ONLINE SEQUENTIAL MONTE CARLO EM ALGORITHM

Olivier Cappé

LTCI, Télécom ParisTech & CNRS
46 rue Barrault, 75013 Paris, France

ABSTRACT

Online (or recursive) estimation of fixed model parameters in general state-space models is a crucial but often difficult task. This paper is about likelihood-based point estimation, showing that an online EM (Expectation-Maximization) algorithm recently proposed for discrete hidden Markov models can be extended to more general settings, including non-linear non-Gaussian state-space models that necessitate the use of sequential Monte Carlo filtering approximations. The performance of the proposed online sequential Monte Carlo EM algorithm is illustrated on numerical examples.

1. INTRODUCTION

State-space models certainly are one of the concepts of statistical signal processing that has had the most profound practical impact in the latest forty years. Recent advances in this field include Sequential Monte Carlo (SMC) filtering approximation techniques which make it possible to perform inference in models that are more general than linear Gaussian models or models with finite number of states (sometimes called hidden Markov models) [6, 5]. Despite these advances, the estimation of fixed model parameters remains a difficult issue. Much effort has focussed on the fully Bayesian approach, in which one sequentially updates a Monte Carlo approximation to the fixed parameter posterior [9, 12, 8, 15]. This is a challenging problem however as the Bayesian posterior for fixed parameters tends to concentrate and hence it is very difficult to prevent degeneracy of the Monte Carlo approximation to this posterior in the long term.

In this contribution, I consider the comparatively simpler problem of determining sequential point estimates that are as close as possible to the maximum likelihood estimates and eventually share the same asymptotic convergence properties. In this context, I believe that it is of some interest to consider methods that are as close as possible to the principle of EM (Expectation-Maximization) algorithm used for batch estimation. Although not necessarily preferable to other numerical optimization approaches, the EM algorithm is largely dominant in practice due to its stability and ease of implementation combined with its ability to explicitly handle parameters constraints. In the context of online estimation, where the parameter estimate much be updated with each new observation, very simple algorithms are indeed required and EM tends to be simpler than gradient based algorithms, the latter often requiring line searches or projections to deal with constraints and matrix weighting to handle multidimensional parameters.

Related works include [1, 2] which capitalize on the idea originally proposed for Hidden Markov Models (hereafter abbreviated to HMMs) by [14] that the exact likelihood can be approximated by fixed-memory pseudo likelihoods, also called split likelihoods. This approach thus only requires fixed memory smoothing, which is provably easier to achieve in the context of SMC. Likewise the batch approach of [5] uses fixed-lag approximation ideas also inspired by an algorithm previously considered for HMMs [11]. Finally, one should also mention [10] which proposes an approach that is more reminiscent of simulated annealing but also capitalizes on EM related-ideas

Recently however, Mongillo and Denève [13] proposed an online version of the EM algorithm which is not based on finite memory or fixed-lag smoothing ideas. The main tool is a recursion which allows for data recursive computation of smoothing functionals required by the EM algorithm. However, this recursion appears to be very specific to the case considered by [13], that is, HMMs for which both the states and observations take a finite number of values. In a companion paper [3], it is shown that this is indeed not the case and that this idea can be seen as a generalization of the online EM algorithm for mixture discussed in [4] combined with a scheme for recursive implementation of smoothing for sum functionals of the state variables which can be traced back to [16, 7].

The purpose of this contribution is to show that the algorithm of Mongillo and Denève can be extended to provide online EM estimation in more general models, including continuous state-space non-linear and non-Gaussian models. In this proposed approach, SMC is also used to approximate an auxiliary recursion that maintains smoothed estimates of the complete-data EM sufficient statistics. The framework of [3] is first extended in Section 2 to the case of continuous state-space dynamic models. In Section 3, the use of SMC is then considered so as to obtain a realistic online estimation algorithm. Section 4 contains some experimental evaluations in two simulated scenarios.

2. ONLINE EM FOR STATE-SPACE MODELS

Consider a state-space model with observations Y_t and associated state variables X_t. Let $q_\theta(x_{t-1}, x_t)$ denote the state probability transition function and $g_\theta(x_t, y_t)$ the observation probability transition function, both of them depending on an unknown parameter θ. It is further assumed that the product $q_\theta(x_{t-1}, x_t)g_\theta(x_t, y_t)$ belongs to an exponential family of distributions with complete-data sufficient statistic $s(x_{t-1}, x_t, y_t)$. With these notations, each iteration of the usual batch-mode EM algorithm consists in

E-step Compute $S_n = \frac{1}{n}\mathrm{E}[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t)|Y_{0:n}]$ for the current value of the parameter estimate;

M-step Update the parameter estimate to $\bar{\theta}(S_n)$, where $\bar{\theta}(S) \mapsto \theta$ is the solution (here assumed to be unique) to the complete-data score equations.

In order to generalize [13, 3], based on [16, 7], observe that S_n can be updated recursively by defining the filtering pdf $\phi_n(x)$ such

that $\int f(x)\phi_n(x)dx = \mathrm{E}[f(X_n)|Y_{0:n}]$ and the auxiliary function $\rho_n(x) = \frac{1}{n}\mathrm{E}[\sum_{t=1}^n s(X_{t-1}, X_t, Y_t)|Y_{0:n}, X_n = x]$ which is such that

$$\int f(x)\rho_n(x)\phi_n(x)dx = \mathrm{E}\left[\frac{f(X_n)}{n}\sum_{t=1}^n s(X_{t-1}, X_t, Y_t)\middle| Y_{0:n}\right]$$

for integrable functions f. In particular, $\int \rho_n(x)\phi_n(x)dx = S_n$. The function $\rho_n(x)$ may be updated updated according to (see Chapter 4 of [4])

$$\rho_{n+1}(x) = \int \left\{\frac{1}{n+1}s(x', x, Y_{n+1}) + \left(1 - \frac{1}{n+1}\right) \times \rho_n(x')\right\}\frac{\phi_n(x')q_\theta(x', x)}{\int \phi_n(x'')q_\theta(x'', x)dx''}dx' \quad (1)$$

Thus, the generalization of [13, 3] based on (1) consists in selecting a sequence of positive step-sizes $\gamma_n \in (0, 1)$ which follow the usual stochastic approximation guidelines that $\gamma_n \equiv n^\alpha$, with $\alpha \in (\frac{1}{2}, 1]$ and to apply the following update.

Algorithm 1

Stochastic Approximation E-step

$$\hat{\phi}_{n+1}(x) = \frac{\int \hat{\phi}_n(x')q_{\hat{\theta}_n}(x', x)g_{\hat{\theta}_n}(x, Y_{n+1})dx'}{\iint \hat{\phi}_n(x')q_{\hat{\theta}_n}(x', x'')g_{\hat{\theta}_n}(x'', Y_{n+1})dx'dx''}$$

$$\hat{\rho}_{n+1}(x) = \int \left\{\gamma_{n+1}s(x', x, Y_{n+1}) + (1 - \gamma_{n+1}) \times \hat{\rho}_n(x')\right\}\frac{\hat{\phi}_n(x')q_{\hat{\theta}_n}(x', x)}{\int \hat{\phi}_n(x'')q_{\hat{\theta}_n}(x'', x)dx''}dx' \quad (2)$$

M-step

$$\hat{\theta}_{n+1} = \bar{\theta}\left(\int \hat{\rho}_{n+1}(x)\hat{\phi}_{n+1}(x)dx\right)$$

3. SEQUENTIAL MONTE CARLO APPROXIMATION

Of course, Algorithm 1 cannot be applied directly in usual settings as the integrals involved in the update equations are not available in closed form. To obtain a practical algorithm, consider a generic SMC method, of the sampling importance resampling type, applied to the model under consideration. The filtering density ϕ_n is approximated by so-called particles $\{\xi_n^i\}_{1\leq i\leq m}$ with associated weights $\{w_n^i\}_{1\leq i\leq m}$, such that $w_n^i \geq 0$ and $\sum_{i=1}^m w_n^i = 1$. Upon observing Y_{n+1}, the particles $\{\xi_{n+1}^i\}_{1\leq i\leq m}$ are simulated independently under the mixture density $\sum_{i=1}^m w_n^i r(\xi_n^i, x)$, that is, one first draws an index J_n^i under the discrete "probability" $w_n^1, \ldots, w_n^m$ and then simulate ξ_{n+1}^i from the density $r(\xi_n^{J_n^i}, x)$. r is an instrumental probability transition function which may be chosen arbitrarily. Then the new weights are determined as

$$w_{n+1}^i \propto \frac{q_\theta(\xi_n^{J_n^i}, \xi_{n+1}^i)g_\theta(\xi_{n+1}^i, Y_{n+1})}{r(\xi_n^{J_n^i}, \xi_{n+1}^i)}$$

where the proportionality constant is determined by the constraint that the weights sum to one. Variants exists in which the resampling weights are not necessarily selected to be equal to the importance weights w_n^i [6, 5] but this is not considered here for simplicity. In addition to the approximation of the filter ϕ_n, one needs a Monte Carlo based approximation to the auxiliary quantity $\hat{\rho}_n$ in (2). It seems natural to use a Monte Carlo approximation for $\hat{\rho}_n$ defined by adjustment values $\{\rho_n^i\}_{1\leq i\leq m}$ such that $\int \hat{\rho}_n(x)f(x)\hat{\phi}_n(x)dx$ is approximated by

$$\sum_{i=1}^m \rho_n^i w_n^i f(\xi_i)$$

Note that when the sufficient statistic $s(x_{t-1}, x_t, y_t)$ is vector valued (which is usually the case in multiparameter models), the adjustment values ρ_n^i also are vector valued and are not normalized in any ways. The proposed online SMC EM algorithm may now be described as follows.

Algorithm 2

Filter update *Draw* $J_n^i \sim w_n^1, \ldots, w_n^m$ *and* $\xi_{n+1}^i \sim r(\xi_n^{J_n^i}, x)$*; let*

$$w_{n+1}^i \propto \frac{q_{\hat{\theta}_n}(\xi_n^{J_n^i}, \xi_{n+1}^i)g_{\hat{\theta}_n}(\xi_{n+1}^i, Y_{n+1})}{r(\xi_n^{J_n^i}, \xi_{n+1}^i)}$$

Adjustment values update

$$\rho_{n+1}^i = \gamma_{n+1}s(\xi_n^{J_n^i}, \xi_{n+1}^i, Y_{n+1}) + (1 - \gamma_{n+1})\rho_n^{J_n^i} \quad (3)$$

Parameter update

$$\hat{S}_{n+1} = \sum_{i=1}^m \rho_{n+1}^i w_{n+1}^i$$

$$\hat{\theta}_{n+1} = \bar{\theta}(\hat{S}_{n+1})$$

To understand why (3) is "properly weighted" [6, 5], ignore the dependence with respect to the parameters estimate $\hat{\theta}_n$ and represent $\sum_{i=1}^m \rho_{n+1}^i w_{n+1}^i f(\xi_{n+1}^i)$ as N_{n+1}/D_{n+1} where $N_{n+1} = \frac{1}{m}\sum_{i=1}^m \rho_{n+1}^i \bar{w}_{n+1}^i f(\xi_{n+1}^i)$, $D_{n+1} = \frac{1}{m}\sum_{i=1}^m \bar{w}_{n+1}^i$, and $\bar{w}_{n+1}^i$ refers to the *unnormalized* weights (i.e., before dividing by the sum). Then, it is easily checked that

$$\mathrm{E}[N_{n+1}|\mathcal{F}_n] = \sum_{i=1}^m w_n^i \int \left\{\gamma_{n+1}s(\xi_n^i, x, Y_{n+1}) + (1 - \gamma_{n+1}) \times \rho_n^i\right\}q(\xi_n^i, x)g(x, Y_{n+1})f(x)dx$$

$$\mathrm{E}[D_{n+1}|\mathcal{F}_n] = \sum_{i=1}^m w_n^i \int q(\xi_n^i, x)g(x, Y_{n+1})dx$$

where $\mathcal{F}_n$ denotes the observations and simulations up to index n as well as the new observation Y_{n+1}. Hence, if one assumes consistency of the usual particle filter, that is, $\sum_{i=1}^m w_n^i f(\xi_n^i) \to \int f(x)\phi_n(x)dx$ as $m \to \infty$,

$$\mathrm{E}[N_{n+1}|\mathcal{F}_n] \to \iint \left\{\gamma_{n+1}s(x', x, Y_{n+1}) + (1 - \gamma_{n+1}) \times \hat{\rho}_n(x')\right\}\phi_n(x')q(x', x)g(x, Y_{n+1})f(x)dx'dx$$

$$\mathrm{E}[D_{n+1}|\mathcal{F}_n] \to \iint \phi_n(x')q(x', x)g(x, Y_{n+1})dx'dx$$

assuming that indeed $\rho_n^i = \hat{\rho}_n(\xi_n^i)$ for an integrable function $\hat{\rho}_n$. Note finally that the ratio of both limits may be interpreted as

$$\mathrm{E}\left[\left\{\gamma_{n+1}s(X_n, X_{n+1}, Y_{n+1}) + (1 - \gamma_{n+1}) \times \hat{\rho}_n(X_n)\right\}f(X_{n+1})\middle|Y_{0:n+1}\right]$$

as expected. □

4. NUMERICAL EXAMPLES

To start with an example where comparison with exact computations is feasible, consider the simple noisy Gaussian AR(1) model observed in noise:

$$X_{t+1} = \phi X_t + U_{t+1}$$
$$Y_t = X_t + V_t$$

with parameters ϕ, $\sigma^2 = \mathrm{E}U_t^2$, and $\kappa^2 = \mathrm{E}V_t^2$. Although very simple, this example is often used for benchmarking estimation techniques as it is indeed rather difficult to estimate the parameters of this model due to the strong ambiguity between X_t and V_t given the observations. This is reflected, for instance, in the slow convergence of the EM algorithm for this model, which requires many iterations to reach accurate estimates of the parameter.

Routine calculations show that each iteration of the EM algorithm[1] may be implemented by computing, in the E-step, $S_n(i) = \sum_{t=1}^n \mathrm{E}\left[\, s_i(X_{t-1}, X_t, Y_t) | \, Y_{0:n}\right]$ for the complete data statistics $s_0 = 1$, $s_1(x_t, y_t) = (y_t - x_t)^2$, $s_2(x_{t-1}) = x_{t-1}^2$; $s_3(x_{t-1}, x_t) = x_{t-1}x_t$ and $s_4(x_t) = x_t^2$. The M-step then consists in updating the parameters through the mapping $\bar\theta$ defined by $\kappa^2 = S_n(1)/S_n(0)$, $\phi = S_n(3)/S_n(2)$ and $\sigma^2 = (S_n(4) - S_n^2(3)/S_n(2))/S_n(0)$. Hence, in this case, the adjustment values ρ_{n+1}^i are four dimensional, corresponding to the statistics s_1 to s_4 (as s_0 is deterministic).

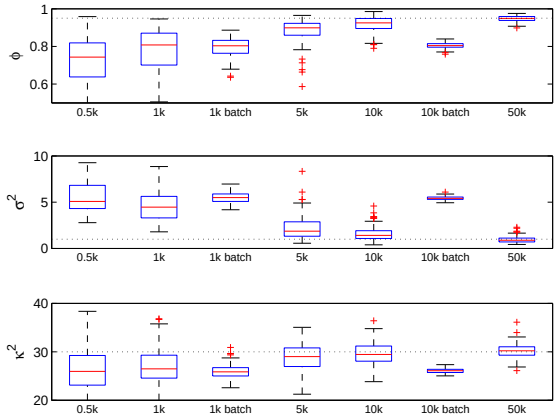

Fig. 1. Estimation results after 500 (0.5k), 1000 (1k), 5000 (5k), 10,000 (10k) and 50,000 (50k) observations together with batch EM estimates after 1000 (1k batch) and 10,000 (10k batch) observations.

In this example, Algorithm 2 was used for 100 independent realizations of length 100,000 of the above model with parameters $\phi = 0.95, \sigma^2 = 10, \kappa^2 = 20$. SMC was performed with $r = q_{\theta_n}$, i.e., using the so-called "bootstrap filter", with the current estimate of the parameters plugged-in. The step-size was decreased[2] as $\gamma_n = n^{-0.6}$ and the algorithm was systematically started from the initial values $\phi_0 = 0.8$, $\sigma_0^2 = 10$ and $\kappa_0^2 = 20$. The estimation results with $m = 100$ particles are summarized as box and whiskers plot in Figure 1. For comparison purpose, Figure 1 also

[1] This is the conditional form of the complete-data likelihood in which the term that depends only on the initial state and observation is discarded.

[2] Following the recommendations of [4] who also suggest using Polyak-Ruppert averaging which is not considered here for reasons of space.

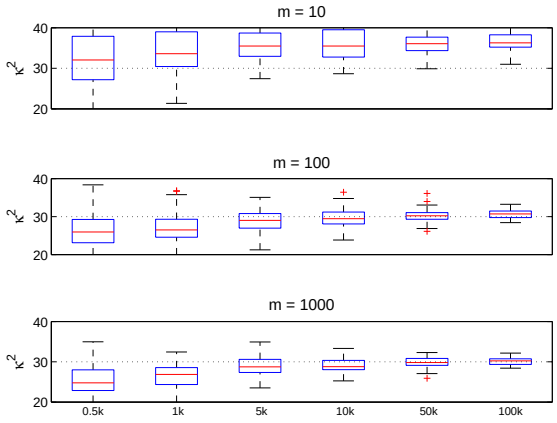

Fig. 2. Estimation results for κ^2 (from 500 to 100,000 observations) for different increasing values of m.

shows the results obtained after 20 iterations of the exact batch EM algorithm applied to the first 1000 and 10000 observations. The word "exact" refers to the fact that for the batch EM algorithm, smoothing is performed using Kalman filtering and disturbance smoothing and the obtained results are thus free from Monte Carlo error.

The proposed algorithm does converge as expected and does not exhibit any type of instability in the long term, even with a moderate number of $m = 100$ particles. As also observed in [4], the online algorithm does become preferable to using a fixed number of batch EM iterations when n increases. The apparent failure of batch EM is due to the fact that as n increases, the batch algorithm converges to a limiting algorithm which does requires several iterations to reach convergence [3]. Hence using a fixed number of batch EM iterations does not provide consistent estimates. In contrast, the online EM algorithm performs satisfactorily, despite the presence of Monte Carlo error due to the particle approximation.

Although barely noticeable on Figure 1, the online EM algorithm does present an asymptotic bias which is due to the use of the particle approximation. This is illustrated on Figure 2 which shows the long term behavior of the estimates of κ^2 for different values of m. The bias is clearly visible for $m = 10$, where the algorithm converges to a variance of about 36 instead of 30. The two other panels of Figure 2 show that this bias is eliminated when m increases. The variability of the estimate is also reduced although it tends to level off as m increases probably because the Monte Carlo errors are then dominated by the statistical variability.

As a second example, consider the standard nonlinear time series model considered, among many others, by [9, 5]:

$$X_t = a_t(X_{t-1}) + U_t, \qquad a_t(x) = \frac{x}{2} + 25\frac{x}{1+x^2} + 8\cos(1.2t)$$
$$Y_t = b(x_t) + V_t, \qquad b(x) = \frac{x^2}{20}$$

where (U_t) and (V_t) are independent Gaussian white noise sequences with standard deviations, σ_U and σ_V, respectively. Although the model is again simple it is nonetheless challenging when σ_V is much smaller than σ_U, as the filtering distribution is then distinctively bimodal rendering state identification very difficult.

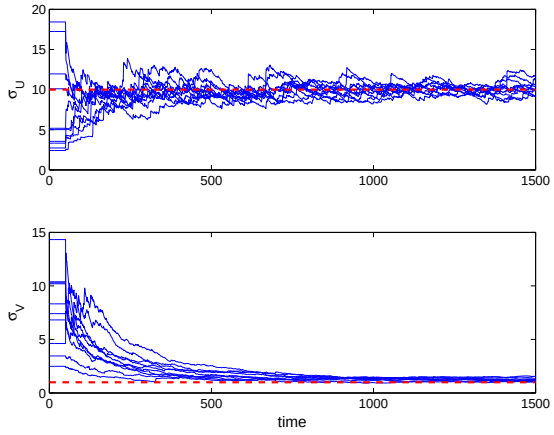

Fig. 3. Estimation results for σ_U and σ_V for 10 independent simulations.

In this case, Algorithm 2 performed best when proposing new particles from a robustified version of the prior chosen as a Student-t distribution with mean $a_t(X_{t-1})$, scale $\hat{\sigma}_{U,n}$ and four degrees of freedom. Alternatives based on using the EKF (Extended Kalman Filter) approximation [5] were found to be much less robust as these tend to be preferable when using accurate estimates of σ_U and σ_V but often turn out to be disastrous when the parameters are poorly matched to the model that generates the observations. Figure 3 displays the estimation results obtained with $m = 1000$ particles for 10 independents runs. Note that as in [4], adaptation is freezed for the first few observations (here 50) to allow for a minimal convergence of the estimates of the smoothed sufficient statistics. On this example, the proposed algorithm appears to be very robust with respect to the choice of the initialization and again converges with a low asymptotic bias.

From these experiments, it is quite remarkable that one can indeed obtain reliable estimates, even when using a non-increasing number of particles fixed to a reasonable value (typically, $m = 100$ in the first example and $m = 500$ in the second). The proposed approach does not seem to be plagued by long-term instabilities or degeneracies which affect some of the methods proposed so far for online Monte Carlo based estimation of fixed parameters. It is likely that the asymptotic bias can only be made arbitrarily small by increasing the number m of particles. But the approach appears to be sufficiently efficient to be usable in practice with values of m that are compatible with the constraints of real-time signal processing.

5. CONCLUSIONS

The online EM algorithm presented in this paper builds on the proposal of Mongillo and Denève (2008) but applies to general continuous state-space models thanks to the use of sequential Monte Carlo approximations. The resulting algorithm is generic, simple to implement and its performance in simulated scenarios is promising. Compared to related approaches, the use of stochastic approximation with a decreasing step-size γ_n does appear to warrant long-term stability, although a complete theoretical analysis of the procedure is still lacking. It is believed that the proposed algorithm could also be useful for HMMs with finite but large state spaces (eg., in digital communications applications) as the implementation cost of the exact online EM algorithm can be prohibitive in such cases [3].

6. REFERENCES

[1] C. Andrieu and A. Doucet. Online expectation-maximization type algorithms for parameter estimation in general state space models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, volume 6, pages 69–72, 2003.

[2] C. Andrieu, A. Doucet, and V. B. Tadic. Online simulation-based methods for parameter estimation in non linear non gaussian state-space models. In *Proc. IEEE Conf. Decis. Control*, 2005.

[3] O. Cappé. Online EM algorithm for hidden Markov models. preprint, 2009.

[4] O. Cappé and E. Moulines. On-line expectation-maximization algorithm for latent data models. *J. Roy. Statist. Soc. B*, 71(3):593–613, 2009.

[5] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[6] A. Doucet, N. De Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[7] R. J. Elliott, L. Aggoun, and J. B. Moore. *Hidden Markov Models: Estimation and Control*. Springer, New York, 1995.

[8] P. Fearnhead. Markov chain Monte Carlo, sufficient statistics and particle filter. *J. Comput. Graph. Statist.*, 11(4):848–862, 2002.

[9] N. Gordon, D. Salmond, and A. F. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Radar Signal Process.*, 140:107–113, 1993.

[10] A. Johansen, A. Doucet, and M. Davy. Maximum likelihood parameter estimation for latent variable models using sequential Monte Carlo. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006.

[11] V. Krishnamurthy and J. B. Moore. On-line estimation of hidden Markov model parameters based on the kullback-leibler information measure. *IEEE Trans. Signal Process.*, 41(8):2557–2573, 1993.

[12] J. Liu and M. West. Combined parameter and state estimation in simulation-based filtering. In A. Doucet, N. De Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[13] G. Mongillo and S. Denève. Online learning with hidden Markov models. *Neural Computation*, 20(7):1706–1716, 2008.

[14] T. Rydén. On recursive estimation for hidden Markov models. *Stochastic Process. Appl.*, 66(1):79–96, 1997.

[15] G. Storvik. Particle filters for state-space models with the presence of unknown static parameters. *IEEE Trans. Signal Process.*, pages 281–289, 2002.

[16] O. Zeitouni and A. Dembo. Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes. *IEEE Trans. Inform. Theory*, 34(4), July 1988.