ON A NEW STOCHASTIC VERSION OF THE EM ALGORITHM

Colin Campbell and Simon Godsill***
Signal Processing Group
Cambridge University Engineering Department
Tel: +44 1223 332600; Fax: +44 1223 332662
* *email:ncc21@eng.cam.ac.uk;* ** *email:sjg@eng.cam.ac.uk*

ABSTRACT

We present an algorithm in which the Maximisation step of the EM algorithm is replaced by a Sampling step. We describe an application of the algorithm to noise reduction for an audio signal. The results of various simulations on synthetic data are presented and compared to the results obtained using the EM algorithm and the Gibbs Sampler. A major limitation of the EM algorithm is that it can converge on local stationary points. The results we present show how our algorithm successfully overcomes this limitation.

1 INTRODUCTION

The EM Algorithm proposed by Dempster *et al.* [5] provides an iterative procedure for Maximum *a posteriori* estimation in the case of incomplete data. A major limitation of the EM algorithm is that whilst convergence to a stationary point of $p(\theta \mid \mathbf{y})$ can be shown [12], this is not necessarily the global maximum. Thus the choice of initial conditions is vital to the convergence of the algorithm [9, 15]. The motivation for implementing a stochastic version of EM is to overcome this limitation. Two stochastic EM algorithms of interest are the SEM Algorithm [4, 9] and the MCEM Algorithm [12, 14]. In MCEM, the analytic calculation of the E-step is replaced by a Monte Carlo approximation. In SEM, the Stochastic Imputation Principle is applied to simulate the unobserved data based on the observed data and the current value of the parameters [3]. Both of these algorithms are effectively replacing the analytic E-step with a stochastic step. The algorithm we propose seeks to replace the M-step with a simulation step. The simulation step should help the algorithm converge to the global maximum of the posterior independently of the initial conditions. We note that for many models of interest, including the model we discuss, it will be possible to calculate the E-step analytically and that this result can be used to give statistical stability. In particular we suggest that the new stochastic version of the EM algorithm can be embedded within a more complex MCMC framework in order to give greater statistical efficiency than standard Gibbs or Metropolis-Hastings samplers.

In Section 2 we describe the EM Algorithm and then propose a method of replacing the M-step with a sampling step. Section 3 describes an application of the proposed algorithm and in Section 4 we compare the results obtained using our method with those obtained using the EM Algorithm and Gibbs sampler. Finally in Section 5 we give our conclusions and propose further study using our algorithm.

2 THE ALGORITHM

The EM algorithm (Dempster *et al.* [5]) is an iterative method for finding the mode of the posterior. EM can, of course, also be used to calculate maximum likelihood estimates. Each iteration of the algorithm consists of an Expectation step (E-step) and a Maximisation step (M-step). Let θ^i be the current estimate of the parameter vector, then the E-step consists of computing,

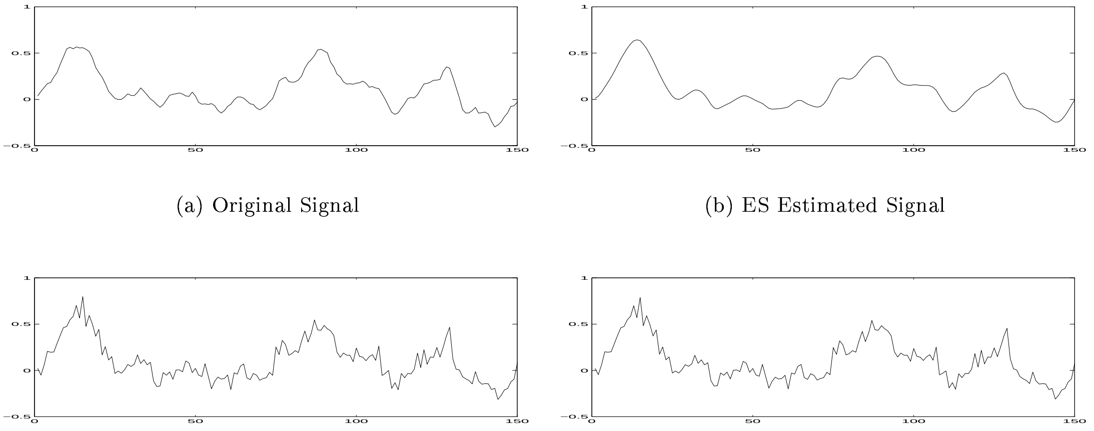
$$Q(\theta, \theta^i) = \int_{\Re^N} \log\left[p(\theta \mid \mathbf{y}, \mathbf{x})\right] p(\mathbf{x} \mid \mathbf{y}, \theta^i) d\mathbf{x} \qquad (1)$$

This is the Expectation of log $p(\theta \mid \mathbf{y}, \mathbf{x})$, the log augmented posterior, with respect to $p\left(\mathbf{x} \mid \mathbf{y}, \theta^i\right)$, the conditional predictive distribution of the latent unobserved data $\mathbf{x}$. In the M-step, $Q\left(\theta, \theta^i\right)$ is maximised with respect to θ to give θ^{i+1}. This process is then iterated until convergence. As stated previously, the EM algorithm converges to a stationary point of $p\left(\theta \mid \mathbf{y}\right)$ and if the posterior has multiple stationary points, the algorithm does not necessarily converge to the global maximum.

In order to overcome this limitation, we propose the following modification to the EM Algorithm which we call the Expectation-Sample (ES) algorithm. Instead of choosing θ^{i+1} to maximise $Q\left(\theta, \theta^i\right)$, we propose drawing a sample ϕ from $q\left(\theta \mid \theta^i\right) \propto \exp\left[Q\left(\theta, \theta^i\right)\right]$. In this way the algorithm becomes,

E-Step

$$Q(\theta, \theta^i) = \int_{\Re^N} \log\left[p(\theta \mid \mathbf{y}, \mathbf{x})\right] p(\mathbf{x} \mid \mathbf{y}, \theta^i) d\mathbf{x} \qquad (2)$$



(a) Original Signal

(b) ES Estimated Signal

(c) EM Estimated Signal

(d) Gibbs Sampler Estimated Signal

Figure 1: Results of using EM and ES Algorithms to estimate the data using the same initial conditions.

S-Step

$$q(\phi \mid \theta^i) \propto \exp\left(Q(\theta, \theta^i)\right) \tag{3}$$

Initially, we choose to assign $\theta^{i+1} = \phi$ and repeat the algorithm until convergence. This scheme can be seen as a variation of the Data Augmentation Algorithm of Tanner and Wong [13]. It is not clear what the stationary distribution of this Markov chain is, but empirical evidence suggests that the chain converges to a distribution similar to the posterior.

In order to ensure convergence to the true posterior $p(\theta \mid \mathbf{y})$, a Metropolis-Hastings step can be used. In this case the proposal density is $q(\theta \mid \theta^i)$ and the target density is the true posterior $p(\theta \mid \mathbf{y})$. Then given θ^i and $\phi \sim q(\theta \mid \theta^i)$ we accept $\theta^{i+1} = \phi$ with probability,

$$\alpha(\phi, \theta^i) = \min\left(1,\ \frac{p\left(\phi \mid \mathbf{y}\right) q\left(\theta^i \mid \phi\right)}{p\left(\theta^i \mid \mathbf{y}\right) q\left(\phi \mid \theta^i\right)}\right) \tag{4}$$

This step then ensures that the stationary distribution of the chain is $p(\theta \mid \mathbf{y})$.

The algorithm we have described here is applicable to conditionally Gaussian state-space models in which the transition equations depend linearly upon the states. Examples of this general class of models include Autoregressive models and the Autoregressive part of an Autoregressive-Moving Average model. For descriptions of such models see [2, 11].

In the next section we outline an application of the proposed algorithm that demonstrates its robustness.

3 APPLICATION

The application we present is that of noise reduction for a signal that can be modelled as an autoregressive process [6, 7, 8, 10].

This system is described by the following equations,

$$\mathbf{y} = \mathbf{x} + \mathbf{v} \tag{5}$$

$$\mathbf{x}_1 = \mathrm{X}\mathbf{a} + \mathbf{e} \tag{6}$$

where $\mathbf{e} = [e_{p+1}, \ldots, e_N]'$ is the excitation vector, $\mathbf{x} = [x_1, \ldots, x_N]'$ is the data vector and p is the order of the AR model. The vector $\mathbf{x}_1$ is $\mathbf{x}$ with the first p elements removed, and $\mathbf{a} = [a_1, \ldots, a_p]'$ is the vector of the AR coefficients. The rows of X are constructed in such a way as to form $x_t = \sum_{i=1}^{p} a_i x_{t-i} + e_t$ for successive samples, x_t.

If we define the augmented data as $\mathbf{s} = [\mathbf{y}'\ \mathbf{x}']'$ and assume conjugate priors [1] for $p(\mathbf{a})$, $p(\sigma_e^2)$ and $p(\sigma_v^2)$, where $\theta = [\mathbf{a}', \sigma_e^2, \sigma_v^2]'$ is the parameter vector, the E-step is then,

$$Q(\theta, \theta^i) \propto \int_{\Re^N} \log\left[p(\mathbf{y} \mid \mathbf{x}, \theta)\right] p(\mathbf{x} \mid \mathbf{y}, \theta^i) d\mathbf{x} + \int_{\Re^N} \log\left[p(\mathbf{x} \mid \theta)\right] p(\mathbf{x} \mid \mathbf{y}, \theta^i) d\mathbf{x} \tag{7}$$

and the S-step is

$${\sigma_v^2}^{i+1} \sim \mathrm{IG}\left(\sigma_v^2 \mid \frac{N}{2} - 1, \frac{\mathbf{y}'\mathbf{y} - 2\mathbf{y}'E[\mathbf{x}] + E[\mathbf{x}'\mathbf{x}]}{2}\right) \tag{8}$$

$${\sigma_e^2}^{i+1} \sim \mathrm{IG}\left(\sigma_e^2 \mid \frac{N}{2} - 1, \frac{E[\mathbf{x}_1'\mathbf{x}_1] - E[\mathbf{x}_1'\mathrm{X}]\mathbf{a}_{\mathrm{MAP}}}{2}\right) \tag{9}$$

$$\mathbf{a}^{i+1} \sim \mathrm{N}_p\left(\mathbf{a} \mid \mathbf{a}_{\mathrm{MAP}}, {\sigma_e^2}^{i+1} E[\mathrm{X}'\mathrm{X}]^{-1}\right) \tag{10}$$

where

$$\mathbf{a}_{\mathrm{MAP}} = E[\mathrm{X}'\mathrm{X}]^{-1} E[\mathrm{X}'\mathbf{x}_1] \tag{11}$$

The expectations in Equations (8)-(11) are implicitly conditioned on the latent unobserved data $\mathbf{x}$ and the current parameter estimate θ^i.

The algorithm is run to convergence (for our example we obtain the number of iterations for convergence empirically) and then we find the minimum mean square error estimate of the parameters using Monte Carlo integration,

$$\hat{\theta} = \frac{1}{N-m} \sum_{i=m+1}^{N} \theta^i \tag{12}$$

where m is the number of iterations required for convergence ('burn-in') and N is the total number of iterations. For our implementation of the algorithm, we have found the use of the Metropolis step results in low acceptance rates when the estimate is far from the mode thus for this application we have not used a Metropolis step. We suggest that to ensure convergence to the posterior $p(\theta \mid \mathbf{y})$, the algorithm can initially be run without a Metropolis step and once converged the Metropolis step can be used to draw samples from the posterior. Future work will involve establishing an expression for the stationary distribution of the Markov chain described by the algorithm.

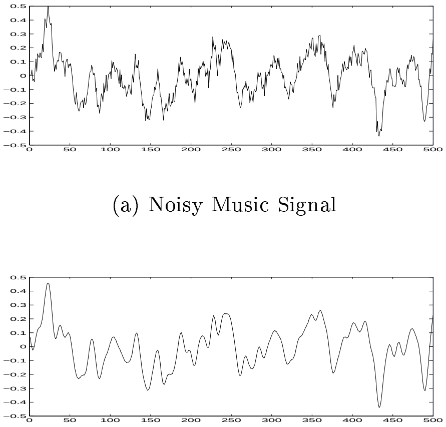
4 RESULTS

We performed a number of simulations using the system described in Section 3. We generated synthetic data using a 2nd order autoregressive process to which we added known white Gaussian noise. The reason for this approach is to enable us to compare the estimated parameters with the known parameter values. We can also compare the ES estimate of the parameters with the estimates obtained using the EM algorithm. Another interesting comparison for our algorithm is the Gibbs Sampler (see [6, 7, 8] for MCMC work with these models). Figure 1 shows the results of estimating the unobserved data for each of the three algorithms using the same initial conditions for the parameters. The results shown in Figure 1 indicate that our algorithm produces a smoothed estimate of the data $\mathbf{x}$, whilst the EM algorithm and the Gibbs sampler have not made any appreciable improvement of the noisy signal.

The results in Table 1 show our algorithm producing better estimates of the parameters than either the EM algorithm or Gibbs sampler for a particular set on initial conditions. The estimates of the parameters for the ES algorithm and the Gibbs sampler were found using Monte Carlo integration as described in Equation (12) where $N = 500$ and $m = 100$. The EM estimates of the parameters were the final estimates obtained by running the EM algorithm for 500 iterations. Once the final estimate of the parameters has been found, we use this to estimate the unobserved data $\mathbf{x}$. In all the simulations we performed our algorithm produced at least comparable results to the results of the other algorithms, and in a number of instances our algorithm performed better than EM or Gibbs sampling. In particular, for each simulation, our algorithm converged close to the true values of the parameters, which was not the case for EM or Gibbs sampling as illustrated by our example. This leads us to believe that our algorithm successfully negotiates local stationary points in the posterior density.

	Real	ES	EM	Gibbs
a_1	1.8	1.77	0.66	0.72
a_2	-0.81	-0.78	0.30	0.24
σ_e^2	1×10^{-3}	0.8×10^{-3}	10.7×10^{-3}	11.2×10^{-3}
σ_v^2	5×10^{-3}	5.2×10^{-3}	0.2×10^{-3}	0.7×10^{-3}

Table 1: Table showing the estimated parameter values compared with the real values for the ES, EM and Gibbs algorithms given the the initial conditions $\mathbf{a}^0 = [-0.18, -0.09]'$, $\sigma_e^{2^0} = 1 \times 10^{-5}$ and $\sigma_v^{2^0} = 1 \times 10^{-2}$.



(a) Noisy Music Signal

(b) 'Clean' ES Estimated Music Signal

Figure 2: Result of using the ES Algorithm for noise reduction of a real music signal.

Figure 2 shows the results of using our algorithm for noise reduction of a real music signal. The degraded signal was obtained by adding known white noise to a real music signal. The 'burn-in' for the algorithm was 10 iterations which we determined by experiment and we then used 20 iterations after convergence to obtain an estimate of the parameter vector - we have found that good results are obtained even for so few iterations. Our results have shown our algorithm to be successful in performing the task of noise reduction for this type of signal.

For the particular application we have described, unless strong prior information is available for σ_v^2 (see [6]) , the Gibbs Sampler fails to give satisfactory results whereas

EM and particularly ES produce much better results. We believe this is a result of the statistical stability that is provided by the analytic calculation of the E-step. This is what motivated us to consider a stochastic scheme in which we made use of this stability, but at the same time attempted to overcome the limitations of deterministic EM.

5 CONCLUSION

The results presented in the previous section show that our algorithm performs at least as well as the EM algorithm and the Gibbs sampler for a particular example. In some instances, our algorithm gives better results than either EM or Gibbs sampling. We conclude that the improvement over EM is a result of the stochastic nature of our algorithm that allows it to negotiate local stationary points of the posterior $p(\theta \mid \mathbf{y})$. Empirical evidence leads us to believe that our algorithm does, in fact, converge on a density that is similar to the posterior. Thus we conclude that our algorithm has successfully overcome the limitation that EM has of converging to local stationary points.

The fact that our algorithm performs better in some instances than the Gibbs sampler we believe is a result of the analytic calculation of the E-step which reduces the dimensionality of the sample space. The usefulness of this result is that it enables us to make use of the statistical stability of EM for certain models of interest. In particular we will extend these ideas to include the proposed algorithm as just one step within a large MCMC framework for parameter estimation. For example, the model can then be elaborated to include outliers, non-Gaussianity and non-stationarity, as in [6, 7, 8, 11].

Future work will involve the use of the algorithm presented here in larger MCMC frameworks. In particular, we are currently investigating the use of the ES algorithm for the source separation problem. Another aspect of future work will be to find the stationary distribution of the Markov chain described by the algorithm.

References

[1] G.E.P. Box and G. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, 1973.

[2] C.K. Carter and R. Kohn. On Gibbs sampling for sate space models. *Biometrika*, 81(3):541–553, 1994.

[3] G. Celeux, D. Chauveau, and J. Diebolt. On Stochastic Versions of the EM Algorithm. PROGRAMME 5 - Traitement du signal, automatique et productique - Rapport de recherche N^o 2514, INRIA Rhône-Alpes, 1995.

[4] G. Celeux and J. Diebolt. The SEM Algorithm: a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. *Computational Statistics Quarterly*, 2:73–82, 1985.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algortihm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[6] S.J. Godsill. Bayesian Enhancement of Speech and Audio Signals which can be Modeled as ARMA Processes. *International Statistical Review*, 65(1):1–21, 1997.

[7] S.J. Godsill. Robust modelling of noisy ARMA signals. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr 1997.

[8] S.J. Godsill and P.J.W. Rayner. Robust noise reduction for speech and audio signals. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1996.

[9] M. Lavielle. A stochastic algorithm for parametric and non-parametric estimation in the case of incomplete data. *IEEE Transactions on Signal Processing*, 42:3–17, 1995.

[10] J.S. Lim and A.V. Oppenheim. Enhancement and Bandwidth Compression of Noisy Speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

[11] N. Shephard. Partial non-Gaussian state space. *Biometrika*, 81(1):115–131, 1994.

[12] M.A. Tanner. *Tools for Statistical Inference.* Springer-Verlag, 3rd edition, 1996.

[13] M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550, 1987.

[14] G.C.G. Wei and M.A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, 85:699–704, 1990.

[15] C.J.F. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.