

PARTICLE FILTER-BASED APPROXIMATE MAXIMUM LIKELIHOOD INFERENCE ASYMPTOTICS IN STATE-SPACE MODELS*

JIMMY OLSSON¹ AND TOBIAS RYDÉN¹

Abstract. To implement maximum likelihood estimation in state-space models, the log-likelihood function must be approximated. We study such approximations based on particle filters, and in particular conditions for consistency of the corresponding approximate maximum likelihood estimator. Numerical results illustrate the theory.

INTRODUCTION

By a state-space model is meant a bivariate process $(X_k, Y_k)_{k \geq 1}$ such that (X_k) is a Markov chain (on some general state space \mathcal{X}) and (Y_k) is a process that depends locally on (X_k) in the sense that given (X_k) , (i) the Y_k are conditionally independent and (ii) the conditional distribution of Y_n depends on X_n but on no other X -variables. Models with finite state space \mathcal{X} are often referred to as hidden Markov models. In a state-space model the Markov chain (X_k) is assumed unobservable (or, latent), while the process (Y_k) is observable. Hence, all inference etc. must be based on the latter process. State-space models are useful in almost any area where statistical modelling is applied; see the monographs [6, 7] for further reading on the subject in general.

The topic of this paper is parameter estimation in state-space models. The transition kernel Q of (X_k) and the conditional densities $g(y|x)$ of Y_k given $X_k = x$ are both assumed to depend on some (finite-dimensional) parameter θ , which we indicate by writing Q_θ and $g_\theta(y|x)$ respectively. Typically θ is naturally divided into two parts, parametrising Q and g respectively. To estimate θ from some observed data $y_{1:n} = (y_1, y_2, \dots, y_n)$, the standard method is maximum likelihood. This approach is non-trivial however, as the likelihood is generally not available in closed form. Indeed, the log-likelihood $\ell_n(\theta)$ is typically decomposed as

$$\ell_n(\theta) = \sum_{k=1}^n \log p_\theta(y_k | y_{1:k-1}),$$

where $p_\theta(y_k | y_{1:k-1})$ is the conditional density of Y_k given $Y_{1:k-1}$. By conditioning on the state X_k and using the structure of a state-space model, we find that

$$\ell_n(\theta) = \sum_{k=1}^n \log \int_{\mathcal{X}} g_\theta(y_k | x) \mathbb{P}_\theta(X_k \in dx | y_{1:k-1}), \quad (1)$$

* This work was supported by grants from the Swedish Research Council and the Swedish Foundation for Strategic Research.

¹ Centre for Mathematical Sciences, Lund University, Box 118, 221 00 Lund, Sweden

where $\mathbb{P}_\theta(X_k \in \cdot | y_{1:k-1})$ is the *predictive distribution*, or *predictor*, of the state X_k given data $y_{1:k-1}$. This distribution is in general not available in closed form. In the following section we shall study the use of so-called *particle filters* to approximate these distributions, the resulting approximation of the log-likelihood and its maximiser.

1. THE PREDICTOR-FILTER RECURSIONS AND PARTICLE FILTERS

Write $\pi_{k|k-1}^\theta(\cdot) = \mathbb{P}_\theta(X_k \in \cdot | y_{1:k-1})$ for the predictor and $\pi_{k|k}^\theta(\cdot) = \mathbb{P}_\theta(X_k \in \cdot | y_{1:k})$ for the so-called *filter distribution*, or simply *filter*. These distributions are related recursively as

$$\pi_{k|k}^\theta(dx) = \frac{g_\theta(y_k|x) \pi_{k|k-1}^\theta(dx)}{\int_{\mathcal{X}} g_\theta(y_k|x') \pi_{k|k-1}^\theta(dx')} \quad (2)$$

and

$$\pi_{k+1|k}^\theta(dx) = \int_{\mathcal{X}} Q_\theta(x', dx) \pi_{k|k}^\theta(dx'), \quad (3)$$

respectively. Here (2) is simply Bayes' formula, while (3) amounts to propagating the filter distribution through the Markov chain dynamics. The recursions are initialised by setting $\pi_{1|0}^\theta = \nu$; the initial distribution of the Markov chain. These recursions in general lack closed-form solutions, with the exceptions of hidden Markov models (the integrals turn into finite sums) and linear Gaussian state-space models (the solution being provided by the Kalman filter).

The literature contains numerous ways to approximate the solution to (2)–(3). Common approaches include the extended Kalman filter, the unscented Kalman filter, spline methods etc. In this paper we will employ so-called particle filters for this purpose. In a particle filter, a collection of particles is used to track the state X_k of the system, and the predictor and filter distributions are approximated by the empirical distributions of such collections. To make this idea precise, assume that for some index k we have available a collection $(\xi_{k|k-1,i}^{\theta,N})_{1 \leq i \leq N}$ of points in \mathcal{X} whose empirical distribution $\pi_{k|k-1}^{\theta,N}$ is an approximation to $\pi_{k|k-1}^\theta$. The transformation (2) is then approximated as follows.

- (i) *Weighting.* Compute unnormalised weights $\tilde{w}_{k,i}^{\theta,N} = g_\theta(y_k | \xi_{k|k-1,i}^{\theta,N})$ and then normalised weights $w_{k,i}^{\theta,N} = \tilde{w}_{k,i}^{\theta,N} / \sum_j \tilde{w}_{k,j}^{\theta,N}$.
- (ii) *Resampling.* Create a sample $(\xi_{k|k,i}^{\theta,N})_{1 \leq i \leq N}$ by sampling N times independently from $(\xi_{k|k-1,i}^{\theta,N})_{1 \leq i \leq N}$ with weights $(w_{k,i}^{\theta,N})_{1 \leq i \leq N}$.

The empirical distribution $\pi_{k|k}^{\theta,N}$ of the sample $(\xi_{k|k,i}^{\theta,N})_{1 \leq i \leq N}$ approximates $\pi_{k|k}^\theta$. The transformation (3) is approximated as follows.

- (iii) *Mutation.* Create a sample $(\xi_{k+1|k,i}^{\theta,N})_{1 \leq i \leq N}$ by independently sampling $\xi_{k+1|k,i}^{\theta,N}$ from $Q_\theta(\xi_{k|k,i}^{\theta,N}, \cdot)$.

The procedure is initialised by letting $(\xi_{1|0,i}^{\theta,N})_{1 \leq i \leq N}$ be an i.i.d. sample of size N from the initial distribution ν .

The procedure described above is known as the *bootstrap particle filter*: the mutation of particles follows the system dynamics, and resampling is done multinomially at each step. There is an abundance of variations of this simplest scheme with other strategies for mutation and resampling, and we refer to [1, 3] for extensive coverage of particle filter theory and algorithms.

2. ASSUMPTIONS

In this section we give conditions that are assumed to hold throughout the paper. The parameter θ is assumed to belong to a parameter set Θ , which is a compact subset of \mathbb{R}^d for some d . The observations (Y_k) arise from a state-space model with parameter θ^0 ; this is thus the ‘true’ parameter. The transition kernels Q_θ are assumed to admit densities $q_\theta(\cdot, \cdot)$ w.r.t. some fixed *finite* measure μ on \mathcal{X} . Gradients w.r.t. θ are denoted by ∇_θ .

- (A1) The function $q_\theta(x, x')$ is bounded away from zero and infinity, uniformly in θ, x and x' . For all y , the function $\theta \mapsto \int_{\mathcal{X}} g_\theta(y|x) \mu(dx)$ is bounded away from zero and infinity.

The first part of this assumption is typically fulfilled only if \mathcal{X} is compact, or at least bounded. It implies that for all θ, Q_θ is positive Harris recurrent with a unique stationary distribution γ_θ . For a stationary version of the Markov chain, obtained with $\nu = \gamma_\theta$, and the corresponding state-space model, we write $\overline{\mathbb{P}}_\theta$ and $\overline{\mathbb{E}}_\theta$ respectively for the corresponding distributions and expectations. Generally we do assume however that the initial distribution ν is fixed and known; letting ν depend on θ introduces no new principal difficulties, but requires some regularity conditions on the mapping $\theta \mapsto \nu_\theta$ similar to the conditions listed below.

- (A2) The function $g_\theta(y|x)$ is bounded uniformly in θ, x and y , and $\overline{\mathbb{E}}_{\theta^0} |\log b_-(Y_1)| < \infty$ where $b_-(y) = \inf_\theta \int_{\mathcal{X}} g_\theta(y|x) \mu(dx)$.
- (A3) For all x, x' and all y , the functions $\theta \mapsto q_\theta(x, x')$, $\theta \mapsto g_\theta(y|x)$ and $\theta \mapsto \log g_\theta(y|x)$ are continuously differentiable.
- (A4) The gradient $\nabla_\theta \log q_\theta(x, x')$ is bounded uniformly in θ, x and x' , and $\overline{\mathbb{E}}_{\theta^0} [\sup_\theta \sup_x \|\nabla_\theta \log g_\theta^{Y_1}(x)\|] < \infty$.
- (A5) For some integer $p \geq 1$, the expectation

$$\mathbb{E}_{\theta^0} \left[\left(\sup_{x', x''} \frac{g_\theta(Y_1|x')}{g_\theta(Y_1|x'')} \right)^p \middle| X_1 = x \right]$$

is bounded in θ .

- (A6) $\theta = \theta^0$ if and only if $\overline{\mathbb{P}}_\theta^Y = \overline{\mathbb{P}}_{\theta^0}^Y$.

3. LIKELIHOOD APPROXIMATION

Replacing the predictor in (1) with its particle filter counterpart, an obvious approximation to the log-likelihood is

$$\ell_n^N(\theta) = \sum_{k=1}^n \log \int_{\mathcal{X}} g_\theta(y_k|x) \pi_{k|k-1}^{\theta, N}(dx) = \sum_{k=1}^n \log \left(\frac{1}{N} \sum_{i=1}^N g_\theta(y_k|\xi_{k|k-1, i}^{\theta, N}) \right). \tag{4}$$

Moreover, finding the point $\hat{\theta}_n^N$ where this function is maximal would produce an approximate maximum likelihood estimator. Questions we address in this paper concern asymptotic properties of such an estimator; is it consistent, asymptotically normal etc.? To achieve this we expect that it is required to increase the size N of the particle filter as the sample size n of the observed data increases, but how fast need this increase be?

First however, we must take a closer look at the function $\ell_n^N(\theta)$. For a fixed θ it is a random variable, the randomness coming from the particle filter (at this point we consider the observations as fixed numbers). When evaluating this function for different θ , which is necessary to find its maximum, it is far from obvious how to treat this randomness for different θ . One option is to use ‘independent randomness’ for different θ . This is simple to implement, but the function $\ell_n^N(\theta)$ so obtained becomes everywhere discontinuous. Another option is to use a common set of random numbers in the particle filter (‘fixed randomness’). This produces a function $\ell_n^N(\theta)$ that is piecewise continuous, but still discontinuous at points where any resampling step in the particle filter goes from selecting one particle to another one. In addition, the stochastic properties of the approximation are more difficult to analyse, because of the dependence across θ . Pitt [5] proposed a smoothed version of the filter with fixed randomness that does give a continuous likelihood approximation; this approach does only work for one-dimensional state spaces \mathcal{X} however. In this paper we shall rather study approximations on a finite grid over Θ , and devise an analysis that allows either independent or fixed randomness.

The starting point of the development is the following result.

Theorem 3.1 ([2, Theorem 7]). *Let $\tilde{\theta}_n$ be an estimator satisfying $\ell_n(\tilde{\theta}_n) \geq \sup_{\theta \in \Theta} \ell_n(\theta) - o_P(n)$ with $P = \mathbb{P}_{\theta^0}$. Then $\tilde{\theta}_n$ is consistent.*

We notice that the randomness in this expression in the first place stems from the observations $Y_{1:n}$, which are not fixed here but considered as a sample of size n from a stochastic process with distribution \mathbb{P}_{θ^0} . However, any additional randomness—such as from a particle filter—needs to be accounted for as well. It is straightforward to check that if $\tilde{\ell}_n$ is an approximation to ℓ_n such that

$$\sup_{\theta \in \Theta} |\tilde{\ell}(\theta) - \ell_n(\theta)| = o_P(n) \quad (5)$$

and $\tilde{\theta}_n$ is the maximiser of $\tilde{\ell}_n$, then $\tilde{\theta}_n$ satisfies the conditions of the theorem and is hence consistent [2, p. 2285].

To construct our particular approximation to the log-likelihood, we introduce a finite set $(\bar{\theta}_i)_{1 \leq i \leq M}$ of points in Θ . This set is typically a regular grid, but does not need to be. With $[\theta]$ denoting the grid point closest to θ , we then put $\tilde{\ell}_n(\theta) = \ell_n^N([\theta])$ with $\ell_n^N(\theta)$ as in (4). The resolution of the grid is denoted by $\Delta = \sup_{\theta} |\theta - [\theta]|$.

4. CONSISTENCY OF THE APPROXIMATE MAXIMUM LIKELIHOOD ESTIMATOR

Let $\hat{\theta}_n^N$ be the maximiser of ℓ_n^N . By construction there is no unique point where ℓ_n^N is maximal, so we choose to pick the grid point where ℓ_n^N is maximal. This choice is arbitrary however, and does not affect the asymptotics. In order to show that this estimator is consistent, we need to prove that ℓ_n^N satisfies (5), where $N = N_n$ and the grid resolution $\Delta = \Delta_n$ both depend on n . The error $\ell_n^N(\theta) - \ell_n(\theta)$ consists of two parts: the error of the log-likelihood approximation at $[\theta]$, and the error in the exact log-likelihood arising from replacing θ with $[\theta]$. Using this decomposition, we find that we must prove that for any $\varepsilon > 0$,

$$\begin{aligned} & \mathbb{P}_{\theta^0} \left(n^{-1} \sup_{\theta} |\ell_n^N([\theta]) - \ell_n(\theta)| \geq \varepsilon \right) \\ & \leq \mathbb{P}_{\theta^0} \left(n^{-1} \sup_{\theta} |\ell_n^N([\theta]) - \ell_n([\theta])| \geq \frac{\varepsilon}{2} \right) + \mathbb{P}_{\theta^0} \left(n^{-1} \sup_{\theta} |\ell_n([\theta]) - \ell_n(\theta)| \geq \frac{\varepsilon}{2} \right) \rightarrow 0. \end{aligned} \quad (6)$$

Here the second part involves randomness from the observations only, while the first part involves randomness from the particle filter as well.

To find a suitable bound on the first part of the decomposition, apply Boole's, Markov's and Minkowski's inequalities in turn to obtain

$$\begin{aligned} & \mathbb{P}_{\theta^0} \left(n^{-1} \sup_{\theta} |\ell_n^N([\theta]) - \ell_n([\theta])| \geq \varepsilon \right) \leq \sum_{i=1}^M \mathbb{P}_{\theta^0} (|\ell_n^N(\bar{\theta}_i) - \ell_n(\bar{\theta}_i)| \geq n\varepsilon) \\ & \leq \frac{1}{(n\varepsilon)^p} \sum_{i=1}^M \mathbb{E}_{\theta^0} |\ell_n^N(\bar{\theta}_i) - \ell_n(\bar{\theta}_i)|^p \\ & \leq \frac{1}{(n\varepsilon)^p} \sum_{i=1}^M \left\{ \sum_{k=1}^n \left(\mathbb{E}_{\theta^0} |\log \pi_{k|k-1}^{\bar{\theta}_i, N} g_{\bar{\theta}_i}(Y_k|\cdot) - \log \pi_{k|k-1}^{\bar{\theta}_i} g_{\bar{\theta}_i}(Y_k|\cdot)|^p \right)^{1/p} \right\}^p, \end{aligned}$$

where p is as in (A5). Provided that each expectation on the right-hand side can be bounded by $C(p)/N^{p/2}$ for some constant $C(p)$ depending on p , the right-hand side will be of order $MC(p)/(N^{1/2}\varepsilon)^p$. The proof that each expectation is indeed bounded by $C(p)/N^{p/2}$ proceeds in three steps. First, use the inequality $|\log a - \log b| \leq |a - b|/(a \wedge b)$ for $a, b > 0$ to bound the difference of logarithms in terms of the difference $|\log \pi_{k|k-1}^{\bar{\theta}_i, N} g_{\bar{\theta}_i}(Y_k|\cdot) - \log \pi_{k|k-1}^{\bar{\theta}_i} g_{\bar{\theta}_i}(Y_k|\cdot)|$. Secondly, condition on $Y_{1:k}$ to focus on the randomness coming from the particle filter only. Third, use available bounds on L^p norms for particle filters [1, Theorem 7.4.4]. Finally all the pieces need to be put together; see [4] for details.

For the second term on the right-hand side of (6), make the Taylor expansion $\ell_n([\theta]) - \ell_n(\theta) = \nabla_{\vartheta} \ell_n(\vartheta)([\theta] - \theta)$, where ϑ is a point on the line segment between θ and $[\theta]$, and employ the Cachy-Schwartz and Markov inequalities

to obtain

$$\mathbb{P}_{\theta^0} \left(n^{-1} \sup_{\theta} |\ell_n([\theta]) - \ell_n(\theta)| \geq \varepsilon \right) \leq \frac{1}{\varepsilon} \sup_{\theta} \|\theta - [\theta]\| \times \mathbb{E}_{\theta^0} \|n^{-1} \sup_{\theta} \nabla_{\theta} \ell_n(\theta)\|.$$

The first factor on the right-hand side is Δ by definition. The second factor can be shown to be bounded in n , the heuristic being that the score function is the sum of n conditional scores $\nabla_{\theta} \log p_{\theta}(Y_k|Y_{1:k-1})$ and hence increasing linearly in n . This heuristic is indeed true, which can be proved using a decomposition of the score function as in [2, Section 6]; again we refer to [4] for details.

Summing up the above, we find that (6) is bounded by an expression of order $MC(p)/(N^{1/2}\varepsilon)^p + \Delta$. Considering the request (5) that this bound must vanish as $n \rightarrow \infty$, we find the following requirements on $N = N_n$, $M = M_n$ and $\Delta = \Delta_n$.

Theorem 4.1. *Let $p \geq 1$ be fixed, satisfying (A5). Let the grid and the size of the particle filter depend on the sample size n in such a way that $\Delta_n \rightarrow 0$ and $M_n/N_n^{p/2} \rightarrow 0$ as $n \rightarrow \infty$. Then the estimator $\hat{\theta}_n^{N_n}$ is consistent.*

We note that at first sight it appears favourable to take p as large as possible, as this relaxes the requirements on the sequence (N_n) . On the other hand the constant $C(p)$ increases in p , whence for a finite sample size n it is not obvious that the largest possible p is to prefer.

5. NUMERICAL EXAMPLE

As an illustration we consider a model with state space $\mathcal{X} = [-D, D]$ and $X_k = \exp(-\beta X_{k-1}) + \eta_k$, where (η_k) are i.i.d. normal variates $N(0, \sigma_{\eta}^2)$. Whenever this addition yields a sum outside \mathcal{X} , the state is reflected into \mathcal{X} . The observed process is given by $Y_k = \theta X_k^2 + \varepsilon_k$, where (ε_k) are i.i.d. normal variates $N(0, \sigma_{\varepsilon}^2)$. We simulated 30 trajectories of length 2,000 for the parameters $(\sigma_{\varepsilon}^2, \sigma_{\eta}^2, \beta, \theta) = (0.1, 0.1, 1, 1/\sqrt{2})$ with the initial distribution $N(0, 0.1)$ (reflected as well) for X_1 . All parameters except θ were considered as known and set to their true values, while optimisation was carried out w.r.t. θ . Theorem 4.1 was applied with $p = 6$, a uniform grid on $\Theta = [0, 1]$ with resolution $\Delta_n = 1/(5n^{1/2})$, and particle filter size $N_n = 5 \lceil n^{23/60} M_n^{1/3} \rceil$; here $\lceil \cdot \rceil$ denotes upwards rounding.

Figure 1 shows an approximate log-likelihood curve obtained for $n = 100$ observations and $N = 110$ particles. The same set of random numbers was used for the particle filter at all θ (fixed randomness). Obviously the approximation is smooth already for this rather small number of particles. Figure 1 also shows box-plots of the approximate maximum likelihood estimates (MLEs) obtained from samples of sizes $n = 100, 1,000$ and $2,000$ respectively, with Δ_n and M_n as above (the resulting N are 110, 385 and 560 respectively). The estimates become increasingly concentrated around the true $\theta^0 = 1/\sqrt{2} \approx 0.707$, although still with a slight bias for the largest sample size.

Figure 2 clearly illustrates the point in not over-dimensioning the particle filter. Here approximate MLEs of θ were computed from 50 samples of size $n = 1,000$, using particle filters of sizes $N = 300$ and $N = 1,500$ respectively and with a five times denser grid in the latter case. The sample standard deviation of the 50 estimates so obtained was 0.013 and 0.012 respectively. Increasing N (and decreasing Δ) even further would decrease this variability only marginally, as the variation in the estimates is then totally dominated by the sample variation intrinsic to the maximum likelihood estimator itself (which decreases only with n). Thus, for a fixed sample size n it is sensible to choose N large enough that the variability of the parameter estimate due to the particle filter variation is smaller than the variability of the maximum likelihood estimator itself, while choosing N much larger is only cost ineffective. Figure 2 also shows that the approximate MLEs are approximately normal. Such an asymptotic result can indeed be verified, provided N_n increases faster than what is required for consistency; see [4] for details.

REFERENCES

[1] P. Del Moral, *Feynman-Kac Formulae. Geneological and Interacting Particle Systems with Applications*, Springer, New York, 2004.

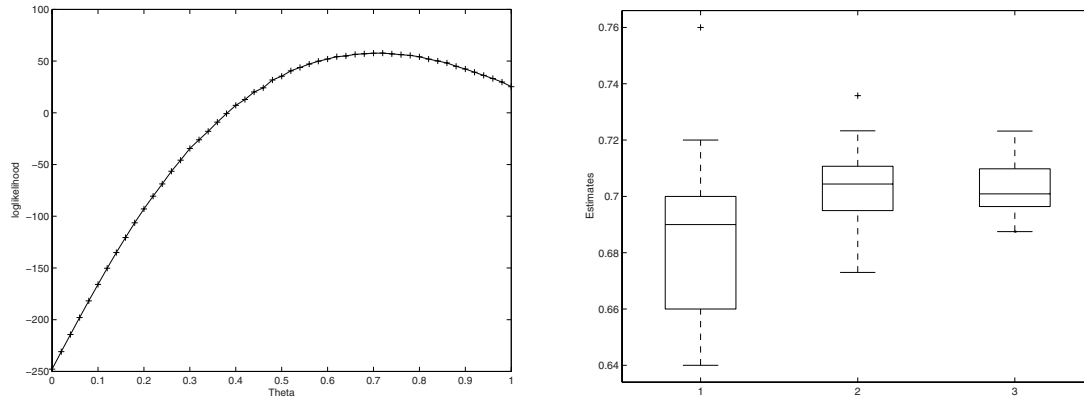


FIGURE 1. Left plot: Likelihood approximation based on a sample of size $n = 100$ and particle filter of size $N = 110$. Right plot: Box-plots of approximate MLEs computed from 30 samples of sizes $n = 100, 1,000$ and $2,000$ respectively (left to right), with corresponding $\Delta_n = 0.02, 0.0063$ and 0.0045 and $M_n = 110, 385$ and 560 respectively.

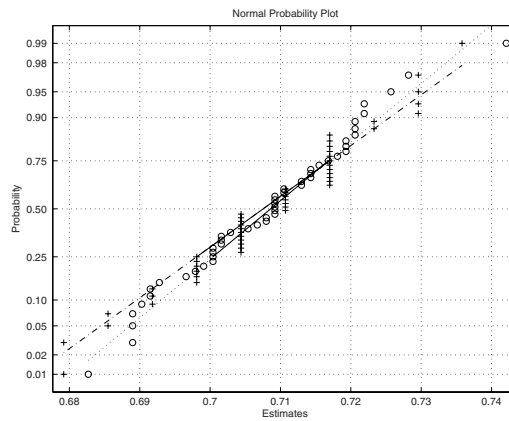


FIGURE 2. Normal probability plots of approximate MLEs computed from 50 samples of size $n = 1,000$ and particle filter sizes $N = 300$ (+) and $N = 1,500$ (o). In addition, the grid is five times denser for $N = 1,500$.

- [2] R. Douc, É. Moulines and T. Rydén, Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime, *Ann. Statist.*, **32**, 2002, pp. 2254–2304.
- [3] A. Doucet, N. de Freitas and N. Gordon, *An Introduction to Sequential Monte Carlo Methods*, Springer, New York, 2001.
- [4] J. Olsson and T. Rydén, Asymptotic properties of the bootstrap particle filter maximum likelihood estimator for state-space models. Preprint, Centre for Mathematical Sciences, Lund University, 2006.
- [5] M.K. Pitt, Smooth particle filters for likelihood evaluation and maximisation. Preprint, Department of Economics, University of Warwick, 2002.
- [6] R.H. Shumway and D.S. Stoffer, *Time Series Analysis and its Applications*, 2nd ed., Springer, New York, 2006.
- [7] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*, 2nd ed., Springer-Verlag, New York, 1997.