MARGINAL MAP ESTIMATION USING MARKOV CHAIN MONTE CARLO

Christian P. Robert

Statistics Laboratory, CREST, INSEE,
92245 Malakoff, Cedex, France
robert@bayes.ensae.fr

Arnaud Doucet *Simon J. Godsill*

Signal Processing Group
Cambridge University Engineering Dept.
{ad2,sjg}@eng.cam.ac.uk

ABSTRACT

Markov chain Monte Carlo (MCMC) methods are powerful simulation-based techniques for sampling from high-dimensional and/or non-standard probability distributions. These methods have recently become very popular in the statistical and signal processing communities as they allow highly complex inference problems in detection and estimation to be addressed. However, MCMC is not currently well adapted to the problem of marginal *maximum a posteriori* (MMAP) estimation. In this paper, we present a simple and novel MCMC strategy, called State-Augmentation for Marginal Estimation (SAME), that allows MMAP estimates to be obtained for Bayesian models. The methodology is very general and we illustrate the simplicity and utility of the approach by examples in MAP parameter estimation for Hidden Markov models (HMMs) and for missing data interpolation in autoregressive time series.

1. INTRODUCTION

When performing complicated statistical signal detection and estimation, probabilistic models are typically specified that involve high-dimensional unknown parameter vectors. Very often the values of some of these parameters are not required in the inference task, for example the amplitude of frequency components in a frequency estimation application. Within a Bayesian setting, these so-called nuisance parameters are integrated out from the posterior distribution using the marginalization identity and inference is then performed in terms of only the required parameters [10], [12].

Consider the following Bayesian model: $\theta = (\theta_1, \theta_2)$ is a random parameter with prior distribution $p(\theta)$ and the likelihood of the observations $\mathbf{y}$ is given by $p(\mathbf{y}|\theta)$. Bayesian inference is based on the posterior distribution $p(\theta|\mathbf{y})$, which is given by Bayes' theorem:

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)\, p(\theta). \tag{1}$$

In this paper, the aim is to obtain the marginal MAP (MMAP) estimator for θ_1:

$$\theta_1^{MMAP} = \underset{\Theta_1}{\operatorname{argmax}}\, p(\theta_1|\mathbf{y}) \tag{2}$$

where the nuisance parameters θ_2 have been integrated out:

$$p(\theta_1|\mathbf{y}) = \int_{\Theta_2} p(\theta_1, \theta_2|\mathbf{y})\, d\theta_2. \tag{3}$$

Estimating θ_1^{MMAP} is a complex problem as typically neither the maximization (2) nor the integration (3) can be performed analytically.

In order to solve problems of this type, algorithms such as Expectation-Maximization (EM) and its stochastic variants [3], [16] are available. These methods aim to maximize the marginal likelihood for the required parameters. They are easily adapted for Bayesian MMAP estimation by the introduction of a prior penalization term in the maximization step. However, these EM-based methods are prone to estimate local rather than global modes of the marginal posterior distribution and are not easily adapted to the highly complex modelling requirements of many statistical signal processing problems. MCMC methods [5], on the other hand, which can be routinely applied to problems of high complexity, have not until now been developed for the solution of the MMAP estimation problem.

MCMC methods are a broad class of algorithms for sampling from complex, non-standard probability distributions. Since their introduction in applied statistics at the beginning of the 90's, they have become very popular as they allow for the solution of complex problems in Bayesian signal processing [10] and statistics [12], [13] where typically the posterior distributions and estimators of interest do not admit any closed form solution. In a Bayesian framework, the key idea of MCMC methods is to run a homogeneous Markov chain whose invariant distribution is the posterior distribution of interest. Under mild conditions on its transition kernel the Markov chain converges towards the required target distribution[15]. One can then use the simulated samples to estimate the posterior distribution and its features such as posterior means and variances. However, in their usual form these methods cannot be used to perform optimization of marginal posterior distributions.

In this article, we propose an original MCMC-based strategy for maximization of marginal posterior distributions. This strategy introduces an artificially augmented probability model, whose sampling leads to MMAP estimation of the required parameters. The algorithm is conceptually very simple and straightforward to implement in most cases, requiring only small modifications to MCMC code written for the original model. In its basic form the algorithm is directly applicable only to the standard MMAP problem. We believe however that it will be easily adaptable to solving more general inference with respect to Bayesian utility functions, using the extended utility variable methods of Müller [9].

The paper is organized as follows. In Section 2 the new MCMC strategy is described for the solution of the MMAP problem, which we call the State Augmentation for Marginal Estimation (SAME) method. In Section 3, the method is applied to MAP parameter

estimation of hidden Markov models for blind equalization of digital communications channels. In Section 4 we give an example of estimating missing data in autoregressive time series, a problem which occurs in speech and audio processing [8][7][6].

2. MCMC STRATEGIES

2.1. Standard MCMC approaches

The standard MCMC approach draws a large number N of (dependent, approximate) samples $\left\{\left(\theta_1^{(i)}, \theta_2^{(i)}\right); i = 1, \ldots, N\right\}$ from the joint posterior distribution $p(\theta_1, \theta_2 | \mathbf{y})$. As a consequence, $\left\{\theta_1^{(i)}; i = 1, \ldots, N\right\}$ are drawn from the marginal distribution $p(\theta_1 | \mathbf{y})$. If $p(\theta_1 | \mathbf{y})$ can be evaluated analytically, a possible estimate for θ_1^{MMAP} is then simply that $\theta_1^{(i)}$ which has maximum posterior probability $p\left(\theta_1^{(i)} \middle| \mathbf{y}\right)$. This method is not efficient in the sense that random samples from $p\left(\theta_1^{(i)} \middle| \mathbf{y}\right)$ only rarely explore the vicinity of the mode, unless the posterior has large probability mass around the mode – much computation is thus wasted exploring areas of no interest for MMAP estimation. Moreover the marginal posterior $p(\theta_1 | \mathbf{y})$ will often be unavailable in closed form; then histogram or kernel methods must be employed. These are unsuitable for high-dimensional parameters, being very sensitive to the choice of grid and kernels functions.

2.2. State Augmentation for Marginal Estimation (SAME)

We present here an alternative simulation-based strategy that is related to the standard simulated annealing algorithm. As in simulated annealing we replace the target distribution $p(\theta_1 | \mathbf{y})$ by the distribution $\overline{p}^{\gamma}(\theta_1 | \mathbf{y}) \propto p^{\gamma}(\theta_1 | \mathbf{y})$. It is well known that as $\gamma \to +\infty$, so $\overline{p}^{\gamma}(\theta_1 | \mathbf{y})$ becomes concentrated on the set of global maxima of $p(\theta_1 | \mathbf{y})$. In the classical simulated annealing framework, sampling from $\overline{p}^{\gamma}(\theta_1 | \mathbf{y})$ is realized by using a Metropolis-Hastings or a Gibbs sampler. However, such an algorithm cannot be developed when one is not able to evaluate $p(\theta_1 | \mathbf{y})$ analytically (up to a normalizing constant), and may be hard to construct effectively even when the marginal is available.

We propose here a novel approach based on a different idea which has very general applicability. It consists of defining an artificial probability model whose marginal distribution is the concentrated distribution $\overline{p}^{\gamma}(\theta_1 | \mathbf{y})$. If samples $\theta_1^{(i)}$ can be drawn from this concentrated distribution, then as γ becomes large the samples will be concentrated around the mode. We now show how to achieve this by artificially replicating the nuisance parameters in the model. Let us augment the model with $\gamma \in \mathbb{N}^*$ artificial replications of θ_2, denoted by $\theta_2(1), \ldots, \theta_2(\gamma)$. Each of these replications is now treated as a distinct random variable in its own right and the following joint distribution is defined:

$$q_{\gamma}(\theta_1, \theta_2(1), \ldots, \theta_2(\gamma) | \mathbf{y}) \propto \prod_{k=1}^{\gamma} p(\theta_1, \theta_2(k) | \mathbf{y}) \qquad (4)$$

By construction, the marginal for θ_1 in this distribution is

$$q_{\gamma}(\theta_1 | \mathbf{y}) = \overline{p}^{\gamma}(\theta_1 | \mathbf{y}) \propto p^{\gamma}(\theta_1 | \mathbf{y}). \qquad (5)$$

So, if we build a MCMC algorithm in the augmented space, with invariant distribution $q_{\gamma}(\theta_1, \theta_2(1), \ldots, \theta_2(\gamma) | \mathbf{y})$, then the simulated sequence $\left\{\theta_1^{(i)}; i \in \mathbb{N}\right\}$ will be drawn from the marginal posterior of interest, $\overline{p}^{\gamma}(\theta_1 | \mathbf{y})$. An important point here is that when a MCMC sampler is available to sample from $p(\theta_1, \theta_2 | \mathbf{y})$ then it is usually very easy to construct a MCMC sampler to sample from $q_{\gamma}(\theta_1, \theta_2(1), \ldots, \theta_2(\gamma) | \mathbf{y})$, especially if $p(\theta_1 | \mathbf{y}, \theta_2)$ is a distribution from the exponential family, or when it can be simulated using *slice sampling* (see [13]).

To implement the algorithm practically, it is likely to be benificial to employ a cooling schedule in a similar fashion to standard simulated annealing. Thus γ would be increased over iterations, i.e. $\gamma = \gamma(i)$ with $\lim_{i \to +\infty} \gamma(i) = +\infty$ and perhaps $\gamma(1) = 1$. Choice of optimal cooling schedules remains a topic for future work. As mentioned already, the scheme we present is not an algorithm but a general strategy that has to be adapted on a case by case basis. In the following sections, we detail possible schemes for its applications to two signal processing problems.

3. APPLICATION TO BLIND EQUALIZATION OF COMMUNICATIONS CHANNELS

3.1. Signal Model and Estimation Objectives

The channel input sequence x_t is assumed to be an i.i.d. sequence with known (discrete) state space $\mathcal{X}$. This signal is passed through a FIR channel of length L. We denote its impulse response by $\mathbf{h} = [h_0, \ldots, h_{L-1}]^{\mathrm{T}}$ and the state vector at time sampling instant t by $\mathbf{x}_t = [x_t, \ldots, x_{t-L+1}]^{\mathrm{T}}$. The observed signal y_t is the baseband output of the channel corrupted by additive noise n_t:

$$y_t = \mathbf{h}^{\mathrm{T}} \mathbf{x}_t + n_t$$

where n_t is assumed to be white and Gaussian with variance σ^2. $(\mathbf{x}_t, y_t)_{t \in \mathbb{N}}$ is a hidden Markov model [11]. The signal x_t and the parameters $\theta_1 = (\mathbf{h}, \sigma^2)$ are assumed unknown.

In a fully Bayesian framework, we assign a prior distribution not only to the signal x_t, but also to the unknown parameters θ_1. For the channel and noise variance, a classical normal-inverse gamma prior distribution is selected, i.e. $\mathbf{h} | \sigma^2 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma_0)$ and $\sigma^2 \sim \mathcal{IG}(\eta_0/2, \nu_0/2)$, with Σ_0 regular.

Given the set of observations $\mathbf{y}_{1:T} \triangleq \{y_1, \ldots, y_T\}$, our aim is to estimate θ_1 in a MMAP sense, i.e. obtaining $\theta_1^{MMAP} = \operatorname{argmax} p(\theta_1 | \mathbf{y}_{1:T})$. In this example, there is no integration problem as $p(\theta_1 | \mathbf{y}_{1:T})$ can be evaluated pointwise up to a normalizing constant. However, it is well-known that maximizing this distribution is a very complex problem. The most popular method is without doubt the EM algorithm. However this deterministic method is quite sensitive to initialization, which is why several stochastic versions of the EM have been proposed in the literature, see [1] for a review of these methods applied to the blind deconvolution problem.

3.2. MAP parameter estimation

We first present a standard MCMC algorithm to sample from $p(\theta_1 | \mathbf{y}_{1:T})$ and then we show how this algorithm can be straightforwardly modified to a SAME algorithm which optimizes $p(\theta_1 | \mathbf{y}_{1:T})$.

3.2.1. Algorithms

To sample from $p(\theta_1 | \mathbf{y}_{1:T})$ we treat the unobserved state sequence as nuisance parameters, i.e. $\theta_2 = \mathbf{x}_{1:T} \triangleq \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$

and perform MCMC to draw samples from the joint distribution $p(\theta_1, \mathbf{x}_{1:T}|\mathbf{y}_{1:T})$. To sample from $p(\theta_1, \mathbf{x}_{1:T}|\mathbf{y}_{1:T})$, we use a Data Augmentation MCMC algorithm:

Standard MCMC

1. Initialization, $i = 0$. Set randomly $\theta_1^{(0)}$.
2. Iteration i, $i \geq 1$

- Sample $\mathbf{x}_{1:T}^{(i)} \sim p\left(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}, \theta^{(i-1)}\right)$.
- Sample $\theta_1^{(i)} \sim p\left(\theta_1|\mathbf{y}_{1:T}, \mathbf{x}_{1:T}^{(i)}\right)$.

Now, following the strategy presented in Section 2.2, the modified algorithm to optimize $p(\theta_1|\mathbf{y}_{1:T})$ proceeds as follows.

SAME algorithm

1. Initialization, $i = 0$. Set randomly $\theta_1^{(0)}$.
2. Iteration i, $i \geq 1$

- Sample $\mathbf{x}_{1:T}^{(i)}(j) \sim p\left(\mathbf{x}_{1:T}(j)|\mathbf{y}_{1:T}, \theta_1^{(i-1)}\right)$ for $j = 1, \ldots, \gamma(i)$.
- Sample $\theta_1^{(i)} \sim q_{\gamma(i)}\left(\theta|\mathbf{y}_{1:T}, \mathbf{x}_{1:T}^{(i)}(1), \ldots, \mathbf{x}_{1:T}^{(i)}(\gamma(i))\right)$.

---

where $\gamma(i)$ is an increasing sequence of positive integers satisfying $\lim_{i \to +\infty} \gamma(i) = +\infty$.

3.2.2. Implementation issues

To implement these algorithms, it is necessary to be able to sample from $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}, \theta_1)$, $p(\theta_1|\mathbf{y}_{1:T}, \mathbf{x}_{1:T})$ and $q_{\gamma(i)}(\theta_1|\mathbf{y}_{1:T}, \mathbf{x}_{1:T}(1), \ldots, \mathbf{x}_{1:T}(\gamma(i)))$. Sampling from $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}, \theta_1)$ can be realized using the efficient forward filtering-backward sampling method [2], [4] whose computational complexity is $O(T)$. For $p(\theta_1|\mathbf{y}_{1:T}, \mathbf{x}_{1:T})$, one obtains

$$\begin{aligned} \mathbf{h}|\sigma^2 &\sim \mathcal{N}(\mathbf{m}, \sigma^2\Sigma) \\ \sigma^2 &\sim \mathcal{IG}\left((\eta_0 + T)/2, (\nu_0 + \mathbf{y}_{1:T}\mathbf{y}_{1:T}^{\mathrm{T}} - \mathbf{m}^{\mathrm{T}}\Sigma^{-1}\mathbf{m})/2\right) \end{aligned}$$

$$\text{where} \quad \Sigma^{-1} = \Sigma_0^{-1} + \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t^{\mathrm{T}} \quad \text{and} \quad \mathbf{m} = \Sigma\sum_{t=1}^{T} \mathbf{x}_t y_t$$

Sampling from $q_{\gamma(i)}(\theta_1|\mathbf{y}_T, \mathbf{x}_{1:T}(1), \ldots, \mathbf{x}_{1:T}(\gamma(i)))$ can be easily done since

$$\begin{aligned} &q_{\gamma(i)}(\theta_1|\mathbf{y}_T, \mathbf{x}_{1:T}(1), \ldots, \mathbf{x}_{1:T}(\gamma(i))) \\ &\propto \prod_{k=1}^{\gamma} p(\mathbf{y}|\theta_1, \mathbf{x}_{1:T}(k))\, p(\theta_1) \end{aligned}$$

One obtains

$$\begin{aligned} \mathbf{h}|\sigma^2 &\sim \mathcal{N}(\mathbf{m}(i), \sigma^2\Sigma(i)) \\ \sigma^2 &\sim \mathcal{IG}(\gamma(i)((\eta_0 + T)/2) + (\gamma(i) - 1)(L/2 + 1), \\ &\quad (\gamma(i)(\nu_0 + \mathbf{y}_{1:T}\mathbf{y}_{1:T}^{\mathrm{T}}) - \mathbf{m}^{\mathrm{T}}(i)\Sigma^{-1}(i)\mathbf{m}(i))/2) \end{aligned}$$

$$\begin{aligned} \text{where} \quad \Sigma^{-1}(i) &= \gamma(i)\Sigma_0^{-1} + \sum_{k=1}^{\gamma(i)}\sum_{t=1}^{T} \mathbf{x}_t(k)\mathbf{x}_t^{\mathrm{T}}(k) \\ \text{and} \quad \mathbf{m}(i) &= \Sigma(i)\sum_{k=1}^{\gamma(i)}\sum_{t=1}^{T} \mathbf{x}_t(k)\, y_t \end{aligned}$$

The resulting stochastic optimization method has a nice interpretation for this simple example and also the interpolation example of the next section. At the first iteration ($\gamma(1) = 1$), we have the standard data augmentation algorithm and, as $\gamma(i) \to +\infty$, the algorithm operates as a stochastic EM algorithm.Simulations show that the method is very effective for this model. For example, see figure 1 in which we have plotted the log-posterior probability $\log(p(\theta_1|\mathbf{y}_{1:T}))$ against iteration number for both standard EM (dotted line) and the SAME (solid line) algorithm. EM has converged very fast to a local probability maximum. Note however that the SAME algorithm is slower to converge for the cooling schedule we chose ($\gamma(i) = i$) but achieves a much more probable final solution, which is indicative of the power of the approach.

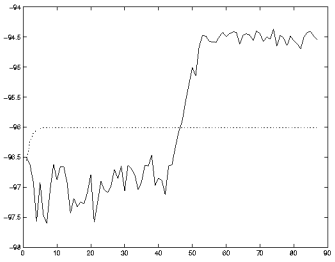

Figure 1: Log posterior probablity vs. iteration number for the communications model using EM (dotted) and SAME (solid) algorithms

4. APPLICATION TO MISSING DATA ESTIMATION IN AUTOREGRESSIVE TIME SERIES

We now consider a problem which finds application in the replacement of missing data packets in speech signals and the restoration of audio time series [8][7][6]. The data sequence x_t is assumed to be drawn from an autoregressive (AR) process with coefficients $\mathbf{a} = [a_0, \ldots, a_{P-1}]$ and the state vector at time sampling instant t is denoted by $\mathbf{x}_t = [x_t, \ldots, x_{t-P+1}]^{\mathrm{T}}$. The model is written as

$$x_t = \mathbf{a}^{\mathrm{T}}\mathbf{x}_{t-1} + e_t,$$

where e_t is assumed to be a white, Gaussian excitation sequence with variance σ^2. The signal x_t is assumed unobserved (missing) at sampling points $\mathcal{I} = \{i_1, ..., i_l\} \subset \{1, ..., T\}$. The observed data is $\mathbf{x}_{-\mathcal{I}} \triangleq \{x_t; t \in \{1, ..., T\} - \mathcal{I}\}$, the missing data is $\mathbf{x}_{\mathcal{I}} \triangleq \{x_t; t \in \mathcal{I}\}$ and the parameters $\theta_2 = (\mathbf{a}, \sigma^2)$ are assumed unknown. Exactly as above, we assign a normal-inverse gamma prior distribution for $\theta_2 = (\mathbf{a}, \sigma^2)$. In this case, however, we wish to estimate the missing data, $\theta_1 = \mathbf{x}_{\mathcal{I}}$, in the MMAP sense. A data augmentation MCMC algorithm is easily constructed for this problem [10][7][6] that has a very similar structure to the above equalization scheme, involving draws from $p(\mathbf{x}_{\mathcal{I}}|\mathbf{x}_{-\mathcal{I}}, \theta_2)$ and $p(\theta_2|\mathbf{x}_{1:T})$, which are multivariate normal and normal-inverse gamma distributions, respectively. $p(\mathbf{x}_{\mathcal{I}}|\mathbf{x}_{-\mathcal{I}}, \theta_2)$ can be sampled efficiently using forward-backward simulators [2] in cases where the number of missing data points is large. In order to obtain

the corresponding SAME algorithm, we augment the probability model with artificial θ_2 variables and draw samples from $q_{\gamma(i)}\left(\mathbf{x}_{\mathcal{I}}|\,\mathbf{x}_{-\mathcal{I}},\theta_2(1),\ldots,\theta_2\left(\gamma\left(i\right)\right)\right)$ and $p\left(\theta_2|\,\mathbf{x}_{1:T}\right)$. The first term $q_{\gamma(i)}\left(\mathbf{x}_{\mathcal{I}}|\,\mathbf{x}_{-\mathcal{I}},\theta_2(1),\ldots,\theta_2\left(\gamma\left(i\right)\right)\right)$ retains a similar multivariate Gaussian form to its counterpart $p\left(\mathbf{x}_{\mathcal{I}}|\mathbf{x}_{-\mathcal{I}},\theta_2\right)$ and so is readily implemented using only minor code modifications. The formulae are similar in structure to the equalization example, so we do not repeat the details here.

For this model we test the new method in a difficult scenario where 50% of the data are missing in the middle of a short block of length $T = 40$, extracted from a real music signal. The AR model order is fixed at $P = 9$ and once again both EM and SAME algorithms are applied to the data. The starting guess for the missing data was the all-zero vector. Very diffuse priors approaching the non-informative limit are chosen. In this case we also compare results with the standard Gibbs sampler data augmentation algorithm, used here to estimate the posterior mean of the missing data. The top plot in figure 2 shows the estimated waveforms for the three methods. The bottom plot shows the posterior probabilities against iteration number, as before. It is clear from this that the three methods give very different results and that the SAME algorithm finds a significantly more probable solution than either of the other techniques. It is also clear from the probability plots that the Gibbs sampler would be quite inappropriate for performing optimisation of this model since it virtually never achieves probabilities close to the maximum. EM has converged to a local but not global stationary point.

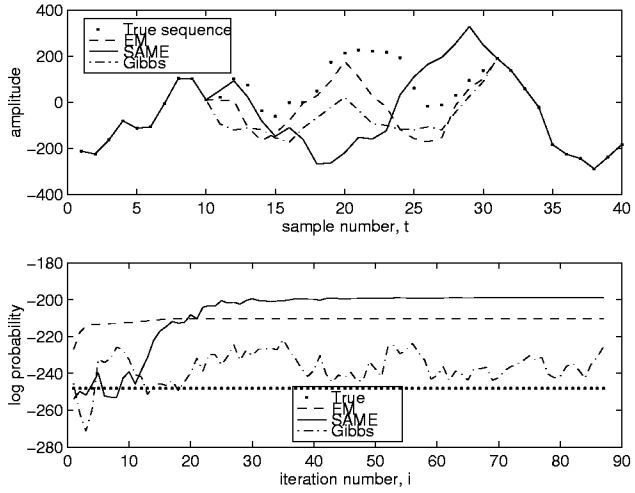

Figure 2: Simulations for missing data in AR model

4.1. Discussion

In this article, we have presented an original simulation-based strategy to maximize marginal posterior distributions. This method is closely linked to simulated annealing. However, contrary to classical annealing algorithms, it is based on the introduction of an artificial augmented probability model. Once a MCMC algorithm is available to sample from a posterior distribution, the proposed algorithms are very simple to implement in most cases. This methodology has been applied to statistical signal processing problems. Computer simulations demonstrate the effectiveness of our method compared with EM and standard MCMC approaches. We have chosen to demonstrate the new methods for models which have an analytic EM algorithm and for which the posterior probability is available in closed form. This allows a direct comparison to be made with EM, which is aimed at solving exactly the same problem as our SAME procedure. Some significant improvements have been demonstrated for testing datasets. However, we envisage that the real benefits of this new method are that it can be routinely applied to more complex models where standard MCMC but not EM algorithms are available. We have tested the SAME algorithm using such models and found the results to be equally promising and will report on these results more fully in a later publication.

5. REFERENCES

[1] O. Cappé, A. Doucet, É. Moulines and M. Lavielle, "Simulation-Based Methods for Blind Maximum-Likelihood Filter Identification", to appear *Signal Processing*, 1998.

[2] C.K. Carter and R. Kohn, "On Gibbs sampling for state space models", *Biometrika*, vol. 81, pp. 541-553, 1994.

[3] G. Celeux and J. Diebolt, "Une version recuit simulé de l'algorithme EM", *C.R.A.S.*, vol. 310, pp. 119-124, 1990.

[4] S. Chib, "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models", *J. Econometrics*, vol. 75, pp. 79-97, 1996.

[5] W. R. Gilks, S. Richardson and D. J. Spiegelhalter eds., *Markov Chain Monte Carlo in Practice,* Chapman and Hall, 1996.

[6] S.J. Godsill and P.J.W. Rayner, *Digital Audio Restoration*, Springer-Verlag, 1998.

[7] S. J. Godsill and P. J. W. Rayner. Robust reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler. IEEE Tr. Sp.Aud. Proc., July 1998, vol. 6, pp. 352-372

[8] S. J. Godsill and P. J. W. Rayner. Robust noise reduction for speech and audio signals. Proc. ICASSP, 1996.

[9] P. Müller, "Simulation-Based Optimal Design", in *Bayesian Statistics 6*, Oxford University Press, 1998 (to appear)

[10] J.J.K. Ó Ruanaidh and W.J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*, Springer-Verlag, 1996.

[11] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, vol. 77, no. 2, pp. 257-285, 1989.

[12] C.P. Robert, *The Bayesian Choice,* Springer-Verlag Series in Statistics, 1996.

[13] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag Series in Statistics, 1998.

[14] C.P. Robert and D.M. Titterington, "Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum-likelihood estimation", *Stat. Comp.*, vol. 8, pp. 145-158, 1998.

[15] L. Tierney, "Markov Chains for Exploring Posterior Distributions", *Ann. Stat.*, vol. 22, pp. 1701-1762, 1994.

[16] G.C.G. Wei and M.A. Tanner, "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithm", *J. Am. Stat. Assoc.*, vol. 85, pp. 699-704, 1990.