The means and variances of these Gaussians are sequentially updated. West and Harrison (1997) is a comprehensive study of models of this type.

There is a large literature about more general filtering algorithms; see Doucet et al. (2001). In this paper, we have used functional approximations of the filtering distributions based on extensions to the Kalman filter that incorporate non-linearity and non-Gaussian error while holding on to the basic principle of updating the first two moments. The extended Kalman filter linearizes a non-linear model to create a Gaussian approximation to the filtering and prediction densities. The unscented Kalman filter (Julier and Uhlmann, 1997) also produces Gaussian approximations to the filtering and prediction densities but propagates the means and covariances through the non-linear function. It tends to be more accurate than the extended Kalman filter, more so for strongly non-linear models.

Other finite dimensional suboptimal filters have also been proposed. Some of them are the Gaussian sum filter (Ito and Xiong, 1999), quadrature filters (Davis and Rabinowitz, 2007) etc. In our examples, we have used the Kalman filter, the unscented Kalman filter and the quadrature filter.

5. UPDATING THE GRID

Recall that θ is of dimension K and assume that there is an initial grid available at time $t = T_{\text{INLA}}$. If INLA is not used then $T_{\text{INLA}} = 0$ and an initial grid Θ_0 is assumed that covers the region of high density of the prior. It is also assumed that the grid is made up of points along a set of basis vectors for \mathbb{R}^K , so that it can be written as the Cartesian product

$$\Theta_t = \Theta_{t1} \times \Theta_{t2} \times \cdots \times \Theta_{tK}, \quad (5.1)$$

where each $\Theta_{tk} = \{\theta_{tk1}, \dots, \theta_{tkn_{tk}}\}$ is a set of n_{tk} ordered values for the k th component of θ .

5.1 Checking the Grid

For $t > T_{\text{INLA}}$, the grid is checked every T_{update} observations to see if it is still an appropriate support for $\tilde{p}(\theta | \mathbf{y}_{1:t})$.

The following simple approach appears to perform well in the examples of the next section. First, the basis that represents the coordinate system for θ is not changed; it is assumed that by $t = T_{\text{start}}$ there is a basis that will be satisfactory for all future t . Second, changes to the grid are based on the univariate marginals of the

posterior along the grid coordinates. For the k th coordinate, the marginal is:

$$\tilde{p}_k(\theta_{tki}|\mathbf{y}_{1:t}) = \sum_{\theta \in \Theta_{t,-k,i}} \tilde{p}(\theta|\mathbf{y}_{1:t}) \left(\prod_{\substack{l=1 \\ l \neq k}}^K \delta(\theta_{tl}) \right), \quad i = 1, \dots, n_{tk},$$

where, to give the appropriate marginalizing sum, $\Theta_{t,-k,i}$ is Θ_t with $\Theta_{tk} = \theta_{tki} \mathbf{1}_{n_k}$, summation is over all coordinates except the k th, and $\delta(\theta_{tl})$ is the step size on the grid at θ along the l th coordinate at time t .

Changes are considered when the support of the marginals is not adequately covered by the grid in one of the following situations:

1. If $\tilde{p}_k(\theta_{tk1}|\mathbf{y}_{1:t})$ or $\tilde{p}_k(\theta_{tkn_{tk}}|\mathbf{y}_{1:t})$ (i.e. the values at each end of the grid support) are greater than a certain proportion, say $\delta_{\text{ext}}^{\text{add}}$ (chosen to be between 0.15 and 0.3) of the value at the mode of the marginal $\tilde{p}_k(\theta|\mathbf{y}_{1:t})$ then an extra point will be added at that end in $\Theta_{t+1,k}$ e.g. at $\theta_{nk1} - \delta_k$ and/or $\theta_{nk n_{tk}} + \delta_k$, where δ_k is a step size along coordinate k . Either addition has the effect, through Equation 5.1, of adding a $K - 1$ dimensional hyperplane to Θ_{t+1} . This is referred to as an external point addition.
2. Similarly, if either edge value $\tilde{p}_k(\theta_{tk1}|\mathbf{y}_{1:t})$ or $\tilde{p}_k(\theta_{tkn_{tk}}|\mathbf{y}_{1:t})$ is less than a certain small proportion, say $\delta_{\text{ext}}^{\text{drop}}$ (chosen to be small, say 0.001) of the marginal mode then either point is removed for $\Theta_{t+1,k}$. Either deletion removes a $K - 1$ dimensional hyperplane from Θ_{t+1} . The resulting new edge points in $\Theta_{t+1,k}$ are then also checked and removed if they also satisfy the criterion, allowing more than 1 point to be removed from each end of each Θ_{tk} . Internal grid points are not dropped in this approach. This is referred to as external deletion.
3. If the change in marginal between two existing points, relative to its modal value,

$$\left| \frac{\tilde{p}(\theta_{tki}|\mathbf{y}_{1:t}) - \tilde{p}(\theta_{tk,i-1}|\mathbf{y}_{1:t})}{\max_{m=1}^{n_{tk}} \tilde{p}(\theta_{tkm}|\mathbf{y}_{1:t})} \right|,$$

is larger than a certain threshold value, say $\delta_{\text{int}}^{\text{add}}$ (usually chosen between 0.3 and 0.4) then an extra point is added to $\Theta_{t+1,k}$ at $(\theta_{tki} + \theta_{tk,i-1})/2$. This adds a $K - 1$ dimensional hyperplane to Θ_{t+1} . This is called internal point addition.

Figure 1 shows examples of these additions and deletions. Note that it is possible to increase or decrease the thresholds over time and can be tweaked by the user in advance; for example keeping in mind the fact that the posterior shrinks over time one can assign a lower value to $\delta_{\text{int}}^{\text{add}}$ later in time such that more points are added to the region of high density.

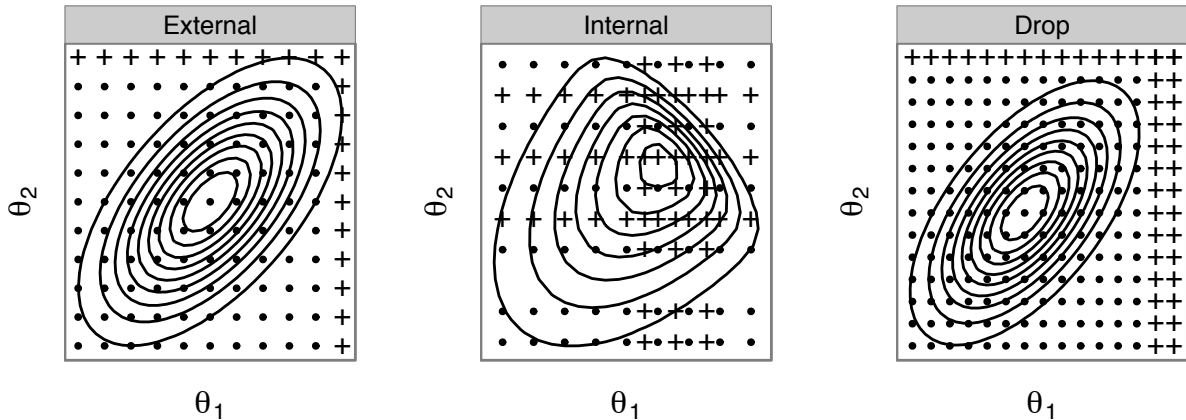


Figure 1: Addition and deletion from the grid, where \bullet denotes points that are not changed and $+$ denotes points that are added or deleted. From left to right: external addition, internal addition and external deletion.

5.2 Computing $\tilde{p}(\theta|\mathbf{y}_{1:t})$ at added points in Θ_t

For a set of newly added points to the grid, the most obvious next step is to compute $\tilde{p}(\theta|\mathbf{y})$ at any θ that has been added from Θ_{t-1} . However, this is computationally infeasible as t gets large and furthermore it would remove the property of the algorithm that it does not require the entire filtering sequence to be rerun from scratch.

Instead, extrapolation or interpolation methods are used to compute a value of $\tilde{p}(\theta|\mathbf{y})$ at any added external and internal point respectively from the existing values of \tilde{p} . Several methods exist e.g. kriging (Cressie, 1993), thin plate splines (Wahba, 1990) etc. We use tensor product spline surfaces (Schumaker, 2007), more specifically tensor product linear interpolation on the log scale. In our experience so far, this method is both fast and produces good values for new $\tilde{p}(\theta|\mathbf{y})$.

Tensor product linear interpolation has also been used to compute the moments of the state estimation filter corresponding to the added external or internal grid points, which in turn provides us with necessary filtering densities such as $p(x_t|\mathbf{y}_{1:t}, \theta)$, $p(x_t|\mathbf{y}_{1:t-1}, \theta)$ and $p(y_t|\mathbf{y}_{1:t-1}, \theta)$. For such points, this simple method works in most cases since we are essentially providing an informative starting value for moments in the Kalman filter framework, even in multivariate settings. Table 1 summarises the complete algorithm.

Table 1: The sequential parameter estimation algorithm.

```

specify  $T_{\text{INLA}}, T_{\text{update}};$ 
 $t = 0;$ 
if  $T_{\text{INLA}} == 0:$ 
    specify grid  $\Theta_0$  to cover high-density-region of  $p(\theta);$ 
repeat:
    if  $t \leq T_{\text{INLA}}:$ 
        compute discrete grid  $\Theta_t$ , i.e.  $\tilde{p}(\theta|\mathbf{y}_{1:t})$  for  $\theta \in \Theta_t$  using INLA;
    else:
        if  $(t - T_{\text{INLA}}) \bmod T_{\text{update}} == 0:$ 
            update grid  $\Theta_t$  using method of Section 5.;
        else:  $\Theta_t = \Theta_{t-1};$ 
        for each  $\theta \in \Theta_t:$ 
            compute  $\tilde{p}(y_t|\mathbf{y}_{1:t-1}, \theta)$ , or  $\tilde{p}(x_t|\mathbf{y}_{1:t-1}, \theta)$  and  $\tilde{p}(x_t|\mathbf{y}_{1:t}, \theta)$ , using
filtering;
            update  $\tilde{p}(\theta|\mathbf{y}_{1:t})$  from  $\tilde{p}(\theta|\mathbf{y}_{1:t-1})$  by Equation 3.3 or 3.4.

```

6. APPROXIMATION ERROR

Theoretical results about the error in approximating $p(\theta | y_{1:t})$ using this approach can be divided amongst the various aspects of the approximation:

1. Error from using a discrete grid;
2. Convergence of the filtering algorithm, including state filtering;
3. Interpolation error in relation to adding grid points.

In this section, existing results around each of these 3 aspects are discussed and a roadmap to proving error bounds is laid out.

It is noted that the unnormalised $\tilde{p}(\theta|\mathbf{y}_{1:t})$ evolves independently at each point in the grid, except for the case of points that are added, where the initial value of $\tilde{p}(\theta|\mathbf{y}_{1:t})$ is an interpolation of neighbouring values. Convergence results for the polynomial interpolation that is used here are well documented (Kress, 1998). For densities with uniformly bounded derivatives, we can ensure that the interpolation is arbitrarily accurate by increasing the order of the polynomial. However, for the rest of this section, we consider a simpler case where adding/dropping points is not permitted, allowing the third source of error to be ignored and making it sufficient to consider the error in $\tilde{p}(\theta|\mathbf{y}_{1:t})$ at one value $\theta \in \Theta_t$ only, and then separately the consequences of that for errors related to the use of a discrete grid.

For the discretization error, were the approximation to be exact on the grid e.g. $\tilde{p}(\theta|\mathbf{y}_{1:t}) = p(\theta|\mathbf{y}_{1:t}), \forall \theta \in \Theta_t$ then the standard theory of Riemann integration gives us that all integrals of $p(\theta|\mathbf{y}_{1:t})$, at least over compact sets, will be approximated arbitrarily well by finite difference summations of $\tilde{p}(\theta|\mathbf{y}_{1:t})$ as the grid density increases, including over irregular grids.

Therefore the key to results about bounding the error in the approximation of $p(\theta|\mathbf{y}_{1:t})$ by $\tilde{p}(\theta|\mathbf{y}_{1:t})$ lies in the choice of starting values (that come from the prior), error from the use of approximate filters and how that error propagates through to $\tilde{p}(\theta|\mathbf{y}_{1:t})$ by repeated application of Equations 3.3 or 3.4 at a single grid point. These results are the most challenging of the 3 error sources to discuss.

At time t , given that our algorithm provides the following filtering distribution

$$\tilde{p}(\theta|y_{1:t}) = C\tilde{p}(\theta) \prod_i \tilde{p}(y_i|y_{1:i-1}, \theta),$$

where C is the accumulated constant of proportionality, the total variation norm between the log of the approximation and the actual data generating process is the following (see Appendix for details):

$$|\log p(\theta|y_{1:t}) - \log \tilde{p}(\theta|y_{1:t})| \leq |\log p(\theta) - \log \tilde{p}(\theta)| + \sum_i |\log p(y_i|y_{1:i-1}, \theta) - \log \tilde{p}(y_i|y_{1:i-1}, \theta)|.$$

Given that the error is bounded by a sum that is infinite in the limit, it seems that at best it will be possible to show that the error $|\log p(\theta|y_{1:t}) - \log \tilde{p}(\theta|y_{1:t})|$ is bounded rather than converging to 0.

Convergence of $\log \tilde{p}(\theta|y_{1:t})$ to $\log p(\theta|y_{1:t})$ can possibly be achieved under the following relatively strong conditions:

- $\tilde{p}(y_t|y_{1:t-1}, \theta)$ converges to $p(y_t|y_{1:t-1}, \theta)$ as $t \rightarrow \infty$ for each θ .
- Further, assuming that both $\tilde{p}(y_t|y_{1:t-1}, \theta)$ and $p(y_t|y_{1:t-1}, \theta)$ are bounded away from both 0 and ∞ on a compact subset of \mathbb{R}^d in the limit is then sufficient for convergence between $\log \tilde{p}(y_t|y_{1:t-1}, \theta)$ and $\log p(y_t|y_{1:t-1}, \theta)$.
- Convergence need to be sufficiently fast between $\log \tilde{p}(y_t|y_{1:t-1}, \theta)$ and $\log p(y_t|y_{1:t-1}, \theta)$ such that $\sum_i |\log p(y_i|y_{1:i-1}, \theta) - \log \tilde{p}(y_i|y_{1:i-1}, \theta)|$ is bounded. For example, geometric convergence of $\log \tilde{p}(y_t|y_{1:t-1}, \theta)$ to $\log p(y_t|y_{1:t-1}, \theta)$ will give a *bounded* error to $|\log p(\theta|y_{1:t}) - \log \tilde{p}(\theta|y_{1:t})|$ by the geometric sum formula.

Convergence results in the literature pertain to the state process x_t and hence to differences between $p(x_t|y_{1:t}, \theta)$ and $\tilde{p}(x_t|y_{1:t}, \theta)$. These results can easily be

extended to the prediction distribution $\tilde{p}(y_t|y_{1:t-1}, \theta)$. Focus is on the stability of the filter, so that the total variation norm

$$\|p(x_t | y_{1:t}, \theta) - \tilde{p}(x_t | y_{1:t}, \theta)\| \xrightarrow[t \rightarrow 0]{} 0.$$

In the simplest setting, convergence results are well researched on:

- For the linear filter, properties such as uniform *detectability* and uniform *stabilizability* provide the conditions for convergence (Anderson and Moore, 1981). Since for $\theta \in \Theta_t$, the Kalman filter converges for any $\tilde{p}(x_0)$, convergence for our algorithm would essentially depend only on $\tilde{p}(\theta)$, where $\tilde{p}(\cdot)$ represents the prior chosen by us, while $p(\cdot)$ is the prior in the true data generating process.
- Finite state space models provide simple settings to study convergence of the filter. Assuming that the actual data generating process arises from $p(\cdot)$, whereas we use $\tilde{p}(\cdot)$ as the prior; one can calculate $p(x_t | y_{1:t}, \theta)$ and $\tilde{p}(x_t | y_{1:t}, \theta)$ respectively. The total variation between $p(x_t | y_{1:t}, \theta)$ and $\tilde{p}(x_t | y_{1:t}, \theta)$ is bounded above under ergodic signal, bounded error density, conditions on initial values and certain “mixing” conditions (Atar and Zeitouni, 1997). The upper bound is dependent on the Hilbert metric between $\tilde{p}(x_0)$ and $p(x_0)$ and mixing parameters (see Appendix for more details). For our algorithm, this would mean that convergence of $p(\theta | y_{1:t})$ would again be dependent on our choice of the prior $\tilde{p}(\theta)$ along with the aforementioned conditions.

Convergence results for non-linear systems are much more difficult to compute; the existing proofs are based on assumptions which are very restrictive, especially in terms of practical application. An illustration of a potential approach in general nonlinear filtering is provided in Chigansky (2006), where convergence is proved along the same lines as that of finite state space models.

Most solutions to general state space filtering are sub-optimal, and initial estimate is found to play a very important role. Kluge et al. (2010) and others before them proved that estimation error in EKF is exponentially bounded in mean square under permitted nonlinearity and very conservative bounds for each of initial error and noise covariance matrix. Other filters like UKF or Gaussian filters require tuning of signal covariance matrices to achieve stability; no quantitative bounds have been computed and verification of existing assumptions prove to be exceptionally complex (Xiong et al., 2009). It is extremely difficult to generalise these results to our method, and hence pointwise convergence of $\tilde{p}(\theta | y_{1:t})$ would possibly depend on our choice of prior $\tilde{p}(\theta)$ along with the conditions associated with each filter, as well as those mentioned under general conditions.

A simulation study we have performed substantiates the claim above, vis a vis a linear Gaussian model. Both for state and parameter, priors have been constructed to be extremely narrow and the true starting values for the data generating process are “very distant” from the region of high density of these priors. Our filter was occasionally found to diverge under these circumstances. An ill-defined state process prior caused more issues than a parameter prior.

7. EXAMPLES AND COMPARISON WITH ALTERNATIVE APPROACHES

Our method has been implemented for four real data sets and its performance is compared to offline methods such as INLA, MCMC, iterated filtering (Ionides et al., 2011) and PMCMC (Andrieu et al., 2010), as well as the online method of Liu and West (henceforth referred to as L&W filter). Computation times have not been compared because of the difficulties in accounting for differences in implementation and code optimization that have been made across all of these methods, even given that they are all written in R. For INLA, the R-INLA package was used (see www.r-inla.org), for MCMC in stochastic volatility model the package `stochvol` was used (Kastner, 2016b) and for iterated filtering, PMCMC and the L&W filter, the R package `pomp` was used (King et al., 2010).

For all the examples the heuristic parameters are set at $\delta_{\text{ext}}^{\text{add}} = 0.2$, $\delta_{\text{ext}}^{\text{drop}} = 0.001$ and $\delta_{\text{int}}^{\text{add}} = 0.35$. The starting grid has been constructed by placing a regular grid on the region of high density of the prior.

7.1 Nile data

The local level model (West and Harrison, 1997) is a simple dlm of the form

$$y_t = x_{t-1} + \eta_t, \tag{7.1}$$

$$x_t = x_{t-1} + \epsilon_t, \tag{7.2}$$

where $\eta_t \sim N(0, \sigma_{\text{obs}}^2)$ and $\epsilon_t \sim N(0, \sigma_{\text{sys}}^2)$. Here it is fitted to the well-known Nile data set, being annual measurements of the river Nile at Ashwan from 1871 to 1970. Parameter inference is done for $\theta = (\log(\sigma_{\text{obs}}^2), \log(\sigma_{\text{sys}}^2))$. These data have been elaborately studied by others e.g. Cobb (1978), Durbin and Koopman (2001) etc. Maximum likelihood estimates are $\hat{\sigma}_{\text{obs}}^2 = 15100$ and $\hat{\sigma}_{\text{sys}}^2 = 1468$.

The sequential inference algorithm is implemented with a variety of informative and non-informative priors for θ and x_0 . The starting grid has been constructed on

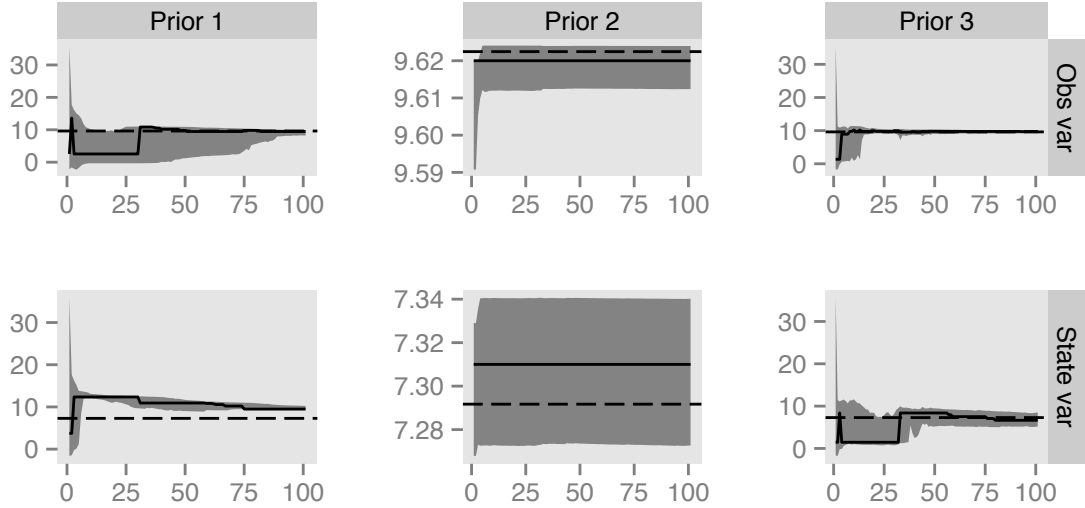


Figure 2: Trace plots of the median (solid line), 2.5% and 97.5% posterior quantiles (shaded) of the log variances. The MLE given all the data is a dashed line.

40 points in each dimension. Performance is evaluated qualitatively through trace plots of the posterior median. We would also like to mention at the outset that average runtime was about 3 secs.

Because of a lack of identifiability between σ_{obs}^2 and σ_{sys}^2 , there is significant sensitivity to the prior and outlier observations. This can cause the approximation to concentrate around a local mode in the posterior.

This is illustrated in Figure 2, where trace plots of parameters with three different prior distributions are given:

Prior 1: Flat log inverse gamma prior with shape = 0.1 and scale = 1 for both log variance parameters and an uninformative prior for x_0 , namely normal with mean 0 and variance 100;

Prior 2: An informative log inverse gamma prior for $\log(\sigma_{sys}^2)$ with shape = 4506 and scale = 6760490, and that for $\log(\sigma_{obs}^2)$ with shape = 45006 and scale = 675015000. Also an informative normal(1000, 100) for x_0 ;

Prior 3: Flat prior for $\log(\sigma_{sys}^2)$ and $\log(\sigma_{obs}^2)$ (same as Prior 1) and a prior for x_0 similar to Prior 2 but with larger variance (10000).

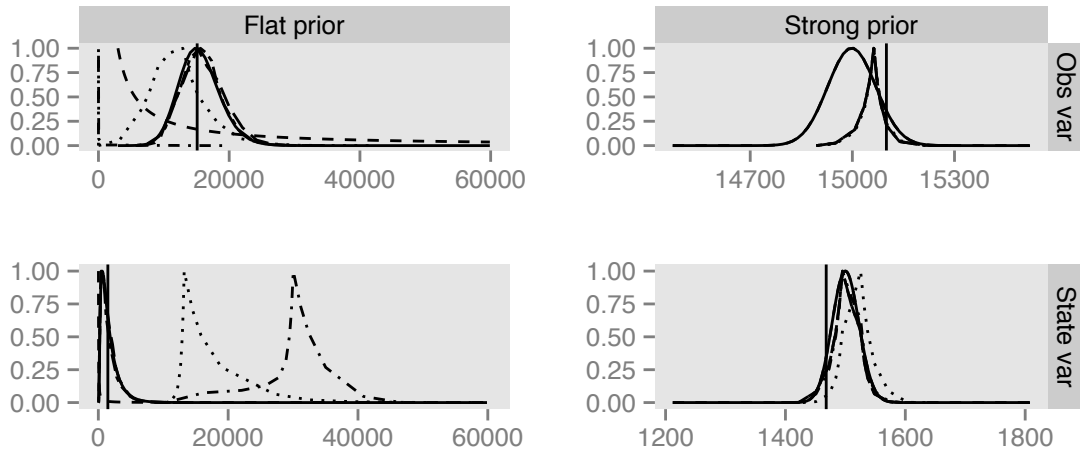


Figure 3: Prior density (dashed line) of θ over-layed on posterior density calculated exactly (solid line) and using our method (all other line types) over various combinations of priors on x_0 . These densities are scaled to fit. The vertical line is the MLE.

It can be seen that in this example, the prior on the state process (or equivalently the starting value for x_0) plays the most significant role. The marginal posterior distributions are largely in agreement with the MLE in spite of the prior influence, as is evident in Figure 3. It should be mentioned here that when we used INLA to provide the starting grid, our algorithm managed to identify the MLE even when the priors were uninformative for all unknowns (prior 1).

As regards model fit, Figure 4 plots the posterior predictions $p(y_{t+1}|\mathbf{y}_{1:t})$ with the data. Predictive performance is good and the effect of the prior is seen in generally tighter prediction intervals where the prior was informative.

It has already been mentioned that the maximum likelihood estimation of the unknown static parameters in this model are $\sigma_{obs}^2 = 15100$ and $\sigma_{sys}^2 = 1468$. The L&W filter was run 50 times for each of the 3 above mentioned priors. Table 2 shows the median of each of the 2.5%, 50% and 97.5% posterior quantiles at $t = 100$. The methods largely agree under priors 2 and 3, but L&W fails to track the MLEs in the uninformative prior case (prior 1).

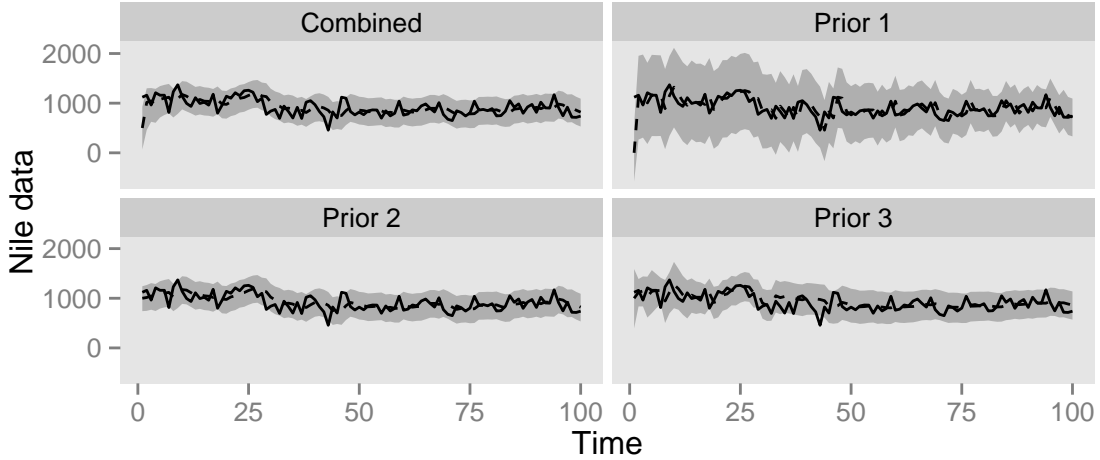


Figure 4: The one-step ahead predictive median (dashed line), 2.5% and 97.5% quantiles (shaded) along with data (solid line).

7.2 Tokyo rainfall data

Figure 5 displays the number of instances of rainfall over 1 mm in Tokyo for each day during the years 1983 – 1984 (Kitagawa, 1987). For each calendar day t , $t = 1, \dots, 366$ the data is defined as the following:

$$\begin{aligned}
 y_t &= 0 && \text{if no rain in either year,} \\
 &= 1 && \text{if rain in only 1 year and,} \\
 &= 2 && \text{if rain in both years,}
 \end{aligned}$$

with only 1 observation at $t = 60$, corresponding to 29th February 1984. These data have been analyzed before by several authors, for example Fahrmeir and Tutz (2001) and Rue and Held (2005).

Both Fahrmeir and Tutz (2001) and Kitagawa (1987) fitted the same model, in which the state process, which is related to the probability $\pi_t = p(\text{rain on day } t)$ through a logit link, is a random walk process. The model is given by,

$$y_t = \begin{cases} \text{Bin}(1, \pi_t), & t = 60 \text{ (February 29)} \\ \text{Bin}(2, \pi_t), & t \neq 60. \end{cases} \quad (7.3)$$

$$\pi_t = 1 / (1 + \exp(-x_t)), \quad (7.4)$$

$$x_t = x_{t-1} + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \quad (7.5)$$

	Sequential BMI		Liu & West	
	σ_{state}^2	σ_{obs}^2	σ_{state}^2	σ_{obs}^2
Prior 1	15253 (12847, 28340)	10863 (3608, 20325)	76302 (15041, 175546)	131621 (26393, 213275)
Prior 2	1494 (1440, 1541)	15061 (14948, 15123)	1499 (1457, 1548)	15002 (14863, 15137)
Prior 3	937 (181, 4832)	15361 (9530, 21165)	1121 (246, 4991)	15479 (10763, 23214)

Table 2: Comparison of the medians of 2.5%, 50% and 97.5% posterior quantiles for the Nile data between the method in this paper and that of L&W.

Rue and Held (2005) use a circular random walk of order 2 for the state process that cannot be directly analyzed using the “forward looking” state space approach employed here. Instead a second order difference random walk, also in Fahrmeir and Knorr-Held (2000), is used:

$$x_t = 2x_{t-1} + x_{t-2} + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \quad (7.6)$$

The state variance parameter σ^2 is the lone parameter in this problem. We have used EKF to compute the estimation in the state process as discussed extensively in Fahrmeir (1992). A flat prior for the state variance parameter is used and INLA is used for the first $T_{INLA} = 10$ iterations which is sufficient to get a grid with good posterior support. Average runtime for this data was approximately 30 secs.

Figure 6 shows the sequential parameter learning and comparison with ‘batch’ methods: INLA and an MCMC procedure. All 3 methods give a similar inference by the time that all the data are observed.

Figure 7 plots the smoothing mean $\hat{\pi}_t | \mathbf{y}_{1:366}, \hat{\sigma}_t^2$ along with bands denoting 2.5% and 97.5% quantiles. The smoothing mean seems to capture the “rainy patches” in Tokyo quite well and matches with similar inferences made in other works.

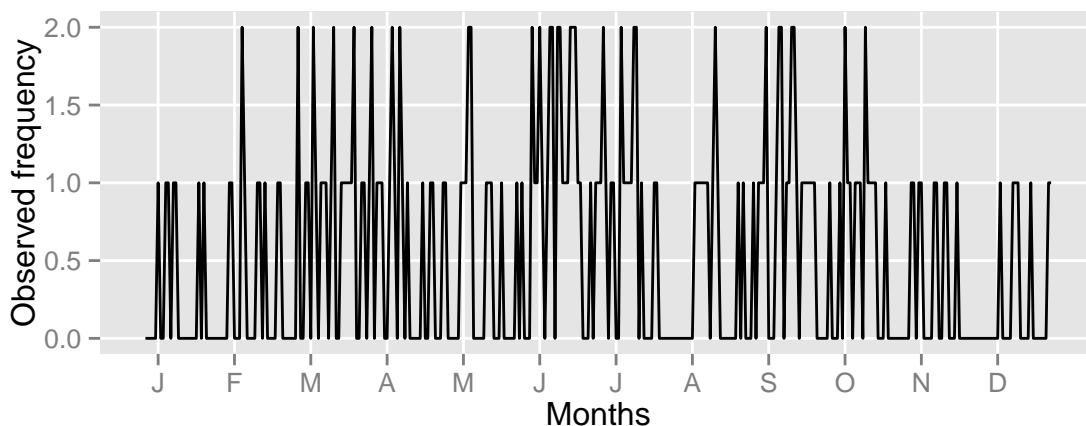


Figure 5: Instances of rainfall over 1 mm in Tokyo for each day from 1983 to 1984. The alphabets in the x-axis denote months of the year.

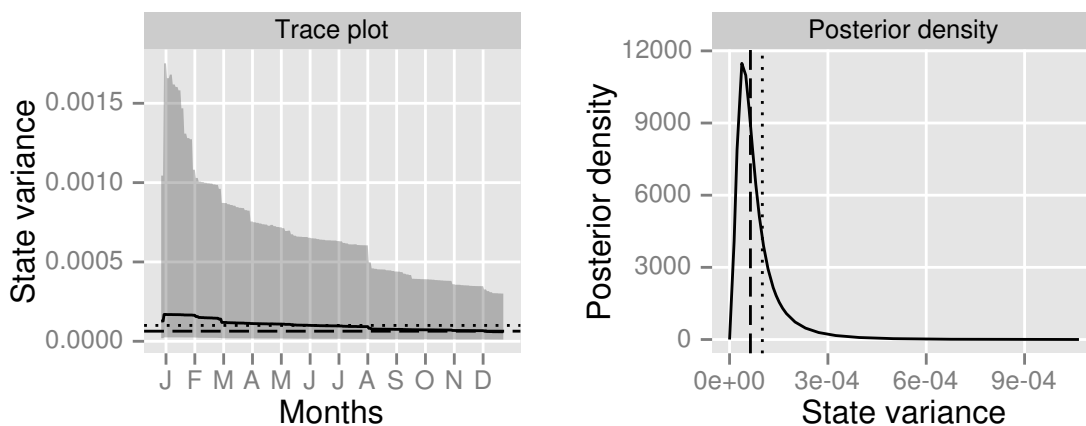


Figure 6: The posterior distribution of σ^2 and comparison to INLA and an MCMC procedure. Left is a summary of $p(\sigma^2 | \mathbf{y}_{1:t})$: posterior median (solid line), 2.5 and 97.5 percentiles (shaded), posterior mean given all the data from INLA (dashed line) and MCMC (dotted line). Right: final posterior $p(\sigma^2 | \mathbf{y}_{1:366})$.

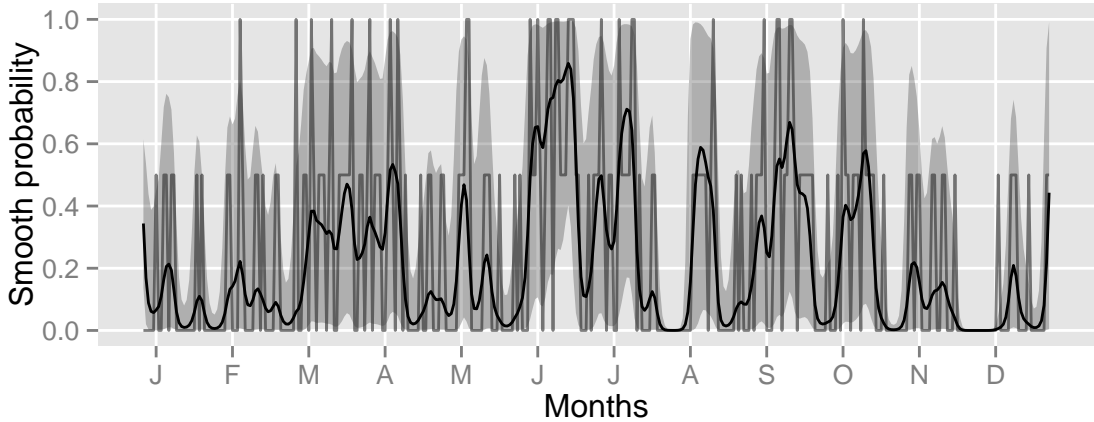


Figure 7: Summary of $p(\pi_t | \mathbf{y}_{1:366}, \hat{\sigma}_t^2)$ where $\hat{\sigma}_t^2$ is the posterior mode: posterior median (solid line) 2.5 and 97.5 percentiles (shaded region) and data (solid vertical lines).

7.3 Stochastic volatility model

The stochastic volatility model (Taylor, 1982) can be written:

$$y_t = \exp(x_t / 2) \epsilon_t, \quad (7.7)$$

$$x_t = \mu + \phi(x_{t-1} - \mu) + \sigma \eta_t, \quad (7.8)$$

$$x_t \sim N(\mu, \sigma^2 / (1 - \phi^2)),$$

where ϵ_t and η_t are independent standard Gaussian variables. The set of unknown static parameters in the model are $\theta = (\mu, \phi, \sigma^2)$: the *level* of log-variance μ , the *persistence* of log-variance ϕ and the *volatility* of log-variance σ .

The daily difference of the dollar-pound exchange rates between October 1st and June 28th, 1985, plotted in Figure 8, are analysed with this model.

The algorithm was applied to these data with a set of 32 different priors, corresponding to a mixture of informative and uninformative cases. The starting grids for each prior were 40 for each parameter.

Medians of the posterior quantiles for all the parameters, across the 32 prior specifications, along with 1-step ahead predictions, are plotted in Figure 9, along with the posterior mean from an analysis of all the data using the MCMC approach of Kastner (2016a) and INLA. It can be seen that there is good robustness to the prior specification, at least once most of the data are analysed. A much larger estimate

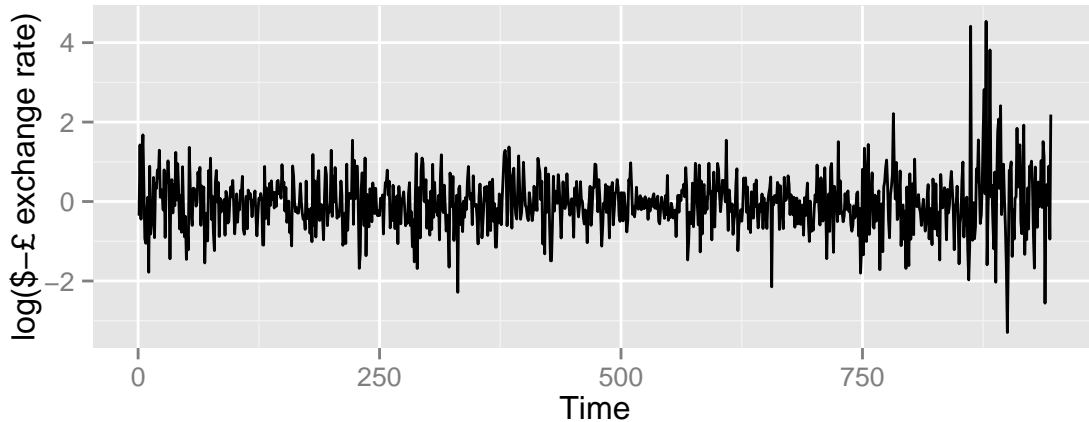


Figure 8: Logarithms of daily difference of the dollar-pound exchange rates between October 1st and June 28th, 1985.

of σ^2 is given by INLA; our method and the MCMC approach are in much better agreement. The bottom-right panel assesses model fit via box plots of posterior p-values (Gelman, 2003), which demonstrates good predictive performance of our method regardless of the prior specifications. The average runtime for our method is approx. 280 sec including calculation of posterior quantiles and p-values.

In addition, Table 3 shows a similar comparison with the method of L&W. The table summarises the posterior distribution across the 32 priors, showing the median of the posterior median, 2.5% and 97/5% quantiles once all the data were analysed. In general, the table shows the difficulties of parameter estimation for this model, with large differences between the approaches. Our method is in rough agreement with L&W, but it is also able to provide a reasonable estimate for the AR parameter. INLA showed the largest sensitivity across the different priors that were used. L&W provide narrower probability bounds than our method.

7.4 Univariate non-stationary growth model

The univariate non-stationary growth model (UNGM) is a very challenging model for parameter estimation and has been discussed in numerous occasions in the literature; see Kitagawa (1996); Andrieu et al. (2010) for discrete time and Xu and Yoon (2009)

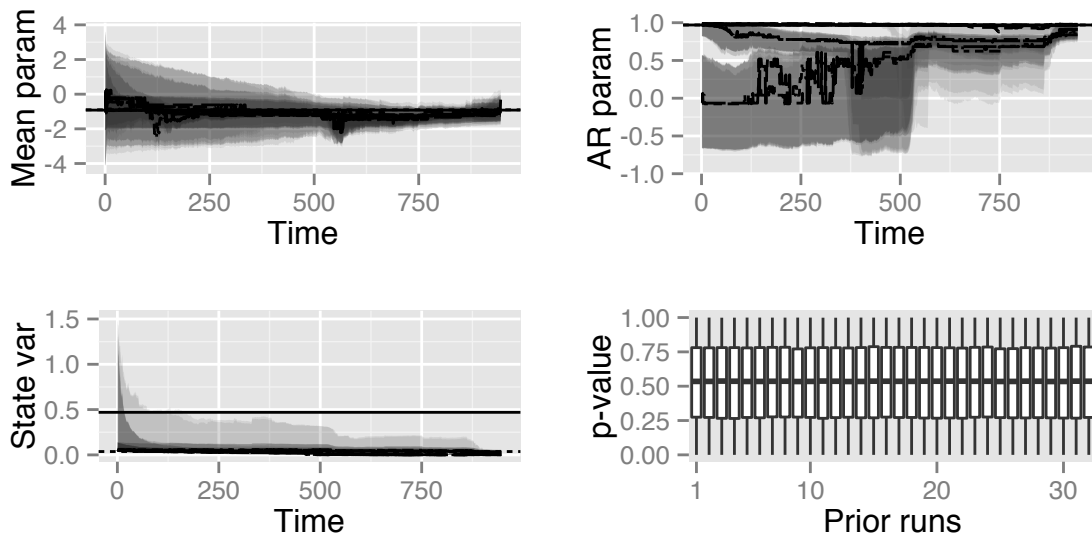


Figure 9: Medians of posterior 2.5%, 50% and 97.5% quantiles over the 32 prior runs for each of the parameters. The solid and dashed lines correspond to the parameter estimate with all the data from INLA and the MCMC approach. The bottom-right panel plots the Bayesian p-values over various priors on x_0 .

Method	$\hat{\mu}$ (95% Intervals)	$\hat{\phi}$ (95% Intervals)	$\hat{\sigma}_y^2$ (95% Intervals)
Our method	-0.99 (-1.24, -0.53)	0.93 (0.84, 0.98)	0.014 (0.002, 0.039)
Liu & West	-0.81 (-1.29, -0.37)	1 (1, 1)	0.016 (0.012, 0.018)
MCMC	-0.9 (-0.96, -0.82)	0.968 (0.943, 0.975)	0.037 (0.03, 0.059)
INLA	-0.9 (-0.95, -0.86)	0.976 (0.956, 0.983)	0.49 (0.25, 0.72)

Table 3: Posterior 2.5%, 50% and 97.5% quantiles for each of the parameters from: our approach, L&W’s, MCMC and INLA.

for continuous time version. The model is given by:

$$y_t = x_t^2 / 20 + \epsilon_t, \quad \text{where } \epsilon_t \sim \mathcal{N}(0, \sigma_y^2), \quad (7.9)$$

$$x_t = 0.5 x_{t-1} + 25 x_{t-1} / (1 + x_{t-1}^2) + 8 \cos(1.2 t) + \eta_t, \quad \text{where } \eta_t \sim \mathcal{N}(0, \sigma_x^2), \quad (7.10)$$

$$x_0 \sim N(0, 1).$$

The parameters are $\theta = (\sigma_y^2, \sigma_x^2)$. Estimation procedures for parameters for this model have focussed on batch analysis rather than online. As well as being highly nonlinear, both in the state and the observation process, the state process is multimodal, making filtering and prediction difficult, particularly when $x_t^2/2$ is small relative to σ_y (Kotecha and Djurić, 2003).

Data were simulated from this model with $\sigma_y^2 = \sigma_x^2 = 1$. For estimation of the state process, both the augmented UKF (Wu et al., 2005) and quadrature methods were applied, and the latter performed better, at the cost of a longer computation time. Updating the posterior using $p(y_t | \mathbf{y}_{1:t-1}, \theta)$ — that is, the sequential version of Bayes Law — produced better performance. Another issue is prior sensitivity, which can be considerable because of the nonlinearity and lack of identifiability in the state process. Because of this, to evaluate our method, multiple data sets from this model were generated, while keeping the prior on the state and parameters to be the same in all the runs with the grid to be 40 for both.

Figure 10 summarises the inference over these 24 runs which ran over an average time of 850 sec again including computation for posterior quantiles and p-values which slows the process down significantly. Note the extremely narrow posterior bounds for each run (top row of the plot), along with consistent under-estimation of the state variance parameter.

An explanation for the under-estimation of σ_x^2 is that while σ_y^2 has been identified

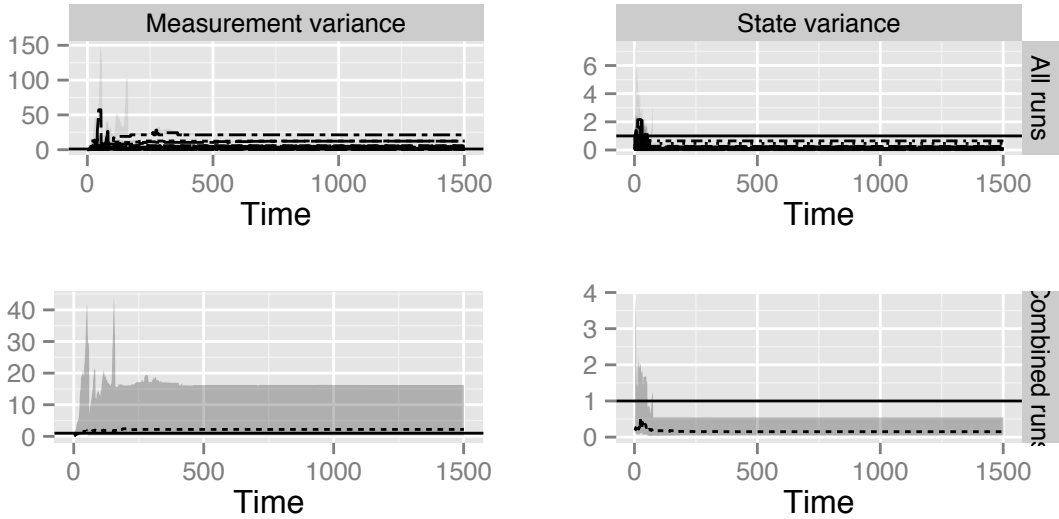


Figure 10: Top: posterior modes, 2.5% and 97.5% quantiles for all 24 runs. Bottom: 2.5%, 50% and 97.5% percentiles of the posterior medians from the 24 runs.

by our method, note the high variability between the runs, mostly indicating over-estimation, an indication of an identifiability problem. In several cases, the posterior distribution diverged considerably from the true value, usually due to the presence of severe nonlinearity. Furthermore, the individual runs were found to have extremely narrow probability bounds, implying that the posterior is concentrated in a narrow region of parameter space. However, Figure 11, shows that predictive performance is good (this is for a single run). It is possible to identify the problem faced in identifying the observation when it is close to 0, i.e. when $x_t^2/2$ is small, in the figure.

Table 4 compares the performance of the L&W method and two other offline methods: iterated filtering and PMCMC. All the algorithms were provided with a similar set of priors and their posterior medians were recorded. The values provided in the table are the 2.5%, 50% and 97.5% percentiles of these medians from multiple runs. L&W algorithm seem to have estimated the state variance parameter correctly, while our method clearly under-estimated it. However, performance of our algorithm is better for the measurement variance parameter. A major disadvantage for the L&W method is that all the runs ended in a single particle, thus degeneracy is certain, a problem our method does not face. Iterated filtering does contain the true parameter values in its bounds, but at the cost of very large standard errors.

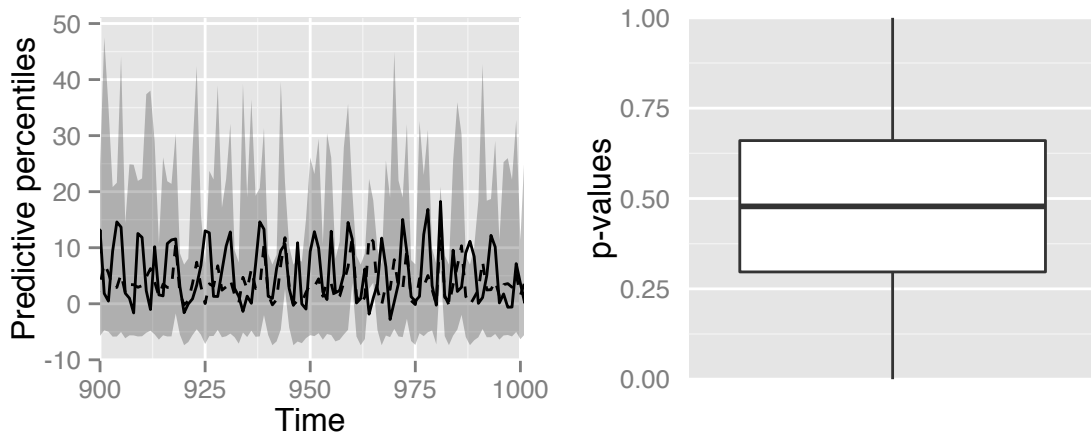


Figure 11: Left: snapshot of the 1-step marginal predictive median (dashed) and probability bounds (band) along with the data (solid) for a single run. Right: box plot of posterior p-values.

PMCMC is the only method which estimates both the parameters correctly, even though it is extremely sensitive to the proposal distribution for the parameters and very easily can get caught in local modes. This goes on to display the extreme complication associated with the UNGM model, because of the nonlinearity in both levels of the state space models. In conclusion one may say that for models with severe nonlinearity and identifiability issues existing online algorithms fail to fit the model well.

Method	$\hat{\sigma}_x^2$ (95% Intervals)	$\hat{\sigma}_y^2$ (95% Intervals)
Our method	0.15 (0.04, 0.54)	2.19 (0.27, 16.25)
Liu & West	1.19 (0.43, 14.6)	9.29 (0.84, 23.27)
Integrated filtering	0.47 (0.47, 76.93)	33.18 (0.62, 34.68)
PMCMC	2.01 (0.67, 9.93)	3.11 (0.52, 5.35)

Table 4: The median, with 2.5% to 97.5% percentiles of posterior medians over multiple runs from our approach, L&W, iterated filtering and PMCMC.

8. CONCLUSION

A method of sequential static parameter estimation for generalized dynamic state space models has been proposed and compared to multiple alternative approaches, both online and offline. Our method achieved similar accuracy to the offline methods in almost all the examples provided and usually performed better than the particle filter. Furthermore, it does not suffer from usual scenarios of degeneracy.

Our method has several appealing properties. It can be applied to models with a non-Gaussian state process, and so is more general than INLA. When the underlying process is Gaussian, it can make use of INLA for a good initialization. It is a very flexible framework in that any filtering and prediction algorithm, not just those used here, can be plugged in as the approximations for posterior predictive or filtering distributions. Dynamic grid updating by tensor product splines is fast, simple and seems to work well, at least in the examples described in the paper; many alternative interpolation methods could be implemented once this approach fails to be sufficiently accurate. Finally, the algorithm is trivially parallelizable over the grid; the computation of $\tilde{p}(\theta_j|\mathbf{y}_{1:T})$ at each θ_j is completely independent.

The principal disadvantage of the approach is that it is restricted to models with a relatively small number of fixed parameters, hence suited for example to inference on hyper-parameters related to the case of hierarchical model in a dynamic setting. Since the number of hyper-parameters in such models are usually much smaller than the number of parameters, which themselves could be time dependent, our algorithm can be fast enough for online Bayesian inference.

Finally, a short discussion on computational complexity of our algorithm is provided. In this paper, our method has not been formally compared to the other online algorithms in terms of computation time, such as L&W method, because the latter is applied from an R package, which would have more optimized coding than ours. This method has computational complexity $\mathcal{O}(G^p)$ where G is the number of grid points in each dimension and $p = \dim(\theta)$. Because of the recursive nature of the algorithm, the computation is independent of t . For each grid point, the complexity also depends on that of the Kalman filter, EKF, UKF or other state estimation methods; total complexity then becomes $\mathcal{O}(G^p \star K)$, where K corresponds to computation in the state estimation process. For the Kalman filter, for example, $K = \mathcal{O}(n^3 + n^2m + nm^2 + m^3)$ where n and m denote the dimensions of the state and observation process respectively. However note that because of the nature of the algorithm, the state estimation steps can be trivially parallelized across points in the grid.

9. Appendix

9.1 Convergence of filtering distribution under Hilbert metric approach

We will first present the general form for a discrete time filter. Let $X = (X_t)$ be a Markov sequence with values in \mathbb{R}^d and transition probability density $p(x_t|x_{t-1}, \theta)$ and initial probability density $p(x_0|\theta)$. The observation process $Y = (Y_t)$ is an i.i.d sequence conditional of X , such that we have $p(y_t|x_t, \theta)$ as the likelihood. We drop the dependence on θ from future references as they don't play a role in the convergence proofs involving state filtering.

The filtering equation of X_t given $Y_{1:t}$ for the general problem can be written as

$$\pi_t(x_t) = \frac{p(y_t|x_t) \int_{\mathbb{R}^d} p(x_t|x_{t-1}) \pi_{t-1}(x_{t-1}) dx_{t-1}}{\int_{\mathbb{R}^d} p(y_t|x_t) \int_{\mathbb{R}^d} p(x_t|x_{t-1}) \pi_{t-1}(x_{t-1}) dx_{t-1} dx_t}.$$

Even though the filtering equation above is nonlinear, its solution at each time point is obtained by using the linear Zakai equation:

$$\rho_t(x_t) = p(y_t|x_t) \int_{\mathbb{R}^d} p(x_t|x_{t-1}) \pi_{t-1}(x_{t-1}) dx_{t-1}, \text{ for } t \geq 0, \quad (9.1)$$

and thus we have

$$\pi_t(x_t) = \frac{\rho_t(x_t)}{\int_{\mathbb{R}^d} \rho_t(u) du}. \quad (9.2)$$

The *Hilbert projective metric* approach has been used to prove the stability result of the state filter for a finite state space model by Atar and Zeitouni (1997) and by LeGland and Oudjane (2004) in particle filters. Let $\mathbb{S} \subseteq \mathbb{R}^d$ be a measurable set and \mathcal{M}_+ be the space of nonnegative measures on $(\mathbb{S}, \mathbb{B}(\mathbb{S}))$ with the partial order relation $p \preceq q$ if $p(A) \leq q(A)$ for any measurable $A \subseteq \mathbb{S}$. The measures p and q are *comparable* if $c_1 p \preceq q \preceq c_2 p$ for some constants $c_1, c_2 > 0$.

Definition (Hilbert projective distance metric) The Hilbert projective distance is defined as

$$h(p, q) = \log \frac{\sup_{A, q(A) > 0} p(A)/q(A)}{\inf_{A, q(A) > 0} p(A)/q(A)}, \quad p, q \in \mathcal{M}_+ \text{ are comparable}$$

and $h(p, q) = \infty$ otherwise.

The following properties of the Hilbert metric are useful:

1. $h(p, q)$ is a non-negative symmetric function;

2. It satisfies the triangle inequality

$$h(p, q) \leq h(p, r) + h(r, q), \quad p, q, r \in \mathcal{M}_+;$$

3. $h(p, q) = 0$ iff $p = cq$ for some $c > 0$;

4. $h(p, q) = h(c_1 p, c_2 q)$ for any $p, q \in \mathcal{M}_+$ and any scalars $c_1, c_2 > 0$;

5. $\|p - q\| \leq \frac{2}{\log(3)} h(p, q)$ for any $p, q \in \mathcal{S}^{d-1}$.

Further let K be a linear operator, mapping \mathcal{M}_+ to itself, then the *Birkhoff contraction coefficient* is defined as

$$\tau(K) = \sup_{0 < h(p, q) < \infty} \frac{h(Kp, Kq)}{h(p, q)}. \quad (9.3)$$

Defining h -diameter as $H(K) = \sup_{p, q \in \mathcal{M}_+} h(Kp, Kq)$, we have

$$\tau(K) = \tanh\left(\frac{H(K)}{4}\right). \quad (9.4)$$

In the filtering concept, the operator is of a particular integral structure and then if the transition kernel satisfies the “mixing conditions”, i.e. if \exists constants κ_\star and κ^\star s.t.

$$0 < \kappa_\star \leq p(x_t | x_{t-1}) \leq \kappa^\star < \infty,$$

it can be shown that

$$\tau(K) \leq \frac{\kappa^\star - \kappa_\star}{\kappa^\star + \kappa_\star} < 1. \quad (9.5)$$

Let $\pi = (\pi_t)_{t \geq 0}$ be the filtering posterior that one can get from the data generating process and $\tilde{\pi} = (\tilde{\pi}_t)_{t \geq 0}$ be the applied model, and $p(x_0)$ and $\tilde{p}(x_0)$ be the respective initializations. The condition $p(x_0) \ll \tilde{p}(x_0)$, i.e. $\tilde{p}_i(x_0) = 0 \implies p_i(x_0) = 0$ is sufficient for $\tilde{p}(x_0)$ to be a valid initialization. Under these assumptions we have the following theorem

Theorem 9..1 *Suppose the signal evolves on a subset $\mathbb{S} \subseteq \mathbb{R}^d$, i.e. $p(X_t \notin \mathbb{S}) = 0$ for all $t \geq 0$. Assume that $h(p(x_0), \tilde{p}(x_0)) < \infty$ and there exists a reference measure on \mathbb{S} , with respect to which the transition law of the signal has a uniformly positive and bounded density, i.e.*

$$0 < \lambda_\star \leq p(x_t | x_{t-1}) \leq \lambda^\star < \infty.$$

Then

$$\|\pi_n - \tilde{\pi}_n\| \leq \frac{2}{\log(3)} h(p(x_0), \tilde{p}(x_0)) \left(\frac{\lambda^\star - \lambda_\star}{\lambda^\star + \lambda_\star}\right)^n, \quad n \geq 1.$$

9.2 Outline for convergence proof in our algorithm

Given that our algorithm provides the following filtering distribution:

$$\begin{aligned} p(\theta|y_{1:t}) &\propto p(\theta|y_{1:t-1})p(y_t|y_{1:t-1}, \theta), \\ &= C_t p(\theta|y_{1:t-1})p(y_t|y_{1:t-1}, \theta), \\ &= Cp(\theta) \prod_j p(y_j|y_{1:j-1}, \theta). \end{aligned}$$

We now have

$$\log p(\theta|y_{1:t}) = \log p(\theta) + \log C + \sum_j \log p(y_j|y_{1:j-1}, \theta).$$

Assuming $\tilde{p}(\cdot)$ is the approximate filter and $p(\cdot)$ is the true posterior coming from the true data generating process, the total variation norm between the corresponding logs is

$$\begin{aligned} &|\log p(\theta|y_{1:t}) - \log \tilde{p}(\theta|y_{1:t})| \\ &= \left| [\log p(\theta) + \log C + \sum_j \log p(y_j|y_{1:j-1}, \theta)] - [\log \tilde{p}(\theta) + \log \tilde{C} + \sum_j \log \tilde{p}(y_j|y_{1:j-1}, \theta)] \right|, \\ &\leq |\log p(\theta) - \log \tilde{p}(\theta)| + \sum_j |\log p(y_j|y_{1:j-1}, \theta) - \log \tilde{p}(y_j|y_{1:j-1}, \theta)|. \end{aligned}$$

We can provide an upper bound to the above inequality provided the conditions mentioned in the main body of the paper hold.

References

- Anderson, B. D. O. and J. B. Moore (1981). Detectability and stabilizability of time-varying discrete-time linear systems. *SIAM Journal on Control and Optimization* 19(1), 20–32.
- Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(3), 269–342.
- Andrieu, C., A. Doucet, and V. B. Tadic (2005). On-line parameter estimation in general state-space models. In IEEE (Ed.), *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05. 44th IEEE Conference on*, pp. 332–337.

- Arulampalam, S., S. Maskell, and N. Gordon (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing* 50, 174–188.
- Atar, R. and O. Zeitouni (1997). Lyapunov exponents for finite state nonlinear filtering. *SIAM Journal on Control and Optimization* 35(1), 36–55.
- Cappé, O. (2011). Online EM algorithm for hidden markov models. *Journal of Computational and Graphical Statistics* 20(3), 728–749.
- Cappé, O., E. Moulines, and T. Rydén (2005). *Inference in hidden Markov models*. Springer.
- Carvalho, C. M., M. Johannes, H. F. Lopes, and N. Polson (2010). Particle learning and smoothing. *Statistical Science* 25(1), 88–106.
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90(432), 1313–1321.
- Chigansky, P. (2006, August). *Stability of nonlinear filters: A survey*. Lecture notes in Petropolis, Brazil.
- Chopin, N., P. E. Jacob, and O. Papaspiliopoulos (2013). Smc2: an efficient algorithm for sequential analysis of state space models. *Journal of Royal Statistical Society: Series B (Methodological)* 75(3), 397–426.
- Cobb, G. W. (1978). The problem of the Nile: conditional solution to a change-point problem. *Biometrika* 65(2), 243–251.
- Cressie, N. A. (1993). *Statistics for spatial data*. Wiley-Interscience.
- Davis, P. J. and P. Rabinowitz (2007). *Methods of numerical integration*. Courier Corporation.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 39(1), 1 – 38.
- Doucet, A., N. D. Freitas, and N. Gordon (2001). *Sequential Monte Carlo methods in practice*. Springer.
- Durbin, J. and S. J. Koopman (2001). *Time series analysis by state space methods*. Oxford University Press.

- Fahrmeir, L. (1992). Posterior mode estimation by Extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association* 87(418), 501–509.
- Fahrmeir, L. and L. Knorr-Held (2000). Dynamic and semiparametric models. In M. G. Schimek (Ed.), *Smoothing and Regression*, pp. 513–544. John Wiley & Sons.
- Fahrmeir, L. and G. Tutz (2001). *Multivariate statistical modelling based on generalized linear models*. Springer.
- Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* 71(2), 369–382.
- Gordon, N., D. J. Salmond, and A. F. M. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F* 140(2), 107–113.
- Haykin, S. (2001). *Kalman filtering and neural networks*. Wiley-Interscience.
- Haykin, S., K. Huber, and Z. Chen (2004). Bayesian sequential state estimation for mimo wireless communications. *Proceedings of the IEEE* 92(3), 439–454.
- Ionides, E. L., A. Bhadra, Y. Atchadé, and A. King (2011). Iterated filtering. *Annals of Statistics* 39(3), 1776–1802.
- Ito, K. and K. Xiong (1999). Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control* 45, 910–927.
- Julier, S. J. and J. Uhlmann (1997). A new extension of the Kalman filter to nonlinear systems. In *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls, Orlando, Florida*, pp. 182–193.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering* 1(82), 35–45.
- Kantas, N., A. Doucet, S. S. Singh, J. M. Maciejowski, and N. Chopin (2015). On particle methods for parameter estimation in state-space models. *Statistical Science* 30(3), 328–351.
- Kastner, G. (2016a). Dealing with stochastic volatility in time series using the r package stochvol. *Journal of Statistical Software* 69(5), 1–30.

- Kastner, G. (2016b). *stochvol: Efficient Bayesian Inference for Stochastic Volatility (SV) Models*.
- King, A. A., E. L. Ionides, C. M. Bretó, S. Ellner, B. Kendall, H. Wearing, M. J. Ferrari, M. Lavine, and D. C. Reuman (2010). *pomp: Statistical inference for partially observed Markov processes (R package)*.
- Kitagawa, G. (1987, December). Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association* 82(400), 1032–1063.
- Kitagawa, G. (1996, March). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics* 5(1), 1–25.
- Kluge, S., K. Reif, and M. Brokate (2010). Stochastic stability of the extended Kalman filter with intermittent observations. *IEEE Transactions on Automatic Control* 55(2), 514–518.
- Kotecha, J. H. and P. M. Djurić (2003, October). Gaussian particle filtering. *IEEE Transactions on Signal Processing* 51(10), 2592–2601.
- Kress, R. (1998). *Numerical Analysis*. Springer.
- LeGland, F. and L. Mever (1997). Recursive estimation in hidden Markov models. *Proceedings of the 36th IEEE Conference on Decision and Control, San Diego 4*, 3468 – 3473.
- LeGland, F. and N. Oudjane (2004). Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *Annals of Applied Probability* 14(1), 144–187.
- Merwe, R. V. D., N. de Freitas, A. Doucet, and E. A. Wan (2001). The unscented particle filter. *Advances in Neural Information Processing Systems* 13.
- Pitt, M. K. and N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association* 94(446), 590–599.
- Poyiadjis, G., A. Doucet, and S. S. Singh (2005). Maximum likelihood parameter estimation in general state-space models using particle methods. In *Procedures of the American Statistical Association*.

- Ristic, B., S. Arulampalam, and N. Gordon (2004). *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House.
- Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and application*. Chapman & Hall.
- Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 71, 319–392.
- Schumaker, L. (2007). *Spline functions: Basic theory*. Cambridge University Press.
- Storvik, G. (2002, February). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing* 50(2), 281–289.
- Taylor, S. (1982). Financial returns modelled by the product of two stochastic processes - a study of the daily sugar prices 1961-75. In O. Anderson (Ed.), *Time series analysis: theory and practice*, Volume 1, pp. 203–226. Amsterdam: North Holland.
- Wahba, G. (1990, January). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.
- Wan, E. A. and A. T. Nelson (1997). Neural dual Extended Kalman filtering: Applications in speech enhancement and monaural blind signal separation. In *Proceedings of IEEE Workshop on Neural Networks and Signal Processing*, pp. 466–475.
- West, M. and J. Harrison (1997). *Bayesian forecasting and dynamic models* (Second ed.). Springer series in Statistics. Springer.
- Wu, Y., D. Hu, M. Wu, and X. Hu (2005). Unscented Kalman filtering for additive noise case: augmented vs. non-augmented. *IEEE Signal Processing Letters* 12(5), 357–359.
- Xiong, K., L. D. Liu, and H. Y. Zhang (2009). Modified unscented Kalman filtering and its application in autonomous satellite navigation. *Aerospace Science and Technology* 13(4-5), 238–246.
- Xu, Y. and J. Yoon (2009). A direct quadrature approach for nonlinear filtering. In *American Control Conference*.