Marginal Maximum A Posteriori Estimation using Markov Chain Monte Carlo

†Arnaud Doucet (corresponding author) - †Simon J. Godsill - ‡Christian P. Robert

†Signal Processing Group, University of Cambridge

Trumpington Street CB2 1PZ Cambridge, UK

‡Laboratoire de Statistique, CREST, INSEE

92245 Malakoff cedex, France

Email: ad2@eng.cam.ac.uk, sjg@eng.cam.ac.uk, robert@ensae.fr

ABSTRACT

Markov chain Monte Carlo (MCMC) methods, while facilitating the solution of many complex problems in Bayesian inference, are not currently well adapted to the problem of marginal *maximum a posteriori* (MMAP) estimation, especially when the number of parameters is large. We present here a simple and novel MCMC strategy, called *State-Augmentation for Marginal Estimation* (SAME), which leads to MMAP estimates for Bayesian models. We illustrate the simplicity and utility of the approach for missing data interpolation in autoregressive time series and blind deconvolution of impulsive processes.

Keywords: Bayesian computation, data augmentation, deconvolution, missing data, simulated annealing.

I. Introduction

When performing Bayesian inference, we are often faced with models that involve high-dimensional unknown parameters. When the marginal MAP (MMAP) estimate is required for inference, that is, when some parameters are nuisance parameters, they must be integrated out. To define notation and terminology, consider the following Bayesian model: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2$ is a random parameter with prior distribution $p(\boldsymbol{\theta})$ associated with the likelihood $p(\mathbf{y}|\,\boldsymbol{\theta})$. The MMAP estimate $\boldsymbol{\theta}_1^{MMAP}$ of the parameter of interest $\boldsymbol{\theta}_1$ is defined as:

$$\boldsymbol{\theta}_1^{MMAP} = \arg\max_{\boldsymbol{\Theta}_1} p(\boldsymbol{\theta}_1|\,\mathbf{y}) \tag{1}$$

where

$$p(\boldsymbol{\theta}_1|\,\mathbf{y}) = \int_{\boldsymbol{\Theta}_2} p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\,\mathbf{y})\, d\boldsymbol{\theta}_2 \tag{2}$$

In cases where a zero-one loss function is applied, the MMAP estimator is optimal. Approximating $\boldsymbol{\theta}_1^{MMAP}$ is a complex problem, however, since, in general, neither the maximization (1) nor the integration (2) can be performed analytically. While the posterior mean estimate is more popular in the statistical literature, it does not always make good sense. For instance, the posterior mean of the parameters of a standard mixture are all equal unless some additional constraints are imposed on them. In a more general setup, the marginal posterior distribution might be multimodal and the MMSE estimate is located between the modes, possibly in a region of very low probability.

When (2) can be performed in closed-form, a classical method for obtaining the MMAP estimate is the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). The EM algorithm is a deterministic algorithm that converges towards a stationary point of the marginal posterior distribution and depends on initialization. It is also limited to certain classes of models for which the expectation and maximization steps can be performed conveniently; stochastic variants of EM such as Stochastic EM (SEM) (Celeux & Diebolt, 1985) and Monte Carlo EM (MCEM) (Wei & Tanner, 1990) have been developed to partially circumvent these limitations[1]. It is worth noting that in these EM-based estimation schemes, the parameter of interest is always updated deterministically in the M step while, in the *State Augmentation for Marginal Estimation* (SAME) algorithm described below, this parameter is updated stochastically. This extra degree of randomness is a key ingredient for preventing convergence to a local mode.

Within a "standard" Monte Carlo framework, marginal inference can be performed by drawing random samples from the joint posterior distribution $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})$ and simply discarding the nuisance parameters: marginalisation is performed implicitly. In principle any random sampling algorithm can be adopted here (Ripley, 1987) but, when the distributions are complex, the most likely sampling method will be Markov chain Monte Carlo (MCMC) (Gilks *et al.*, 1996, Robert & Casella, 1999). In many cases, it is possible to design an efficient MCMC algorithm to sample from the joint distribution $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})$. However, while MCMC methods allow us to perform approximate MMAP estimation by histogram or density estimation applied to the MCMC sample, these methods are more suited to integration than to optimization problems. As an alternative, simulated annealing (SA) methods might be considered for maximizing $p(\boldsymbol{\theta}_1 | \mathbf{y})$ (Van Laarhoven & Arts, 1987). SA methods are a non-homogeneous variant of MCMC which perform global optimization. However, classical SA methods require evaluation of $p(\boldsymbol{\theta}_1 | \mathbf{y})$ up to a normalizing constant and do not introduce $\boldsymbol{\theta}_2$, whereas the introduction of $\boldsymbol{\theta}_2$ is useful for the design of efficient Bayesian computational algorithms.

In this article we propose a new Monte Carlo method for performing MMAP estimation in general Bayesian models. The method is related to SA in that we also simulate from a distribution proportional to the marginal posterior raised to a power γ, but the means of

[1]EM and stochastic variants are adapted to the Bayesian setting by inclusion of a prior penalization term in the M step.

achieving this are quite different: we employ an augmented probability model constructed in such a way that the marginal distribution of $\boldsymbol{\theta}_1$ is proportional to $p^{\gamma}(\boldsymbol{\theta}_1|\mathbf{y})$. The algorithm is thus conceptually very simple and straightforward to implement in most cases, requiring only small modifications to MCMC code written for sampling from $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$.

The paper is organized as follows. In Section 2 the new MCMC strategy for performing MMAP estimation is described, which we call SAME. In Section 3, the method is applied to MMAP estimation of missing data in autoregressive time series and blind deconvolution of impulsive sequences. These examples demonstrate both the importance of performing MMAP estimation in certain problems and the effectiveness of the SAME method.

II. MCMC strategies for MMAP estimation

Before presenting the proposed scheme, we consider how standard MCMC approaches might be adapted for MMAP estimation.

A. Standard MCMC approaches

Assume that we have generated via MCMC a set of (approximate, dependent) samples $\left\{\left(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}\right); i = 1, \ldots, N\right\}$ from the joint posterior distribution $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$. Then, if $p(\boldsymbol{\theta}_1|\mathbf{y})$ is available up to a normalizing constant, it is possible to propose the following estimate of $\boldsymbol{\theta}_1^{MMAP}$

$$\widehat{\boldsymbol{\theta}}_1^{MMAP} = \underset{\boldsymbol{\theta}_1^{(i)}; i=1,\ldots,N}{\arg\max} \; p(\boldsymbol{\theta}_1|\mathbf{y})$$

This method is not efficient in the sense that random samples approximately distributed from $p(\boldsymbol{\theta}_1|\mathbf{y})$ 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. When $p(\boldsymbol{\theta}_1|\mathbf{y})$ is not available up to a normalizing constant, more sophisticated approaches might consider kernel density estimation to find the mode from the samples. These methods, however, are unsuitable for high-dimensional parameters.

B. State Augmentation for Marginal Estimation (SAME)

We present here an alternative simulation-based strategy that is formally related to the SA algorithm. SA methods are a non-homogeneous variant of MCMC used to perform global optimization where the invariant distribution at iteration i of the algorithm is the distribution proportional to $p^{\gamma(i)}(\boldsymbol{\theta}_1|\,\mathbf{y})$, $\gamma(i)$ being a positive increasing sequence tending to infinity. The basic idea is that as $\gamma(i)$ goes to infinity then $p^{\gamma(i)}(\boldsymbol{\theta}_1|\,\mathbf{y})$ concentrates itself upon the set of global modes. As in SA, our iterative algorithm replaces the target distribution $p(\boldsymbol{\theta}_1|\,\mathbf{y})$ by the distribution $\overline{p}_{\gamma(i)}(\boldsymbol{\theta}_1|\,\mathbf{y}) \propto p^{\gamma(i)}(\boldsymbol{\theta}_1|\,\mathbf{y})$ at iteration i. In the SA literature, it has been shown under various assumptions that convergence to the set of global maxima is ensured for a sequence $\gamma(i)$ growing logarithmically (Van Laarhoven & Arts, 1987). However, in practice, the logarithmic function grows too slowly to be useful; that is the distribution $\overline{p}_{\gamma(i)}(\boldsymbol{\theta}_1|\,\mathbf{y})$ does not concentrate quickly enough upon the global modes. Sequences $\gamma(i)$ with a polynomial or an exponential growth are thus preferred.

For the sake of clarity we first assume that $\gamma = \gamma(i)$ is fixed and does not depend on the iteration number. In the classical SA framework, sampling from $\overline{p}_{\gamma}(\boldsymbol{\theta}_1|\,\mathbf{y})$ is realized by using a Metropolis-Hastings or Gibbs sampler. However, such an algorithm cannot be developed when one is not able to evaluate $p(\boldsymbol{\theta}_1|\,\mathbf{y})$ straightforwardly (up to a normalizing constant), and may be hard to construct effectively even when the marginal is available. A novel approach based on a different idea is proposed here.

We define an artificially augmented probability model whose marginal distribution is $\overline{p}_{\gamma}(\boldsymbol{\theta}_1|\,\mathbf{y})$ where γ is a positive integer. If samples $\boldsymbol{\theta}_1^{(i)}$ can be drawn from this concentrated distribution, then as γ becomes large the samples will be concentrated around the global modes of $p(\boldsymbol{\theta}_1|\,\mathbf{y})$. This can be achieved by means of artificial replications of the nuisance parameters in the model. We thus augment the model by replacing $\boldsymbol{\theta}_2$ with γ artificial replications, denoted by $\boldsymbol{\theta}_2(1), \ldots, \boldsymbol{\theta}_2(\gamma)$. Each of these replications is now treated as a distinct random variable in its own right and the following joint distribution is defined

$$q_\gamma\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(1), \ldots, \boldsymbol{\theta}_2(\gamma) \middle| \mathbf{y}\right) \propto \prod_{k=1}^{\gamma} p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(k) \middle| \mathbf{y}\right) \tag{3}$$

The marginal distribution for $\boldsymbol{\theta}_1$ in (3) is obtained by integration over all the replications of $\boldsymbol{\theta}_2$

$$\begin{aligned} q_\gamma\left(\boldsymbol{\theta}_1 \middle| \mathbf{y}\right) &= \int \ldots \int q_\gamma\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(1), \ldots, \boldsymbol{\theta}_2(\gamma) \middle| \mathbf{y}\right) d\boldsymbol{\theta}_2(1) \ldots d\boldsymbol{\theta}_2(\gamma) \\ &\propto \int \ldots \int \prod_{k=1}^{\gamma} p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(k) \middle| \mathbf{y}\right) d\boldsymbol{\theta}_2(1) \ldots d\boldsymbol{\theta}_2(\gamma) \\ &= \overline{p}_\gamma\left(\boldsymbol{\theta}_1 \middle| \mathbf{y}\right) \end{aligned}$$

So, if we build a MCMC algorithm in the augmented space, with invariant distribution $q_\gamma\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(1), \ldots, \boldsymbol{\theta}_2(\gamma) \middle| \mathbf{y}\right)$, then the simulated sequence $\left\{\boldsymbol{\theta}_1^{(i)}; i \in \mathbb{N}\right\}$ will be drawn from the marginal posterior of interest, $\overline{p}_\gamma\left(\boldsymbol{\theta}_1 \middle| \mathbf{y}\right)$: this is the general SAME strategy.

An important point here is that when a MCMC sampler is available for the distribution $p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \middle| \mathbf{y}\right)$ then it is usually easy to construct a MCMC sampler to sample from $q_\gamma\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(1), \ldots, \boldsymbol{\theta}_2(\gamma) \middle| \mathbf{y}\right)$, as the replications of the missing data set are statistically independent conditional upon $\boldsymbol{\theta}_1$, i.e.

$$q_\gamma\left(\boldsymbol{\theta}_2(1), \ldots, \boldsymbol{\theta}_2(\gamma) \middle| \mathbf{y}, \boldsymbol{\theta}_1\right) = \prod_{k=1}^{\gamma} p\left(\boldsymbol{\theta}_2(k) \middle| \mathbf{y}, \boldsymbol{\theta}_1\right) \tag{4}$$

and for $\boldsymbol{\theta}_1$ the full conditional distribution satisfies

$$q_\gamma\left(\boldsymbol{\theta}_1 \middle| \mathbf{y}, \boldsymbol{\theta}_2(1), \ldots, \boldsymbol{\theta}_2(\gamma)\right) \propto \prod_{k=1}^{\gamma} p\left(\boldsymbol{\theta}_1 \middle| \mathbf{y}, \boldsymbol{\theta}_2(k)\right) \tag{5}$$

According to (4), the sampling step for $\boldsymbol{\theta}_2(k)$ is identical to its counterpart in a standard data augmentation sampler with target distribution $p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(k) \middle| \mathbf{y}\right)$ while the sampling step for $\boldsymbol{\theta}_1$ involves a draw from $q_\gamma\left(\boldsymbol{\theta}_1 \middle| \mathbf{y}, \boldsymbol{\theta}_2^{(i)}(1), \ldots, \boldsymbol{\theta}_2^{(i)}(\gamma)\right)$. If $p\left(\boldsymbol{\theta}_1 \middle| \mathbf{y}, \boldsymbol{\theta}_2\right)$ is a member of the regular exponential family, then sampling from $q_\gamma\left(\boldsymbol{\theta}_1 \middle| \mathbf{y}, \boldsymbol{\theta}_2^{(i)}(1), \ldots, \boldsymbol{\theta}_2^{(i)}(\gamma)\right)$ is straightforward as the product of conditionals in (5) is also a member of this exponential family. In more general settings, (5) can be simulated via a slice sampler, a random walk or a Langevin diffusion Metropolis–Hastings sampler, since this density is available in closed-form (Robert & Casella, 1999).

We have assumed a fixed $\gamma(i) = \gamma$ so far. In practice, if we run our algorithm with a high value of γ, then the simulated sequence $\boldsymbol{\theta}_1^{(i)}$ will most likely get stuck in a local maximum located close to $\boldsymbol{\theta}_1^{(0)}$ as the marginal distribution $\overline{p}_\gamma(\boldsymbol{\theta}_1|\mathbf{y})$ is more concentrated around its (local and global) maxima than $p(\boldsymbol{\theta}_1|\mathbf{y})$. It is thus beneficial to use an increasing sequence $\gamma(i)$ in a fashion similar to standard SA such that $\lim_{i\rightarrow+\infty} \gamma(i) = +\infty$. Contrary to standard SA, however, $\gamma(i)$ has to be a sequence of strictly positive integers in the SAME algorithm. Typically, we start at iteration $i = 1$ with the most diffuse possible (marginal) invariant distribution $p(\boldsymbol{\theta}_1|\mathbf{y})$, that is we set $\gamma(1) = 1$, in order to facilitate the exploration of the state-space. At this point, we have a MCMC transition kernel with (marginal) invariant distribution $p(\boldsymbol{\theta}_1|\mathbf{y})$; we then progressively increase $\gamma(i)$ so that, at iteration i, we have $\overline{p}_{\gamma(i)}(\boldsymbol{\theta}_1|\mathbf{y})$ as (marginal) invariant distribution. As mentioned before, proper sequences for regular SA schemes have a logarithmic growth rate but sequences with faster growth rate have to be used in practice (Van Laarhoven & Arts, 1987).

Example. If a data augmentation algorithm can be applied to sample from $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$ by sampling iteratively and successively from $p(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1)$, then a non-homogeneous SAME version of this algorithm to maximize $p(\boldsymbol{\theta}_1|\mathbf{y})$ proceeds as follows:

Non-homogeneous SAME method

1. $i = 0$. Initialize $\boldsymbol{\theta}_1^{(0)}$.

2. Iteration i, $i \geq 1$

- For $k = 1, \ldots, \gamma(i)$, sample $\boldsymbol{\theta}_2^{(i)}(k) \sim p\left(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1^{(i-1)}\right)$.

- Sample $\boldsymbol{\theta}_1^{(i)} \sim q_{\gamma(i)}\left(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2^{(i)}(1), \ldots, \boldsymbol{\theta}_2^{(i)}(\gamma)\right)$.

This algorithm is run until the sequence of the $\boldsymbol{\theta}_1^{(i)}$'s stabilises, that is, for N iterations, producing an approximation of $\boldsymbol{\theta}_1^{MMAP}$ as $\widehat{\boldsymbol{\theta}}_1^{MMAP} = \boldsymbol{\theta}_1^{(N)}$.

III. Applications

A sequence $\gamma(i)$ going to infinity is theoretically required for $\overline{p}_{\gamma(i)}(\boldsymbol{\theta}_1|\mathbf{y})$ to concentrate itself on the set of global maxima. In practice, as for classical SA algorithms, this cannot be implemented. In all the applications we addressed, a maximum value of $\gamma(i)$ around 100 is apparently sufficient to observe convergence of the SAME procedure. Thus, in all the examples addressed here, we implemented, as is usually done for classical SA algorithms in practice (Van Laarhoven & Arts, 1987), N iterations of the SAME algorithm with a sequence growing linearly , *i.e.* $\gamma(i) = \lfloor ai + b \rfloor$, satisfying $\gamma(0) = 1$ and $\gamma(N) = 100$. Other polynomial and exponential growing sequences have been implemented but the SAME procedure appeared fairly robust to the selected sequence. A rigorous convergence assessment procedure for a non-homogeneous Markov chain such as the SAME algorithm is unfortunately not available currently, given the difficulties that beset convergence assessment even of homogeneous chains (Robert, 1998). Practical guidelines include graphical monitoring of the stability of the simulated MMAP estimates as well as multiple pilot runs from overdispersed random starting points.

A. Missing data estimation in autoregressive time series

1. Signal Model and Estimation Objectives

We first consider a problem which applies in the replacement of missing data packets in speech signals and the restoration of audio time series (Godsill & Rayner, 1998). The data sequence x_t is assumed to be drawn from an autoregressive (AR) process with coefficients $\mathbf{a} = [a_0, \ldots, a_{L-1}]^{\mathrm{T}}$ and the state vector at time sampling instant t is denoted by $\mathbf{x}_t = [x_t, \ldots, x_{t-L+1}]^{\mathrm{T}}$. The prior distribution for the initial state $\mathbf{x}_L$ is diffuse. 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 $\mathcal{I} = \{i_1, ..., i_l\} \subset \{1, ..., T\}$, but fully observed elsewhere in the interval $\{1, ..., T\}$. 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 nuisance parameters $\left(\mathbf{a}, \sigma^2\right)$ are unknown. We assign the conjugate normal-inverted gamma prior distribution to $\mathbf{a}$ and σ^2:

$$\mathbf{a}|\,\sigma^2 \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\Sigma_0\right) \text{ and } \sigma^2 \sim \mathcal{IG}\left(\frac{\eta_0}{2}, \frac{\nu_0}{2}\right)$$

with Σ_0 a regular matrix. Given the set of observations $\mathbf{x}_{-\mathcal{I}}$, our aim is to estimate $\boldsymbol{\theta}_1 = \mathbf{x}_{\mathcal{I}}$ in a MMAP sense, i.e. obtaining $\boldsymbol{\theta}_1^{MMAP} = \arg\max p\left(\boldsymbol{\theta}_1|\,\mathbf{x}_{-\mathcal{I}}\right)$. This is a model which allows an exact EM implementation for direct comparison with SAME.

2. MMAP parameter estimation

To maximize $p\left(\boldsymbol{\theta}_1|\,\mathbf{x}_{-\mathcal{I}}\right)$, we introduce the parameters $\boldsymbol{\theta}_2 = \left(\mathbf{a}, \sigma^2\right)$ as nuisance parameters and then use the SAME strategy. To implement this algorithm, we must sample from $p\left(\mathbf{a}, \sigma^2\middle|\,\mathbf{x}_{-\mathcal{I}}, \mathbf{x}_{\mathcal{I}}\right)$ and $q_{\gamma(i)}\left(\boldsymbol{\theta}_1|\,\mathbf{x}_{-\mathcal{I}}, \boldsymbol{\theta}_2\left(1\right), \ldots, \boldsymbol{\theta}_2\left(\gamma\left(i\right)\right)\right)$. One obtains by conjugacy calculations (Bernardo & Smith, 1994, Appendix A.2; Godsill & Rayner, 1998, Sections 12.4 and 12.5)

$$\mathbf{a}|\,\sigma^2 \sim \mathcal{N}\left(\mathbf{m}_{\mathbf{a}}, \sigma^2\Sigma_{\mathbf{a}}\right) \text{ and } \sigma^2 \sim \mathcal{IG}\left(\left(\eta_0 + T - L\right)/2, \left(\nu_0 + \mathbf{x}_{1:T}^{\mathrm{T}}\mathbf{x}_{1:T} - \mathbf{m}_{\mathbf{a}}^{\mathrm{T}}\Sigma_{\mathbf{a}}^{-1}\mathbf{m}_{\mathbf{a}}\right)/2\right)$$

where $\mathbf{x}_{1:T} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_T\right)^{\mathrm{T}}$ and

$$\Sigma_{\mathbf{a}}^{-1} = \Sigma_0^{-1} + \sum_{t=L+1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}^{\mathrm{T}} \text{ and } \mathbf{m}_{\mathbf{a}} = \Sigma_{\mathbf{a}} \sum_{t=L+1}^{T} \mathbf{x}_{t-1}x_t$$

The required simulations from $q_{\gamma(i)}\left(\mathbf{x}_{\mathcal{I}}|\,\mathbf{x}_{-\mathcal{I}}, \boldsymbol{\theta}_2\left(1\right), \ldots, \boldsymbol{\theta}_2\left(\gamma\left(i\right)\right)\right)$ are readily obtained from the basic model as follows

$$\mathbf{x}_{\mathcal{I}}|\mathbf{x}_{-\mathcal{I}}, \boldsymbol{\theta}_2(1), \ldots, \boldsymbol{\theta}_2\left(\gamma\left(i\right)\right) \sim \mathcal{N}\left(\mathbf{m}_{\mathbf{x}}(i), \Sigma_{\mathbf{x}}(i)\right)$$

where

$$\Sigma_{\mathbf{x}}^{-1}(i) = \sum_{k=1}^{\gamma(i)} \frac{\mathbf{A}(k)_{\mathcal{I}}^{\mathsf{T}}\mathbf{A}(k)_{\mathcal{I}}}{\sigma(k)^2} \quad \text{and} \quad \mathbf{m}_{\mathbf{x}}(i) = -\Sigma_{\mathbf{x}}(i) \sum_{k=1}^{\gamma(i)} \frac{\mathbf{A}(k)_{\mathcal{I}}^{\mathsf{T}}\mathbf{A}(k)_{-\mathcal{I}}}{\sigma(k)^2}\mathbf{x}_{-\mathcal{I}}$$

and the indices within parentheses '(k)' refer to the k^{th} augmented parameter. Here $\mathbf{A}$ is the matrix formed from the AR coefficients such that $\mathbf{e}_{L+1:T} = (e_1, \ldots, e_T)^{\mathsf{T}} = \mathbf{A}\,\mathbf{x}_{1:T}$, and $[\mathbf{A}_{\mathcal{I}}, \mathbf{A}_{-\mathcal{I}}]$ forms a columnwise partition of $\mathbf{A}$ with columns selected according to $\mathcal{I}$.

3. Simulations

In this example we compare the performance of three possible methods for estimating the MMAP solution to a high dimensional problem: EM, standard MCMC and the SAME algorithm. We have chosen a model in which it is possible to perform EM exactly and also to evaluate exactly the posterior probability for the desired parameters. In this way it is possible to make an objective comparison between the three methods in terms of the highest marginal posterior probability value achieved by each. Note that the dataset has been carefully chosen in order to highlight the differences between the SAME algorithm and its competitors for MMAP estimation. We do not here claim that MMAP estimation is always the correct procedure for data of this type, since it could potentially lead to problems of overfitting and neglecting solutions which represent the overall probability mass better. However, it is worth noting that the posterior distribution turns out to be strongly multi-modal in this case, and so extreme caution would have to be used in choosing other estimators such as the posterior mean.

For this model we tested the new method in a challenging situation where 50% of the data are missing in the middle of a short block of length $T = 40$, extracted from a digitised audio signal (Godsill & Rayner, 1998). The AR model order is fixed at $L = 9$. The following prior parameters have been adopted: $\Sigma_0 = 100\,\mathbf{I}_L$, $\eta_0 = \nu_0 = 0.01$. $N = 200$ iterations of the SAME algorithm were run.

We compared, via numerical simulations, the SAME algorithm with EM and the homogeneous MCMC sampler, i.e. $\gamma(i) = 1$ for all i. To perform a fair comparison with the SAME algorithm in terms of computational complexity, the EM and MCMC algorithms were run for $N_0 = \sum_{i=1}^{N} \gamma(i)$ iterations[2]. In all cases, the algorithms were initialized with the same random parameter estimate $\boldsymbol{\theta}^{(0)}$ and we took as final estimate of $\boldsymbol{\theta}_1^{MMAP}$ the parameter $\boldsymbol{\theta}_1^{(N)}$ for the SAME, $\boldsymbol{\theta}_1^{(N_0)}$ for the EM and $\underset{i=1,\ldots,N_0}{\arg\max}\, p\left(\boldsymbol{\theta}_1^{(i)} \middle| \mathbf{x}_{-\mathcal{I}}\right)$ for the MCMC sampler. Figure 1 shows the posterior density values against iteration number 1 to N for the three algorithms. Table 1 displays the maximum posterior density values achieved by each method.

Figure 1 about here

Table 1 about here

It is clear from this that the three methods give very different results and that the SAME algorithm finds a more probable solution than the other techniques. It is also clear from the probability plots that the Gibbs sampler would be quite inappropriate for performing MMAP estimation since it virtually never achieves probabilities close to the maximum. EM has converged to a stationary point of the posterior distribution with a significantly lower posterior density value than the SAME algorithm.

In 100 simulations carried out from randomly chosen initialisations with the same dataset, two principal modes of the distribution are identified by EM and SAME, see Figure 2. EM is very prone to converge to the lower probability mode, while SAME is much more likely to reach the higher probability solution: the average improvement in log-probability over 100 iterations from using SAME was 4.33; SAME reached the same or a higher probability mode than EM in 93 out of 100 trials.

Figure 2 about here

[2] In fact this comparison is favourable for MCMC, since the SAME method requires many fewer draws from $\boldsymbol{\theta}_1$.

B. Blind deconvolution of impulsive processes

1. Signal Model and Estimation Objectives

The observed signal y_t is modelled as the convolution of a sequence v_t with a MA model $\widetilde{\mathbf{h}} = [1, h_1, \ldots, h_L]^{\mathrm{T}} = \left[1, \mathbf{h}^{\mathrm{T}}\right]^{\mathrm{T}}$ observed in white Gaussian noise. If we denote $\widetilde{\mathbf{v}}_t = [v_t, v_{t-1}, \ldots, v_{t-L}]^{\mathrm{T}} = \left[v_t, \mathbf{v}_{t-1}^{\mathrm{T}}\right]^{\mathrm{T}}$, then

$$y_t = \widetilde{\mathbf{h}}^{\mathrm{T}} \widetilde{\mathbf{v}}_t + w_t$$

where $w_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_w^2\right)$. In this model v_t is a Bernoulli-Gauss sequence, i.e. v_t is an i.i.d. sequence such that

$$v_t \sim \lambda \mathcal{N}\left(0, \sigma_v^2\right) + (1-\lambda)\,\delta_0, \;\; 0 \leq \lambda \leq 1$$

where δ_0 denotes the delta-Dirac measure at 0. The sequence v_t and the parameters $\boldsymbol{\theta}_1 = \left(\mathbf{h}, \sigma_w^2, \sigma_v^2, \lambda\right)$ are unknown. It is convenient from an algorithmic point of view to introduce the latent Bernoulli process $r_t \in \{0, 1\}$ such that $\Pr\left(r_t = 1\right) = \lambda$ and

$$v_t |\, r_t = 1 \sim \mathcal{N}\left(0, \sigma_v^2\right), \;\; v_t |\, r_t = 0 \sim \delta_0$$

This statistical model finds application in seismic signal processing (Cheng *et al.*, 1996). We assign a prior distribution to the unknown parameters $\boldsymbol{\theta}_1$ such that

$$p\left(\boldsymbol{\theta}_1\right) = p\left(\mathbf{h}, \sigma_w^2, \sigma_v^2, \lambda\right) = p\left(\mathbf{h}|\, \sigma_w^2\right) p\left(\sigma_w^2\right) p\left(\sigma_v^2\right) p\left(\lambda\right)$$

For the MA model and noise variance, a normal-inverse gamma prior distribution is selected, i.e.

$$\mathbf{h}|\, \sigma_w^2 \sim \mathcal{N}\left(\mathbf{0}, \sigma_w^2 \Sigma_0\right) \text{ and } \sigma_w^2 \sim \mathcal{IG}\left(\frac{\eta_w}{2}, \frac{\nu_w}{2}\right)$$

with Σ_0 a regular matrix, and

$$\sigma_v^2 \sim \mathcal{IG}\left(\frac{\eta_v}{2}, \frac{\nu_v}{2}\right) \text{ and } \lambda \sim \mathcal{U}\left[0, 1\right]$$

Given the set of observations $\mathbf{y}_{1:T} \triangleq \{y_1, \ldots, y_T\}$, our aim is to estimate $\boldsymbol{\theta}_1$ in a MMAP sense, i.e. obtaining $\boldsymbol{\theta}_1^{MMAP} = \arg\max p\left(\boldsymbol{\theta}_1 |\, \mathbf{y}_{1:T}\right)$. A simpler version of this problem has been addressed using several stochastic versions of the EM in Cappé *et al.* (1999). A homogeneous MCMC sampler to estimate the posterior distribution for a similar problem was proposed in Cheng *et al.* (1996).

2. MMAP parameter estimation

To maximize $p(\boldsymbol{\theta}_1|\, \mathbf{y}_{1:T})$, we introduce the unobserved sequences $\mathbf{r}_{1:T} \triangleq \{r_1, \ldots, r_T\}$ and $\mathbf{v}_{1:T} \triangleq \{v_1, \ldots, v_T\}$ as nuisance parameters, i.e. $\boldsymbol{\theta}_2 = (\mathbf{r}_{1:T}, \mathbf{v}_{1:T})$ and then use the SAME strategy. To implement this algorithm, we choose a strategy in which we sample from the reduced conditional $p(r_t|\, \mathbf{y}_{1:T}, \boldsymbol{\theta}_1, \mathbf{r}_{-t})$ where $\mathbf{r}_{-t} \triangleq (r_1, \ldots, r_{t-1}, r_{t+1}, \ldots r_T)^{\mathsf{T}}$ for any $t = 1, \ldots, T$, $p(\mathbf{v}_{1:T}|\, \mathbf{y}_{1:T}, \boldsymbol{\theta}_1, \mathbf{r}_{1:T})$ and

$$q_{\gamma(i)}(\boldsymbol{\theta}_1|\, \mathbf{y}_{1:T}, \mathbf{r}_{1:T}(1), \mathbf{v}_{1:T}(1), \ldots, \mathbf{r}_{1:T}(\gamma(i)), \mathbf{v}_{1:T}(\gamma(i))) . \tag{6}$$

Sampling from $p(r_t|\, \mathbf{y}_{1:T}, \boldsymbol{\theta}_1, \mathbf{r}_{-t})$ is realized using the algorithm described in Cappé *et al.* (1999) whose computational complexity is $O(T)$. Sampling from $p(\mathbf{v}_{1:T}|\, \mathbf{y}_{1:T}, \boldsymbol{\theta}_1, \mathbf{r}_{1:T})$ is implemented using the simulation smoother of DeJong and Shephard (1995). To sample from (6), one obtains by standard conjugacy calculations (Bernardo & Smith, 1994, Appendix A.2)(Cheng *et al.*, 1996)

$$\begin{aligned}
\mathbf{h}|\, \sigma_w^2 &\sim \mathcal{N}\left(\mathbf{m}(i), \sigma_w^2 \Sigma(i)\right) \\
\sigma_w^2 &\sim \mathcal{IG}\left(\gamma(i)(\eta_w + T)/2 + (\gamma(i) - 1)(L/2 + 1),\right. \\
&\qquad \left.\left(\gamma(i)\nu_w + \varepsilon(i) - \mathbf{m}(i)^{\mathsf{T}}(i)\,\Sigma^{-1}(i)\,\mathbf{m}(i)\right)/2\right\}
\end{aligned}$$

where

$$\begin{aligned}
\Sigma^{-1}(i) &= \gamma(i)\,\Sigma_0^{-1} + \sum_{k=1}^{\gamma(i)} \sum_{t=1}^{T} \mathbf{v}_{t-1}(k)\,\mathbf{v}_{t-1}^{\mathsf{T}}(k) \\
\mathbf{m}(i) &= \Sigma(i) \sum_{k=1}^{\gamma(i)} \sum_{t=1}^{T} \mathbf{v}_{t-1}(k)\,(y_t - v_t(k)) \\
\varepsilon(i) &= \sum_{k=1}^{\gamma(i)} (\mathbf{y}_{1:T} - \mathbf{v}_{1:T}(k))^{\mathsf{T}} (\mathbf{y}_{1:T} - \mathbf{v}_{1:T}(k))
\end{aligned}$$

and

$$
\begin{aligned}
\lambda &\sim \mathcal{B}\left(\gamma(i)+\sum_{k=1}^{\gamma(i)}\sum_{t=1}^{T} r_t(k), \gamma(i)(1+T)-\sum_{k=1}^{\gamma(i)}\sum_{t=1}^{T} r_t(k)\right) \\
\sigma_v^2 &\sim \mathcal{IG}\left\{\left(\gamma(i)(\eta_v+2)+\sum_{j=1}^{\gamma(i)}\sum_{t=1}^{T} r_t(j)\right)/2-1,\right. \\
&\qquad\left.\left(\gamma(i)\nu_v+\sum_{k=1}^{\gamma(i)} \mathbf{v}_{1:T}(k)\,\mathbf{v}_{1:T}^{\mathrm{T}}(k)\right)/2\right\}
\end{aligned}
$$

3. Simulations

The following signal has been simulated using the parameters: $T = 500$, $L = 3$, and $\boldsymbol{\theta}_1$ as given in Table 2. The observations are displayed in Figure 3.

Figure 3 about here

The following prior parameters have been adopted: $\Sigma_0 = 100\,\mathbf{I}_L$, $\eta_v = \eta_w = \nu_w = \nu_w = 0.01$. In this case, the EM algorithm cannot be applied as the E-step does not admit a closed-form expression. The SAME algorithm was run for $N = 200$ iterations. The algorithm was initialized with the following parameters $\mathbf{h}^{(0)} = [0, 0, 0]^{\mathrm{T}}$, $\lambda = 0.5$ and $\sigma_w^2 = \sigma_v^2 = 1$. Figures 3 and 4 show the simulated parameters against iteration number. We observe convergence of the algorithm after a short transient regime. Table 2 gives the estimated parameters $\widehat{\boldsymbol{\theta}}_1^{MMAP}$ for $T = 500$. The results for $T = 1000$ are also presented.

Figure 4 about here

Figure 5 about here

Table 2 about here

The algorithm estimates very well the observation noise σ_w^2 and the occurrence rate λ. Given the level of observation noise and the low occurrence rate of the Bernoulli process, it appears very difficult to estimate σ_v^2 and $\mathbf{h}$ very accurately for $T = 500$, better results are naturally obtained for $T = 1000$.

We performed tests from 100 randomly chosen starting points with the same dataset. The SAME method attained an average log-posterior probability of -345.9 as compared with -343.2 for MCEM (Cappé *et al.*, 1999)(Wei & Tanner, 1990). The standard deviation of these values is 0.1 for SAME and 1.2 for MCEM. Moreover SAME reached the same or a higher probability mode than MCEM in all cases.

IV. Conclusion

In this article, we have presented an original simulation-based strategy to maximize marginal posterior distributions. This method is closely linked to SA. However, contrary to classical SA algorithms, it is based on the introduction of an artificial augmented probability model and allows us to handle models that cannot be addressed by standard SA methods. Once a MCMC algorithm is available to sample from a posterior distribution, the proposed algorithms are very simple to implement in all cases we have considered. Computer simulations demonstrate the effectiveness of our method compared with EM in a multimodal interpolation problem and with MCEM in a blind deconvolution problem.

References

[1] Bernardo J.M. and Smith A.F.M. (1994) *Bayesian Theory*, John Wiley & Sons.

[2] Cappé O., Doucet A., Moulines E. and Lavielle M. (1999) Simulation-based Methods for Blind Maximum-Likelihood Filter Identification. *Signal Processing*, 73, 3-25.

[3] Carter C.K. and Kohn R. (1994) On Gibbs Sampling for State Space Models. *Biometrika*, 81, 541-553.

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

[5] Celeux G. and Diebolt J. (1990) Une Version Recuit Simulé de l'Algorithme EM. *Comptes Rendus de l'Académie des Sciences,* 310, 119-124 (in French).

[6] Cheng Q., Chen, R. and Li, T. (1996) Simultaneous Wavelet Estimation and Deconvolution of Reflection Seismic Signals via Gibbs sampler. *IEEE Transactions on Geoscience and Remote Sensing*, 34, 377-384.

[7] DeJong P. and Shephard N. (1995) The Simulation Smoother for Time Series Models. *Biometrika*, 82, 339-350.

[8] Dempster A.P., Laird N.M. and Rubin D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, 39, 1-38.

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

[10] Godsill S.J. and Rayner P.J.W. (1998) *Digital Audio Restoration - A Statistical Model-Based Approach*, Springer-Verlag.

[11] Ripley B.D. (1987) *Stochastic Simulation*, Wiley, New York.

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

[13] Robert C.P. (1998) *Discretization and MCMC Convergence Assessment*, Lecture Notes in Statistics no. 135, Springer-Verlag, New York.

[14] Robert C.P. and Casella G. (1999) *Monte Carlo Statistical Methods*, Springer-Verlag Series in Statistics.

[15] Van Laarhoven P.J. and Arts E.H.L. (1987) *Simulated Annealing: Theory and Applications*, Reidel Pub., Amsterdam.

[16] Wei G.C.G. and Tanner M.A. (1990) 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.

V. Tables

Algorithm	EM	MCMC	SAME
Maximum posterior density value	-210.34	-203.85	-198.93

Table 1: Performances of the EM, MCMC and SAME algorithms for missing data interpolation

Parameter	True value	MMAP Estimates $T = 500$	MMAP Estimates $T = 1000$
h_1	-1.50	-1.87	-1.42
h_2	0.50	0.68	0.50
h_3	-0.20	-0.33	-0.20
λ	0.15	0.14	0.17
σ_w^2	0.10	0.10	0.10
σ_v^2	0.50	0.34	0.41

Table 2: True values and estimated MMAP values for blind deconvolution of impulsive processes

VI. Figures

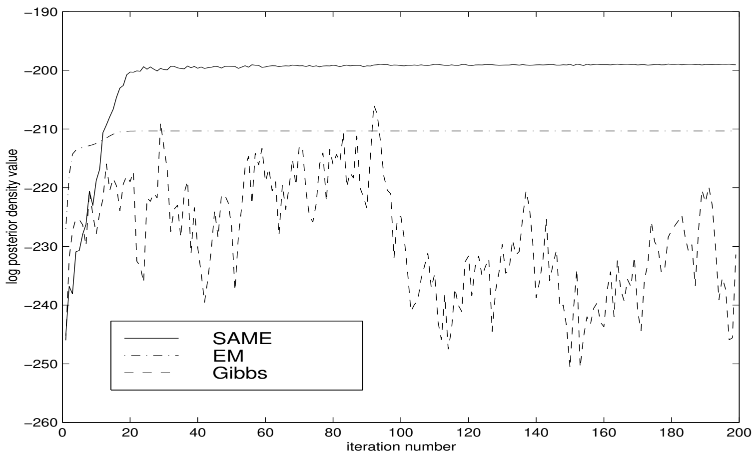

Figure 1: Log-posterior density values $\log(p(\mathbf{x}_{\mathcal{I}}|\mathbf{x}_{-\mathcal{I}}))$ against iteration number for the three algorithms

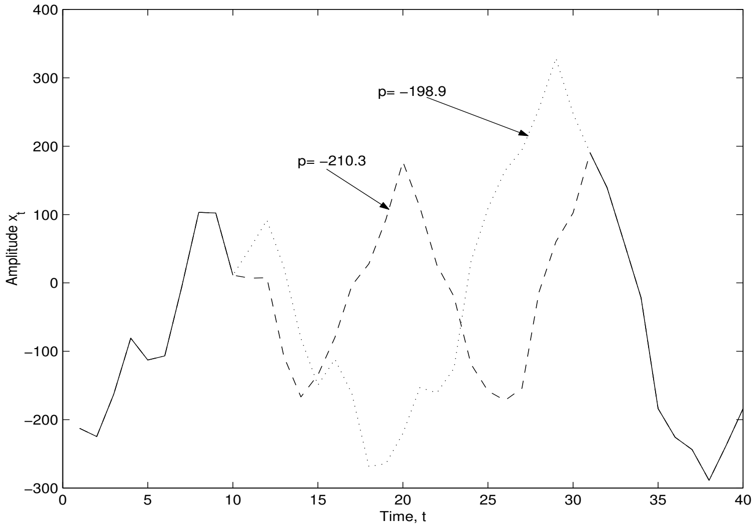

Figure 2: Principal modes estimated by EM and SAME, shown dotted/dashed. Marginal probability density values $\log(p(\mathbf{x}_{\mathcal{I}}|\mathbf{x}_{-\mathcal{I}}))$ are indicated by '$p =$'. Observed data $\mathbf{x}_{-\mathcal{I}}$ are shown as solid lines.

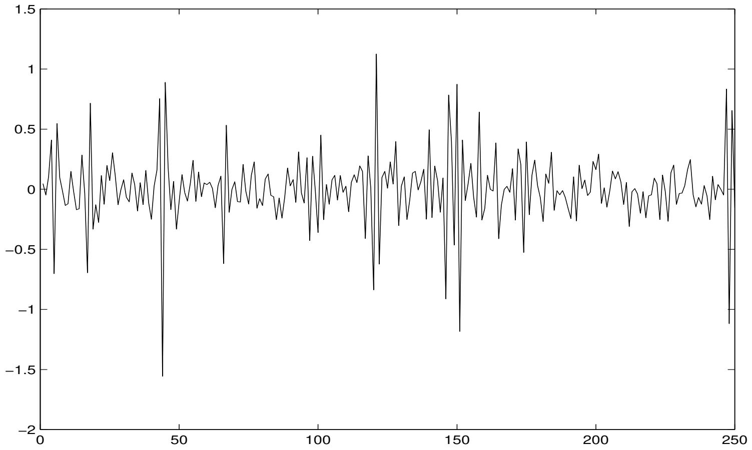

Figure 3: Observations y_t.

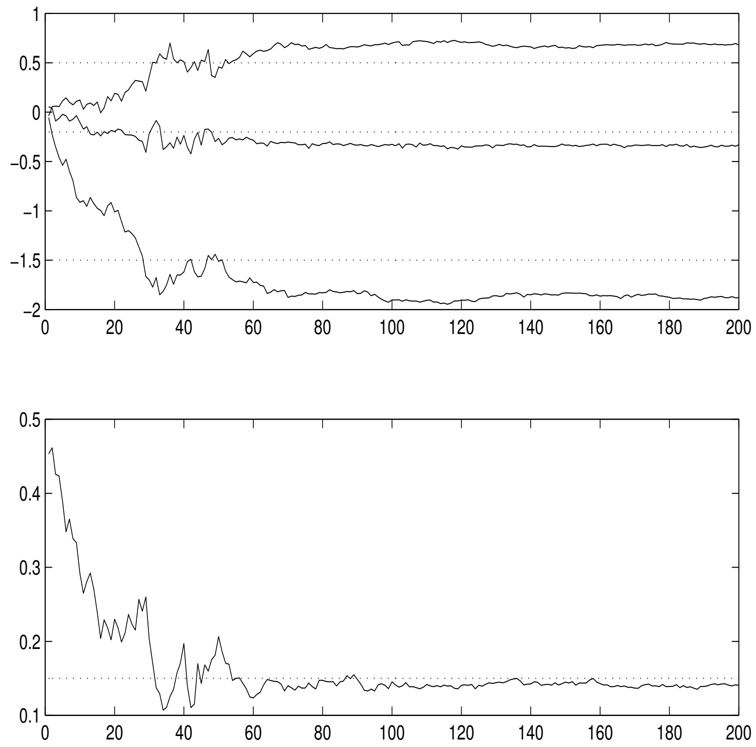

Figure 4: Parameters values against iteration number. True values are displayed in dotted line. Top: filter $\mathbf{h}$. Bottom: occurrence rate λ.

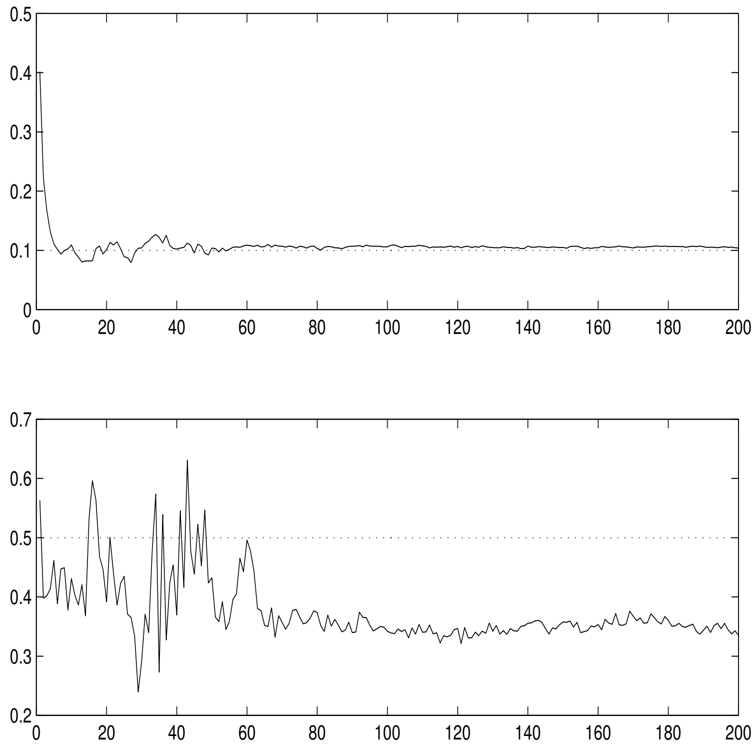

Figure 5: Parameters values against iteration number. True values are displayed in dotted line. Top: standard deviation σ_w. Bottom: standard deviation σ_v.