Monotonicity properties of the Monte Carlo EM algorithm and connections with simulated likelihood

By OMIROS PAPASPILIOPOULOS AND GIORGOS SERMAIDIS

Department of Statistics, Warwick University, Coventry, CV4 7AL, U.K.

O.Papaspiliopoulos@warwick.ac.uk, G.Sermaidis@warwick.ac.uk

SUMMARY

In this note we show that the Monte Carlo EM algorithm, appropriately constructed with importance re-weighting, monotonically increases a corresponding simulated likelihood. This is result is formally proved but also intuitively explained by a formulation of the problem using auxiliary variables.

Some keywords: Importance sampling, fixed random seeds, incomplete data, latent stochastic processes

1. Introduction

A fairly general description of the incomplete data framework is as follows. Let $\mathbb{Q}_\theta$ be the common probability distribution of a pair of processes (Y, X), assumed to be absolutely continuous with respect to a dominating measure $\mathbb{W}$ with density

$$\mathrm{d}\mathbb{Q}_\theta(Y, X) = \pi(Y, X \mid \theta)\, \mathrm{d}\mathbb{W}(Y, X)\,.$$

Without loss of generality we will assume that $\mathbb{W}$ is a probability measure. Only Y is observed, with corresponding marginal density

$$\pi(Y \mid \theta) = E[\pi(Y, X \mid \theta) \mid Y]\,, \tag{1}$$

where the expectation is taken with respect to $d\mathbb{W}(X \mid Y)$ and (1) is a density with respect to the marginal law $\mathrm{d}\mathbb{W}(Y)$. Y, X and (Y, X) are referred to as the observed, missing and complete data respectively, and $L(\theta) := \pi(Y \mid \theta)$ and $\pi(Y, X \mid \theta)$ as the observed and complete likelihood respectively. The aim is to infer θ on the basis of the

observed likelihood, the complication being that the latter is typically intractable due to the expectation involved in (1).

A popular deterministic algorithm for finding the maximum likelihood estimator and the observed information matrix is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). For a current estimate θ_{i-1}, the E-step of the ith iteration of the EM algorithm requires the computation of the following function of θ,

$$Q(\theta, \theta_{i-1}) = E[\log \pi(Y, X \mid \theta) \mid Y] \,, \tag{2}$$

where the expectation is taken with respect to $\mathrm{d}\mathbb{Q}_{\theta_{i-1}}(X \mid Y)$. The M-step replaces θ_{i-1} with some θ_i, such that $Q(\theta_i, \theta_{i-1}) > Q(\theta_{i-1}, \theta_{i-1})$. Each iteration of the algorithm increases the observed likelihood, $L(\theta_i) \geq L(\theta_{i-1})$, thus under regularity conditions (Wu, 1983) the θ_is are guaranteed to converge to local maximizer of the likelihood. The EM algorithm is a very useful tool when both the E- and M-step can be performed analytically, e.g. when the complete likelihood is in the regular exponential family. However, in many applications, although the complete likelihood is explicit, it is not possible to perform the E-step analytically. A variety of stochastic algorithms which involve simulation from $\mathrm{d}\mathbb{Q}_\theta(X \mid Y)$ have been developed for this case (see for example Jank, 2006, for a recent review). We focus on the simulated likelihood (Geyer, 1994) and the Monte Carlo EM algorithm (Wei & Tanner, 1990).

Let $X_1, \ldots, X_N$ be a stationary sequence with marginal law $\mathrm{d}\mathbb{W}(X \mid Y)$ and define

$$L^{(N)}(\theta) := \sum_{j=1}^{N} \pi(Y, X_j \mid \theta) \,, \tag{3}$$

$$Q^{(N)}(\theta, \theta_{i-1}) := \sum_{j=1}^{N} \log \pi(Y, X_j \mid \theta)\, \pi(Y, X_j \mid \theta_{i-1}) \Big/ L^{(N)}(\theta_{i-1}) \ . \tag{4}$$

Notice that $L^{(N)}(\theta)/N$ is an unbiased estimator of (1), thus it forms a simulated likelihood in the sense of Geyer (1994). Additionally, (4) is an unbiased estimator of (2) thus it defines a Monte Carlo EM algorithm, where samples from $\mathrm{d}\mathbb{Q}_{\theta_{i-1}}(X \mid Y)$ are obtained by importance sampling with proposals from $\mathrm{d}\mathbb{W}(X \mid Y)$. The maximizer of (3) under certain conditions (Geyer, 1994; Beskos et al., 2007) converges to the maximum

likelihood estimator as $N \to \infty$. The Monte Carlo EM algorithm does not anymore monotonically increase the likelihood due to the extra randomness introduced by the simulations in the E-step. As a result it is necessary to increase appropriately N with the iterations to achieve convergence (Chan & Ledolter, 1995; Fort & Moulines, 2003). Additionally, it is hard to devise stopping rules for the algorithm, because of its random oscillations once it reaches a high likelihood region. McCulloch (1997) compares the two methods empirically in the context of generalised linear mixed models and advocates a combination of both as a good strategy.

Notice that we have used importance re-weighting to construct the Monte Carlo E-step. This has been considered before, see in particular Quintana et al. (1999); Levine & Casella (2001) for computational reasons when Markov chain Monte Carlo is used to produce the X_js, since the same draws can be used for multiple iterations. Our main result contained in Theorem 1 states that if the X_js are fixed throughout the iterations of the Monte Carlo EM algorithm, then $L^{(N)}(\theta_i) \geq L^{(N)}(\theta_{i-1})$. Thus, this Monte Carlo EM monotonically increases the corresponding simulated likelihood. We obtain a similar result for monotonic increase of the likelihood ratio $L^{(N)}(\theta)/L^{(N)}(\theta_0)$ when we choose $\mathbb{W} = \mathbb{Q}_{\theta_0}$ for some fixed θ_0.

2. The main result

Theorem 1. *Let $X_1, \ldots, X_N$ be a stationary sequence with marginal law $\mathrm{d}\mathbb{W}(X \mid Y)$, and $L^{(N)}$ and $Q^{(N)}$ defined as in (3) and (4) respectively. We assume that the X_js are kept fixed throughout the iterations of the Monte Carlo EM algorithm. Then, for any pair (θ_{i-1}, θ_i) such that $Q^{(N)}(\theta_i, \theta_{i-1}) > Q^{(N)}(\theta_{i-1}, \theta_{i-1})$ it holds that $L^{(N)}(\theta_i) > L^{(N)}(\theta_{i-1})$.*

Proof. Let

$$w_j(\theta) := \pi(Y, X_j \mid \theta)\,, \quad \pi_j(\theta) = w_j(\theta) / \sum_{j=1}^{N} w_j(\theta)\,.$$

Then, a direct calculation gives

$$Q^{(N)}(\theta, \theta_{i-1}) = \sum_{j=1}^{N} \log \pi_j(\theta)\pi_j(\theta_{i-1}) + \log L^{(N)}(\theta)\,.$$

Then, any pair (θ_{i-1}, θ_i) such that $Q^{(N)}(\theta_i, \theta_{i-1}) > Q^{(N)}(\theta_{i-1}, \theta_{i-1})$ implies that

$$\sum_{j=1}^{N} \log \frac{\pi_j(\theta_i)}{\pi_j(\theta_{i-1})} \pi_j(\theta_{i-1}) > \log \frac{L^{(N)}(\theta_{i-1})}{L^{(N)}(\theta_i)},$$

which by Jensen's inequality proves the result. □

We have two observations on this main result. Firstly, a similar argument can be employed when $\mathbb{W} = \mathbb{Q}_{\theta_0}$ for some fixed θ_0, to yield that the Monte Carlo EM increases monotonically the Monte Carlo likelihood ratio $L^{(N)}(\theta)/L^{(N)}(\theta_0)$. Secondly, there is an alternative way to prove and motivate this monotonicity property. One can view the simulated likelihood $L^{(N)}(\theta)/N$ as a marginal of the following "complete" likelihood

$$\pi(Y, X_1, \ldots, X_N, J = j \mid \theta) = \frac{1}{N} \pi(Y, X_j \mid \theta), \tag{5}$$

where this density is with respect to the product of the counting measure on $\{1, \ldots, N\}$, $\mathrm{d}\mathbb{W}(Y)$ and $(\mathrm{d}\mathbb{W}(X \mid Y))^N$. A direct calculation verifies that (5) indeed defines a probability measure (this is particularly easy to check if one assumes that both $\mathbb{Q}_\theta$ and $\mathbb{W}$ have densities with respect to another dominating measure, say the Lebesgue measure). In this formulation, J can be treated as missing data, whereas Y and $X_1, \ldots, X_N$ as observed data. Then, one can check that up to a constant (4) is the corresponding Q function of an ordinary EM algorithm for finding the maximizer of (5).

Acknowledgements

The second author would like to thank the Greek State Scholarship Foundation (IKY) for financial support.

References

Beskos, A., Papaspiliopoulos, O. & Roberts, G. O. (2007). Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *Ann. Statist.* To appear.

Chan, K. S. & Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *J. Amer. Statist. Assoc.* 90 242–252.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39 1–38. With discussion.

Fort, G. & Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.* 31 1220–1259.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc. Ser. B* 56 261–274.

Jank, W. (2006). The EM algorithm, its stochastic implementation and global optimization: Some challenges and opportunities for OR. Available from http://www.smith.umd.edu/faculty/wjank/GA-EM-SaulGass.pdf.

Levine, R. A. & Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Statist.* 10 422–439.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* 92 162–170.

Quintana, F., Liu, J. & del Pino, G. (1999). Monte Carlo EM with importance reweighting and its applications in random effects models. *Computational Statistics & Data Analysis* 29 429–444.

Wei, G. C. G. & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* 85 699–704.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* 11 95–103.