MLE for partially observed diffusions: direct maximization vs. The EM algorithm

Fabien Campillo, François Le Gland

INRIA

UNITÉ DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 884

MLE FOR PARTIALLY OBSERVED DIFFUSIONS : DIRECT MAXIMIZATION vs. THE EM ALGORITHM

Programme 5

Fabien CAMPILLO
François LE GLAND

AOUT 1988

RR-0884

MLE FOR PARTIALLY OBSERVED DIFFUSIONS: DIRECT MAXIMIZATION vs. THE EM ALGORITHM*

Estimateur du Maximum de Vraisemblance
pour des Processus de Diffusion Partiellement Observés:
Maximisation Directe vs. Algorithme EM

Fabien CAMPILLO and François LE GLAND
INRIA
Route des Lucioles
06565 VALBONNE Cedex
FRANCE

*Research partially supported by USACCE under Contract DAJA45-87-M-0296, and by CNRS-GRECO *"Traitement du Signal et Image"*

Abstract

Two algorithms are compared for maximizing the likelihood function associated with parameter estimation in partially observed diffusion processes

- the EM algorithm, investigated by Dembo and Zeitouni [2], an iterative algorithm where, at each iteration, an auxiliary function is computed and maximized,
- the direct approach where the likelihood function itself is computed and maximized.

This yields to a comparison of nonlinear smoothing and nonlinear filtering for the computation of a class of conditional expectations related to the problem of estimation (Section 3). In particular, it is shown that smoothing is indeed necessary for the EM algorithm approach to be efficient.

Time–discretization schemes for the stochastic PDE's involved in the algorithms are given, and the link with the discrete–time case (hidden Markov model) is explored.

Numerical results are presented (Section 6) with the conclusion that direct maximization should be prefered whenever some noise covariances associated with the parameters to be estimated are small.

Keywords: *parameter estimation, maximum likelihood, EM algorithm, diffusion processes, nonlinear filtering, nonlinear smoothing, Skorokhod integral, time–discretization.*

Résumé

On compare deux algorithmes de maximisation de la fonction de vraisemblance associée à un problème d'estimation de paramètres pour des processus de diffusion avec observation partielle

- l'algorithme EM, étudié par Dembo and Zeitouni [2]; il s'agit d'un algorithme itératif où, à chaque itération, une fonction auxiliaire est calculée et maximisée,
- l'approche directe où la fonction de vraisemblance est calculée et maximisée.

On aboutit à la comparaison du filtrage non linéaire et du lissage non linéaire pour le calcul d'une classe d'espérances conditionnelles associées au problème d'estimation (paragraphe 3). En particulier, on montre que l'algorithme EM n'est intéressant qu'à condition d'utiliser le lissage non linéaire.

On introduit les schémas de discrétisation en temps des EDP stochastiques utilisées pour les algorithmes. On présente le lien avec le cas du temps discret (chaînes de Markov cachées).

Des résultats numériques (paragraphe 6) on déduit que l'approche maximisation directe doit être retenue dès que les covariances de certains bruits associés aux paramètres à estimer sont petits.

Mots–clé: *estimation de paramètres, maximum de vraisemblance, algorithme EM, processus de diffusion, filtrage non linéaire, lissage non linéaire, intégral de Skorokhod, discrétisation en temps.*

Contents

1 Introduction: the EM algorithm

The EM algorithm is an iterative algorithm for maximizing a likelihood function, in a context of partial information [3]. Indeed, let $(P_\theta : \theta \in \Theta)$ be a family of mutually absolutely continuous probability measures on a measurable space $(\Omega, \mathcal{F})$, with $P_\theta \sim R$ and let $\mathcal{Y} \subset \mathcal{F}$ be the σ–algebra containing all the available information. Then, the log–likelihood function for the estimation of the parameter θ can be defined as

$$L(\theta) \triangleq \log \mathbf{E}_R \left(\frac{dP_\theta}{dR} \mid \mathcal{Y} \right) , \tag{1}$$

and the MLE (maximum likelihood estimate) as

$$\widehat{\theta} \in \arg\max_{\theta \in \Theta} L(\theta) .$$

The EM algorithm is based on the following straightforward application of Jensen's inequality

$$L(\theta) - L(\theta') = \log \mathbf{E}_{\theta'} \left(\frac{dP_\theta}{dP_{\theta'}} \mid \mathcal{Y} \right) \geq \mathbf{E}_{\theta'} \left(\log \frac{dP_\theta}{dP_{\theta'}} \mid \mathcal{Y} \right) \triangleq Q(\theta, \theta') , \tag{2}$$

which gives, for each value θ' of the parameter, a global minoration of the log–likelihood function $\theta \mapsto L(\theta)$ by means of an auxiliary function $\theta \mapsto L(\theta') + Q(\theta, \theta')$, with equality at $\theta = \theta'$. The algorithm iterations are described by the following steps

1. $p = 0$, initial guess $\widehat{\theta}_0$,
2. set $\theta' = \widehat{\theta}_p$,
3. (E–step) compute $Q(\cdot, \theta')$,
4. (M–step) find $\widehat{\theta}_{p+1}$ such that $Q(\widehat{\theta}_{p+1}, \theta') \geq Q(\theta, \theta')$ for all $\theta \in \Theta$,
5. if a stopping test is satisfied,
   then set final estimate $\theta^* = \widehat{\theta}_{p+1}$,
   else repeat from step 2 with $p = p + 1$.

An interesting feature of the algorithm is that it generates a maximizing sequence $\{\widehat{\theta}_p : p = 0, 1, \cdots\}$ in the sense that $L(\widehat{\theta}_{p+1}) > L(\widehat{\theta}_p)$ unless $\widehat{\theta}_{p+1} = \widehat{\theta}_p$. Some general convergence results about the sequences $\{L(\widehat{\theta}_p) : p = 0, 1, \cdots\}$ and $\{\widehat{\theta}_p : p = 0, 1, \cdots\}$ are proved in [13], under mild regularity assumptions on $L(\cdot)$ and $Q(\cdot, \cdot)$ – see also [2, Theorem 2]. To prove the existence of smooth enough – in the a.s. sense – versions of $\theta \mapsto L(\theta)$ and $(\theta, \theta') \mapsto Q(\theta, \theta')$, as well as to get the expression of the corresponding derivatives, one can rely e.g. on [12, Lemma 1].

To decide whether this algorithm is interesting from a computational point of view, the following three questions should be answered

[E] *how expensive is the computation of the auxiliary function $Q(\cdot, \theta')$?*

[M] *how easy is the maximization of the auxiliary function $Q(\cdot, \theta')$?*

[EM] *how fast is the convergence of this sub–optimal iterative algorithm towards the MLE ?*

In [2], the EM algorithm has been applied in the context of continuous–time partially observed stochastic processes. In the particular case of diffusion processes, the general expression of $Q(\theta, \theta')$ has been derived and said to involve a nonlinear smoothing problem. The purpose of this work is to address the following three points

- discuss the expression in [2] giving $Q(\theta, \theta')$ in terms of a nonlinear smoothing problem – this will involve generalized stochastic calculus (Skorokhod integral).
- get an equivalent expression, in terms of a nonlinear filtering problem, for $Q(\theta, \theta')$ and its gradient $\nabla^{1,0}Q(\theta, \theta')$ – it will turn out that smoothing is indeed necessary for the point **[M]** introduced above to be satisfied, although filtering is enough to compute $Q(\theta, \theta')$ for a given pair (θ, θ').
- get similar expressions for the original log–likelihood function $L(\theta)$ and its gradient $\nabla L(\theta)$.

This will allow to compare, from a computational point of view, the two possible methods for maximum likelihood estimation

- direct maximization of the likelihood function [4],
- the EM algorithm.

In particular, the point **[M]** will receive a positive answer, which is indeed the main motivation for the EM algorithm. On the other hand, it will be proved that computing the auxiliary function $Q(\cdot, \theta')$ is a more heavy task than computing the original log–likelihood function $L(\cdot)$. As for the point **[EM]**, numerical examples will show that the convergence of the EM algorithm may be very slow. This typically occurs in those cases where, for each $\theta' \in \Theta$ the function $L(\theta') + Q(\cdot, \theta')$ is very sharp below the log–likelihood function $L(\cdot)$. In such cases indeed, maximizing the auxiliary function does not allow to update significantly enough the current estimate at each M–step.

The statistical model is presented in Section 2, where expressions are given for $L(\theta)$, $\nabla L(\theta)$, $Q(\theta, \theta')$ and $\nabla^{1,0}Q(\theta, \theta')$ in terms of conditional expectations. It turns out that the last three expressions all belong to a certain class of conditional expectations. Two methods are then proposed in Section 3 for the computation of conditional expectations in this class – one based on nonlinear filtering, the other on nonlinear smoothing and involving generalized stochastic calculus (Skorokhod integral). These results are applied in Section 4 to the computation of $L(\theta)$, $\nabla L(\theta)$, $Q(\theta, \theta')$ and $\nabla^{1,0}Q(\theta, \theta')$ in terms of nonlinear filtering and nonlinear smoothing conditional densities. Section 5 is devoted to the time–discretization of the stochastic PDE's introduced in Section 4, and the link with MLE of parameters in partially observed Markov chains (hidden Markov models) is explored. A numerical example is presented in Section 6, and the influence of noise covariances is investigated.

2 Statistical model

In this section, expressions for the log–likelihood function $L(\cdot)$ and the auxiliary function $Q(\cdot,\cdot)$ will be derived in the following context [2, Section 3].

Suppose that on a measurable space $(\Omega, \mathcal{F})$ are given

- a family $(P_\theta : \theta \in \Theta)$ of probability measures,
- a pair of stochastic processes $(X_t : t \geq 0)$ and $(Y_t : t \geq 0)$ taking values in $\mathbf{R}^m$ and $\mathbf{R}^d$ respectively,

such that under P_θ

$$\begin{aligned} dX_t &= b_\theta(X_t)\,dt + \sigma(X_t)\,dW_t^\theta\,, \qquad X_0 \sim p_0^\theta(\cdot)\,, \\ dY_t &= h_\theta(X_t)\,dt + d\overline{W}_t^\theta\,, \end{aligned} \tag{3}$$

where $(W_t^\theta : t \geq 0)$ and $(\overline{W}_t^\theta : t \geq 0)$ are independent Wiener processes, with covariance matrix I and r respectively, and the pair is independent from the r.v. X_0.

The following hypotheses are made

(H_1) *$\sigma(\cdot)$ is a continuous and bounded function on $\mathbf{R}^m$ such that $a(\cdot) \stackrel{\Delta}{=} \sigma\sigma^*(\cdot)$ is a uniformly elliptic $m \times m$ matrix, i.e. $a(\cdot) \geq \alpha I$,*

for all $\theta \in \Theta$ open subset of $\mathbf{R}^p$ (the set of parameters)

(H_2) *$p_0^\theta(\cdot)$ is a density on $\mathbf{R}^m$,*

(H_3) *$b_\theta(\cdot)$ is a measurable and bounded function from $\mathbf{R}^m$ to $\mathbf{R}^m$,*

(H_4) *$h_\theta(\cdot)$ is a measurable and bounded function from $\mathbf{R}^m$ to $\mathbf{R}^d$,*

and in addition

(H_5) *the probability measures on $\mathbf{R}^m$ with densities $(p_0^\theta(\cdot) : \theta \in \Theta)$ are mutually absolutely continuous.*

Moreover, it is assumed that $p_0^\theta(\cdot)$, $b_\theta(\cdot)$ and $h_\theta(\cdot)$ are continuously differentiable with respect to the parameter θ and that, for all $\theta \in \Theta$ the derivatives $\nabla b_\theta(\cdot)$ and $\nabla h_\theta(\cdot)$ are measurable and bounded functions from $\mathbf{R}^m$ to $\mathbf{R}^m$ and $\mathbf{R}^d$ respectively (throughout this paper, ∇ will denote the derivation with respect to the parameter θ).

The existence and uniqueness of a weak solution to the stochastic differential equation (3) follows from hypotheses $(H_1 - H_3)$. If moreover hypotheses $(H_4 - H_5)$ hold, then for all $T \geq 0$, $(P_\theta : \theta \in \Theta)$ when restricted to $[0,T]$ are mutually absolutely continuous probability measures on $(\Omega, \mathcal{F})$ with Radon–Nikodym derivative

$$\Lambda_{\theta\theta'} \stackrel{\Delta}{=} \frac{dP_\theta}{dP_{\theta'}} =$$

$$= \frac{p_0^\theta}{p_0^{\theta'}}(X_0) \cdot \exp\left\{ \int_0^T [b_\theta(X_s) - b_{\theta'}(X_s)]^* a^{-1}(X_s)\sigma(X_s)\, dW_s^{\theta'} \right. \tag{4}$$
$$\left. -\tfrac{1}{2}\int_0^T [b_\theta(X_s) - b_{\theta'}(X_s)]^* a^{-1}(X_s)[b_\theta(X_s) - b_{\theta'}(X_s)]\, ds \right\}$$
$$\exp\left\{ \int_0^T [h_\theta(X_s) - h_{\theta'}(X_s)]^* r^{-1}\, d\overline{W}_s^{\theta'} \right.$$
$$\left. -\tfrac{1}{2}\int_0^T [h_\theta(X_s) - h_{\theta'}(X_s)]^* r^{-1}[h_\theta(X_s) - h_{\theta'}(X_s)]\, ds \right\} .$$

Consider also the probability measure $P_\theta^\dagger$ defined on $(\Omega, \mathcal{F})$ by

$$Z^\theta \triangleq \frac{dP_\theta}{dP_\theta^\dagger} = \exp\left\{ \int_0^T h_\theta^*(X_s) r^{-1}\, dY_s - \tfrac{1}{2}\int_0^T h_\theta^*(X_s) r^{-1} h_\theta(X_s)\, ds \right\} ,$$

so that, under $P_\theta^\dagger$

$$dX_t = b_\theta(X_t)\, dt + \sigma(X_t)\, dW_t^\theta , \qquad X_0 \sim p_0^\theta(\cdot) ,$$

where $(W_t^\theta : t \geq 0)$ and $(Y_t^\theta : t \geq 0)$ are independent Wiener processes, with covariance matrix I and r respectively, and the pair is independent from the r.v. X_0. The Radon–Nikodym derivative $\Lambda_{\theta\theta'}$ can then be decomposed as

$$\Lambda_{\theta\theta'} = \Lambda_{\theta\theta'}^\dagger \frac{Z^\theta}{Z^{\theta'}} , \quad \text{with} \quad \Lambda_{\theta\theta'}^\dagger \triangleq \frac{dP_\theta^\dagger}{dP_{\theta'}^\dagger} .$$

It is assumed that only $(Y_t : 0 \leq t \leq T)$ is observed. Let $(\mathcal{Y}_t : 0 \leq t \leq T)$ denote the associated filtration. The likelihood function for the estimation of the parameter θ can be expressed as

$$\mathbf{E}_\alpha^\dagger\left(\frac{dP_\theta}{dP_\alpha^\dagger} \mid \mathcal{Y}_T \right) = \mathbf{E}_\alpha^\dagger(Z^\theta \Lambda_{\theta\alpha}^\dagger \mid \mathcal{Y}_T)$$

with the particular choice $R = P_\alpha^\dagger$ (α fixed in Θ) in (1). By Bayes formula

$$\mathbf{E}_\alpha^\dagger(Z^\theta \Lambda_{\theta\alpha}^\dagger \mid \mathcal{Y}_T) = \mathbf{E}_\theta^\dagger(Z^\theta \mid \mathcal{Y}_T) \cdot \mathbf{E}_\alpha^\dagger(\Lambda_{\theta\alpha}^\dagger \mid \mathcal{Y}_T) = \mathbf{E}_\theta^\dagger(Z^\theta \mid \mathcal{Y}_T)$$

since $\Lambda_{\theta\alpha}^\dagger$ is independent of $\mathcal{Y}_T$ under $P_\alpha^\dagger$. This gives the following expression for the log-likelihood function $L(\cdot)$

$$L(\theta) = \log \mathbf{E}_\theta^\dagger(Z^\theta \mid \mathcal{Y}_T) . \tag{5}$$

For the auxiliary function $Q(\cdot,\cdot)$ defined by (2), one has immediately

$$Q(\theta, \theta') = \mathbf{E}_{\theta'}(\lambda^{\theta,\theta'} \mid \mathcal{Y}_T) = \frac{\mathbf{E}_{\theta'}^\dagger(\lambda^{\theta,\theta'} Z^{\theta'} \mid \mathcal{Y}_T)}{\mathbf{E}_{\theta'}^\dagger(Z^{\theta'} \mid \mathcal{Y}_T)} \tag{6}$$

where

$$\lambda^{\theta,\theta'} \triangleq \log \Lambda_{\theta,\theta'}$$

$$= \log \frac{p_0^\theta}{p_0^{\theta'}}(X_0) + \int_0^T [b_\theta(X_s) - b_{\theta'}(X_s)]^* a^{-1}(X_s)\sigma(X_s)\, dW_s^{\theta'} \tag{7}$$
$$-\tfrac{1}{2}\int_0^T [b_\theta(X_s) - b_{\theta'}(X_s)]^* a^{-1}(X_s)[b_\theta(X_s) - b_{\theta'}(X_s)]\, ds$$
$$+ \int_0^T [h_\theta(X_s) - h_{\theta'}(X_s)]^* r^{-1}\, d\overline{W}_s^{\theta'}$$
$$-\tfrac{1}{2}\int_0^T [h_\theta(X_s) - h_{\theta'}(X_s)]^* r^{-1}[h_\theta(X_s) - h_{\theta'}(X_s)]\, ds \ .$$

Under additional regularity asumptions on the data $p_0^\theta(\cdot)$, $b_\theta(\cdot)$ and $h_\theta(\cdot)$, it is easy to prove, using results in [12], that both $\theta \mapsto L(\theta)$ and $\theta \mapsto Q(\theta, \theta')$ have a.s. differentiable versions, with gradients given by

$$\nabla L(\theta) \ = \ \mathbf{E}_\theta(\lambda^\theta \mid \mathcal{Y}_T) = \frac{\mathbf{E}_\theta^\dagger(\lambda^\theta Z^\theta \mid \mathcal{Y}_T)}{\mathbf{E}_\theta^\dagger(Z^\theta \mid \mathcal{Y}_T)} \ , \tag{8}$$

$$\nabla^{1,0} Q(\theta, \theta') \ = \ \mathbf{E}_{\theta'}(\lambda^\theta \mid \mathcal{Y}_T) = \frac{\mathbf{E}_{\theta'}^\dagger(\lambda^\theta Z^\theta \mid \mathcal{Y}_T)}{\mathbf{E}_{\theta'}^\dagger(Z^\theta \mid \mathcal{Y}_T)} \ , \tag{9}$$

respectively, where

$$\lambda^\theta \triangleq \nabla^{1,0} \log \Lambda_{\theta,\theta'} = \nabla^{1,0} \lambda^{\theta,\theta'}$$
$$= \frac{\nabla p_0^\theta}{p_0^\theta}(X_0) + \int_0^T [\nabla b_\theta(X_s)]^* a^{-1}(X_s)\sigma(X_s)\, dW_s^\theta + \int_0^T [\nabla h_\theta(X_s)]^* r^{-1}\, d\overline{W}_s^\theta \tag{10}$$

is independent of θ'.

Remark. One can check from (8) and (9) that

$$\nabla^{1,0} Q(\theta, \theta') \mid_{\theta=\theta'} = \nabla L(\theta') \ ,$$

as expected.

In the next section, two different methods will be given – by means of stochastic PDE's – to compute the various quantities introduced so far: $L(\theta)$, $\nabla L(\theta)$, $Q(\theta, \theta')$ and $\nabla^{1,0} Q(\theta, \theta')$. This will make possible the numerical implementation of algorithms for the maximization of the likelihood function.

3 Smoothing vs. filtering for the computation of a class of conditional expectations

For the sake of simplicity, any reference to the parameter θ will be dropped throughout this section. In particular, P will denote the probability measure under which

$$\begin{aligned} dX_t &= b(X_t)\,dt + \sigma(X_t)\,dW_t\ , \qquad X_0 \sim p_0(\cdot)\ , \\ dY_t &= h(X_t)\,dt + d\overline{W}_t\ , \end{aligned}$$

where $(W_t\ :\ 0 \le t \le T)$ and $(\overline{W}_t\ :\ 0 \le t \le T)$ are independent Wiener processes, with covariance matrix I and r respectively, and the pair is independent from the r.v. X_0, whereas under $P^\dagger$

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t\ , \qquad X_0 \sim p_0(\cdot)\ ,$$

where $(W_t\ :\ 0 \le t \le T)$ and $(Y_t\ :\ 0 \le t \le T)$ are independent Wiener processes, with covariance matrix I and r respectively, and the pair is independent from the r.v. X_0. Therefore $P = Z_T \cdot P^\dagger$, where the process $(Z_t\ :\ 0 \le t \le T)$ is defined by

$$Z_t = \exp\left\{\int_0^t h^*(X_s) r^{-1}\,dY_s - \tfrac{1}{2}\int_0^t h^*(X_s) r^{-1} h(X_s)\,ds\right\}\ .$$

The purpose of this section is to provide two different methods – one based on nonlinear filtering, the other on nonlinear smoothing – for the computation of the following class of conditional expectations

$$A \triangleq \mathbf{E}\left(\beta(X_0) + \int_0^T \xi(X_s)\,ds + \int_0^T \eta^*(X_s)\,d\overline{W}_s + \int_0^T \chi^*(X_s)\sigma(X_s)\,dW_s \mid \mathcal{Y}_T\right)\ , \tag{11}$$

where β, ξ, η and χ are measurable and bounded functions from $\mathbf{R}^m$ to $\mathbf{R}$, $\mathbf{R}$, $\mathbf{R}^d$ and $\mathbf{R}^m$ respectively. It is readily seen from (6–10) that the computation of either $\nabla L(\theta)$, $Q(\theta,\theta')$ or $\nabla^{1,0}Q(\theta,\theta')$ involves such conditional expectations.

It is clear from the definition that A depends linearly on (β,ξ,η,χ). It will turn out that nonlinear smoothing is the only way to make this dependence explicit, although nonlinear filtering – which is simpler – is enough to just compute A.

Rewriting A as

$$\begin{aligned} A \;=\; & \mathbf{E}(\beta(X_0) \mid \mathcal{Y}_T) + \int_0^T \mathbf{E}(\xi(X_s) - \eta^*(X_s)h(X_s) \mid \mathcal{Y}_T)\,ds + \mathbf{E}\left(\int_0^T \eta^*(X_s)\,dY_s \mid \mathcal{Y}_T\right) \\ & + \mathbf{E}\left(\int_0^T \chi^*(X_s)\sigma(X_s)\,dW_s \mid \mathcal{Y}_T\right)\ , \end{aligned} \tag{12}$$

one would like to interchange conditional expectation and stochastic integral in the third term of (12). However, the resulting expression

$$\text{“}\int_0^T \mathbf{E}(\eta^*(X_s) \mid \mathcal{Y}_T)\,dY_s\text{”} \tag{13}$$

is not an Itô integral, since the integrand is obviously not adapted to the filtration $(\mathcal{Y}_t : 0 \leq t \leq T)$, and needs to be given a rigorous meaning. Although the natural generalization of Itô integral that allows anticipating integrands is Skorokhod integral, it will be proved in Proposition 3.3 below that the correct statement is

$$\mathbf{E}\left(\int_0^T \eta^*(X_s)\, dY_s \mid \mathcal{Y}_T\right) = \mathbf{E}\left(\int_0^T \eta^*(X_s) \circ dY_s \mid \mathcal{Y}_T\right)$$
$$= \int_0^T \mathbf{E}(\eta^*(X_s) \mid \mathcal{Y}_T) \circ dY_s \neq \int_0^T \mathbf{E}(\eta^*(X_s) \mid \mathcal{Y}_T)\, dY_s ,$$

where the non–adapted stochastic integrals are respectively a generalized Stratonovitch integral and a Skorokhod integral [6].

In addition, there seems to be no computable expression available for the last term of (12). However, in the particular case where χ derives from a scalar potential function, one has the following

Proposition 3.1. *Assume there exists a scalar function $U \in C_b^2(\mathbf{R}^m)$ such that $\chi = DU$. Then*

$$\mathbf{E}\left(\int_0^T \chi^*(X_s)\sigma(X_s)\, dW_s \mid \mathcal{Y}_T\right) =$$
$$\mathbf{E}(U(X_T) \mid \mathcal{Y}_T) - \mathbf{E}(U(X_0) \mid \mathcal{Y}_T) - \int_0^T \mathbf{E}(\mathcal{L}U(X_s) \mid \mathcal{Y}_T)\, ds , \tag{14}$$

whose proof follows immediately from Itô's lemma.

At this point, it is necessary to introduce some notations and definitions related to nonlinear filtering and smoothing.

Notations and definitions

- *Filtering*

π_t (resp. p_t) will denote the normalized (resp. unnormalized) conditional density of the r.v. X_t given $\mathcal{Y}_t$, i.e.

$$(\pi_t, \phi) \triangleq \mathbf{E}(\phi(X_t) \mid \mathcal{Y}_t) , \qquad (p_t, \phi) \triangleq \mathbf{E}^\dagger(\phi(X_t)Z_t \mid \mathcal{Y}_t) \tag{15}$$

for any test–function ϕ. By Bayes formula

$$(\pi_t, \phi) = \frac{(p_t, \phi)}{(p_t, 1)} . \tag{16}$$

The equation for $(p_t : 0 \leq t \leq T)$ is Zakaï equation [8]

$$dp_t = \mathcal{L}^* p_t\, dt + h^* p_t r^{-1}\, dY_t , \tag{17}$$

where $\mathcal{L}^*$ denotes the adjoint operator of the infinitesimal generator $\mathcal{L}$ of the diffusion process $(X_t : 0 \leq t \leq T)$, defined by

$$\mathcal{L} \triangleq \tfrac{1}{2} \sum_{i,j=1}^m a^{ij}(\cdot) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^m b^i(\cdot) \frac{\partial}{\partial x_i} .$$

• *Smoothing (fixed-interval)*

Let $T > 0$ denote the fixed end-time. ρ_t (resp. q_t) will denote the normalized (resp. unnormalized) conditional density of the r.v. X_t given $\mathcal{Y}_T$, i.e.

$$(\rho_t, \phi) \triangleq \mathbf{E}(\phi(X_t) \mid \mathcal{Y}_T) , \qquad (q_t, \phi) \triangleq \mathbf{E}^\dagger(\phi(X_t) Z_T \mid \mathcal{Y}_T) .$$

Again

$$(\rho_t, \phi) = \frac{(q_t, \phi)}{(q_t, 1)} . \tag{18}$$

Introducing the backward Zakaï equation

$$dv_t + \mathcal{L} v_t \, dt + h^* v_t r^{-1} \, dY_t = 0 , \qquad v_T \equiv 1 , \tag{19}$$

one has [8,9] that (p_t, v_t) is independent of t, and $q_t = p_t v_t$ is differentiable with

$$\dot{q}_t + p_t \mathcal{L} v_t = v_t \mathcal{L}^* p_t . \tag{20}$$

Note that

$$(q_t, 1) = (p_T, 1) , \quad 0 \le t \le T . \tag{21}$$

3.1 Filtering approach

Define

$$\lambda_t \triangleq \beta(X_0) + \int_0^t \xi(X_s) \, ds + \int_0^t \eta^*(X_s) \, d\overline{W}_s + \int_0^t \chi^*(X_s) \sigma(X_s) \, dW_s$$

so that, by Bayes formula

$$A = \mathbf{E}(\lambda_T \mid \mathcal{Y}_T) = \frac{\mathbf{E}^\dagger(\lambda_T Z_T \mid \mathcal{Y}_T)}{\mathbf{E}^\dagger(Z_T \mid \mathcal{Y}_T)} .$$

A first method would be to compute the joint conditional law of (X_T, λ_T) given $\mathcal{Y}_T$, and then integrate over the first variable to get the marginal conditional law of λ_T given $\mathcal{Y}_T$. An alternative method is to find an equation for $(w_t : 0 \le t \le T)$ defined by

$$(w_t, \phi) \triangleq \mathbf{E}^\dagger(\phi(X_t) \lambda_t Z_t \mid \mathcal{Y}_t) .$$

Indeed, by Itô's lemma

$$\begin{aligned} d[\phi(X_t)\lambda_t Z_t] = \; & \lambda_t Z_t \, \mathcal{L}\phi(X_t) \, dt + \lambda_t Z_t \, (D\phi(X_t))^* \, \sigma(X_t) \, dW_t \\ & + \phi(X_t) Z_t \, \xi(X_t) \, dt + \phi(X_t) Z_t \, \eta^*(X_t) \, d\overline{W}_t + \phi(X_t) Z_t \, \chi^*(X_t) \, \sigma(X_t) \, dW_t \\ & + \phi(X_t) \lambda_t Z_t \, h^*(X_t) \, r^{-1} \, dY_t + \phi(X_t) \, \eta^*(X_t) \, h(X_t) \, Z_t \, dt \\ & + Z_t \, \chi(X_t)^* \, a(X_t) \, D\phi(X_t) \, dt . \end{aligned}$$

Using properties of conditional expectation given the observation σ–algebra under the reference probability measure $P^\dagger$, and the definition (15), gives

$$\begin{aligned}(w_t,\phi) \;=\; & (p_0,\beta\phi) + \int_0^t (w_s,\mathcal{L}\phi)\,ds + \int_0^t (w_s,h^*\phi)r^{-1}\,dY_s \\ & + \int_0^t (p_s,\xi\phi)\,ds + \int_0^t (p_s,\eta^*\phi)\,dY_s + \int_0^t (p_s,\mathcal{J}(\chi)\phi)\,ds\;,\end{aligned}$$

where $\mathcal{J}(\chi)\phi \triangleq \chi^* a D\phi$, so that $(w_t\;:\;0\le t\le T)$ solves

$$dw_t = \mathcal{L}^* w_t\,dt + h^* w_t r^{-1}\,dY_t + \xi p_t\,dt + \eta^* p_t\,dY_t + \mathcal{J}^*(\chi)p_t\,dt\;, \qquad w_0 = \beta p_0\;. \tag{22}$$

Theorem 3.2. *Let $(p_t\;:\;0\le t\le T)$ and $(w_t\;:\;0\le t\le T)$ be the unique solution of (17) and (22) respectively. Then, the following expression holds for A defined in (11)*

$$A = \frac{(w_T,1)}{(p_T,1)}\;. \tag{23}$$

This expression is actually computable. Unfortunately, the linear dependence of $(w_T,1)$ on (β,ξ,η,χ) is not made explicit, which should be the case for the point **[M]** introduced in the Introduction to be satisfied. Therefore, the next step will be to make this dependence more explicit. This will involve nonlinear smoothing and generalized stochastic calculus (Skorokhod integral). Actually

- the stochastic integral in (13) will be given a rigorous meaning,
- the last term in (12) will also be given a computable expression, whether or not χ derives from a scalar potential function.

3.2 Smoothing approach

The idea here is to compute the stochastic differential of the scalar product (w_t,v_t), where $(v_t\;:\;0\le t\le T)$ is the solution of the backward Zakaï equation (19). Since (22) is a forward stochastic PDE and (19) is a backward stochastic PDE, one must use the two–sided stochastic calculus introduced in [10,11]. This gives

$$\begin{aligned} d(w_t,v_t) \;=\; & (\mathcal{L}^* w_t,v_t)\,dt + (h^* w_t,v_t)r^{-1}\,dY_t \\ & + (\xi p_t,v_t)\,dt + (\eta^* p_t,v_t)\,dY_t + (\mathcal{J}^*(\chi)p_t,v_t)\,dt \\ & - (w_t,\mathcal{L}v_t)\,dt - (w_t,h^* v_t)r^{-1}\,dY_t \\ \;=\; & (q_t,\xi)\,dt + (q_t,\eta^*)\,dY_t + (p_t,\chi^* a\,Dv_t)\,dt\;. \end{aligned}$$

Integrating from 0 to T gives

$$(w_T,1) = (q_0,\beta) + \int_0^T (q_s,\xi)\,ds + \int_0^T (q_s,\eta^*)\,dY_s + \int_0^T (p_s,\chi^* a\,Dv_s)\,ds\;,$$

where the stochastic integral is a two–sided stochastic integral.

Using (21) gives an expression for A in terms of normalized conditional densities

$$A = (\rho_0,\beta) + \int_0^T (\rho_s,\xi)\,ds + A' + A''\;.$$

• *Study of A'*

$$A' \triangleq \frac{\int_0^T (q_s, \eta^*)\, dY_s}{(p_T, 1)} \,. \tag{24}$$

One has

$$\begin{aligned} \mathbf{E}(A') &= \mathbf{E}^\dagger(Z_T\, A') = \mathbf{E}^\dagger(\mathbf{E}^\dagger(Z_T \mid \mathcal{Y}_T)\, A') \\ &= \mathbf{E}^\dagger\left(\int_0^T (q_s, \eta^*)\, dY_s\right) = 0 \,, \end{aligned}$$

where the last equality follows from results on two–sided stochastic integrals. This was expected, since

$$A' = \mathbf{E}\left(\int_0^T \eta^*(X_s)\, d\overline{W}_s \mid \mathcal{Y}_T\right) \,.$$

Expressions in terms of normalized conditional densities are given by the following

Proposition 3.3. *Let $(\rho_t : 0 \le t \le T)$ denote the normalized smooothing density. Then*

$$\begin{aligned} A' &= \int_0^T (\rho_s, \eta^*)\, dY_s - \int_0^T (\rho_s, \eta^*)\, (\rho_s, h)\, ds \\ &= \int_0^T (\rho_s, \eta^*) \circ dY_s - \int_0^T (\rho_s, \eta^* h)\, ds \,, \end{aligned}$$

where the non–adapted stochastic integrals are respectively a Skorokhod integral and a generalized Stratonovitch integral [6].

PROOF. The idea is to get the denominator $F \triangleq 1/(p_T, 1)$ inside the stochastic integral in (24).

Let first $D_\cdot$ denote, on the probability space $(\Omega, \mathcal{Y}_T, P^\dagger)$, the derivative with respect to the d–dimensional Wiener process $(Y_t : 0 \le t \le T)$ in the direction of the vector space $H^1(0,T;\mathbf{R}^d)$. Since the two–sided integral is a particular case of the Skorokhod integral, it follows from [6, Proposition 3.2] that

$$\begin{aligned} A' &= F \int_0^T (q_s, \eta^*)\, dY_s \\ &= \int_0^T F\, (q_s, \eta^*)\, dY_s + \int_0^T (q_s, \eta^*)\, D_s F\, ds \\ &= \int_0^T \frac{(q_s, \eta^*)}{(q_s, 1)}\, dY_s - \int_0^T \frac{(q_s, \eta^*)}{(q_s, 1)^2}\, D_s(p_T, 1)\, ds \,, \end{aligned}$$

where the stochastic integral is a Skorokhod integral.

For s fixed in $[0,T]$, consider the d–dimensional random process $(z_t : 0 \le t \le T)$ defined by $z_t \triangleq D_s p_t$. Clearly $z_t \equiv 0$ for $0 \le t < s$. For $i = 1, \cdots, d$, the process $(z_t^i : s \le t \le T)$ is the unique solution of the forward stochastic PDE [7]

$$dz_t^i = \mathcal{L}^* z_t^i\, dt + h^* z_t^i r^{-1}\, dY_t \,, \qquad z_s^i = h^i p_s \,.$$

Introducing the solution $(v_t : 0 \le t \le T)$ of the backward Zakaï equation (19) and using again the two–sided stochastic calculus gives $d(z_t, v_t) = 0$ for $s \le t \le T$. Therefore

$$(z_T, 1) = (z_s, v_s) = (hp_s, v_s) = (q_s, h) \ ,$$

so that

$$\begin{aligned} A' &= \int_0^T \frac{(q_s, \eta^*)}{(q_s, 1)} \, dY_s - \int_0^T \frac{(q_s, \eta^*)}{(q_s, 1)} \frac{(q_s, h)}{(q_s, 1)} \, ds \\ &= \int_0^T (\rho_s, \eta^*) \, dY_s - \int_0^T (\rho_s, \eta^*) \, (\rho_s, h) \, ds \ . \end{aligned}$$

To get the second expression, consider the d–dimensional random process $(u_t : 0 \le t \le T)$ defined by $u_t \triangleq (\rho_t, \eta)$. The Skorokhod–Stratonovitch transformation for generalized stochastic integrals gives [6, Theorem 7.3]

$$\int_0^T u_s^* \, dY_s = \int_0^T u_s^* \circ dY_s - \tfrac{1}{2} \int_0^T (D_s^+ u_s + D_s^- u_s) \, ds \ ,$$

where

$$D_s^+ u_s \triangleq \lim_{t \downarrow s} \sum_{i=1}^d D_s^i u_t^i \ , \qquad D_s^- u_s \triangleq \lim_{t \uparrow s} \sum_{i=1}^d D_s^i u_t^i \ .$$

It turns out that

$$\begin{aligned} D_s^i u_t^i &= D_s^i (\rho_t, \eta^i) = D_s^i \frac{(q_t, \eta^i)}{(q_t, 1)} \\ &= \frac{D_s^i (q_t, \eta^i)}{(q_t, 1)} - \frac{(q_t, \eta^i) \, D_s^i (q_t, 1)}{(q_t, 1)^2} \ . \end{aligned}$$

Next $D_s q_t = (D_s p_t) \, v_t + p_t \, (D_s v_t)$. In particular $D_s p_t$ has already been studied, and a similar argument for $D_s v_t$ shows that $D_s^+ q_s = D_s^- q_s = h q_s$. Therefore

$$\begin{aligned} D_s^+ u_s = D_s^- u_s &= \frac{(q_s, \eta^* h)}{(q_s, 1)} - \frac{(q_t, \eta^*) \, (q_t, h)}{(q_t, 1)^2} \\ &= (\rho_s, \eta^* h) - (\rho, \eta^*) \, (\rho_s, h) \ . \end{aligned}$$

This finaly gives

$$A' = \int_0^T (\rho_s, \eta^*) \circ dY_s - \int_0^T (\rho_s, \eta^* h) \, ds \ ,$$

where the stochastic integral is now a generalized Stratonovitch integral. □

Remark. In terms of conditional expectations

$$\begin{aligned} A' &= \int_0^T \mathbf{E}(\eta^*(X_s) \mid \mathcal{Y}_T) \, dY_s - \int_0^T \mathbf{E}(\eta^*(X_s) \mid \mathcal{Y}_T) \, \mathbf{E}(h(X_s) \mid \mathcal{Y}_T) \, ds \\ &= \int_0^T \mathbf{E}(\eta^*(X_s) \mid \mathcal{Y}_T) \circ dY_s - \int_0^T \mathbf{E}(\eta^*(X_s) h(X_s) \mid \mathcal{Y}_T) \, ds \ . \end{aligned}$$

• *Study of A''*

$$A'' \triangleq \frac{\int_0^T (p_s, \chi^* a\, Dv_s)\, ds}{(p_T, 1)} \,. \tag{25}$$

One has

$$\begin{aligned} \mathbf{E}(A'') &= \mathbf{E}^\dagger(Z_T\, A'') = \mathbf{E}^\dagger(\mathbf{E}^\dagger(Z_T \mid \mathcal{Y}_T)\, A'') \\ &= \mathbf{E}^\dagger\left(\int_0^T (p_s, \chi^* a\, Dv_s)\, ds\right) = \int_0^T (\mathbf{E}^\dagger(p_s), \chi^* a\, \mathbf{E}^\dagger(Dv_s))\, ds \,, \end{aligned}$$

where the last equality follows from the independence of p_s and v_s under the probability measure $P^\dagger$. Now $\mathbf{E}^\dagger(Dv_s) = D\mathbf{E}^\dagger(v_s) \equiv 0$ since $\mathbf{E}^\dagger(v_s) \equiv 1$. Therefore $\mathbf{E}(A'') = 0$, which was expected since

$$A'' = \mathbf{E}\left(\int_0^T \chi^*(X_s)\sigma(X_s)\, dW_s \mid \mathcal{Y}_T\right) \,.$$

The identities

$$\frac{p_s\, Dv_s}{(q_s, 1)} = \pi_s\, D\left(\frac{\rho_s}{\pi_s}\right) = \rho_s\, D\left(\log \frac{\rho_s}{\pi_s}\right)$$

give the following two other expressions for A'', in terms of normalized conditional densities

$$A'' = \int_0^T (\pi_s, \chi^* a\, D\left(\frac{\rho_s}{\pi_s}\right))\, ds = \int_0^T (\rho_s, \chi^* a\, D\left(\log \frac{\rho_s}{\pi_s}\right))\, ds \,.$$

In the particular case where χ derives from a scalar potential function, it can be checked that (25) reduces to the expression (14) given in Proposition 3.1. Indeed

Proposition 3.4. *Assume there exists a scalar function $U \in C_b^2(\mathbf{R}^m)$ such that $\chi = DU$. Then*

$$A'' = (\rho_T, U) - (\rho_0, U) - \int_0^T (\rho_s, \mathcal{L}U)\, ds \,.$$

PROOF. It follows from the identity $\mathcal{L}(Uv_s) = U\mathcal{L}v_s + v_s\mathcal{L}U + \chi^* a Dv_s$, and from (20 that

$$\begin{aligned} (p_s, \chi^* a Dv_s) &= (p_s, \mathcal{L}(Uv_s)) - (p_s, U\mathcal{L}v_s) - (p_s, v_s\mathcal{L}U) \\ &= (v_s\mathcal{L}^* p_s - p_s\mathcal{L}v_s, U) - (p_s v_s, \mathcal{L}U) = (\dot{q}_s, U) - (q_s, \mathcal{L}U) \,. \end{aligned}$$

Integrating from 0 to T gives

$$\int_0^T (p_s, \chi^* a\, Dv_s)\, ds = (q_T, U) - (q_0, U) - \int_0^T (q_s, \mathcal{L}U)\, ds \,.$$

Dividing by $(p_T, 1)$ and using (21) finishes the proof. □

Remark. In terms of conditional expectations

$$A'' = \mathbf{E}(U(X_T) \mid \mathcal{Y}_T) - \mathbf{E}(U(X_0) \mid \mathcal{Y}_T) - \int_0^T \mathbf{E}(\mathcal{L}U(X_s) \mid \mathcal{Y}_T)\, ds \,,$$

which is exactly (14).

The following theorem has been proved

Theorem 3.5. *Let $(\pi_t : 0 \leq t \leq T)$ and $(\rho_t : 0 \leq t \leq T)$ be the normalized filtering and smoothing density (e.g. obtained from the unique solution $(p_t : 0 \leq t \leq T)$ and $(v_t : 0 \leq t \leq T)$ of (17) and (19) respectively). Then, the following two expressions hold for A defined in (11)*

$$A = (\rho_0, \beta) + \int_0^T (\rho_s, \xi)\, ds + \int_0^T (\rho_s, \chi^* a\, D\left(\log \frac{\rho_s}{\pi_s}\right))\, ds + A' \ ,$$

$$A' = \begin{cases} \displaystyle\int_0^T (\rho_s, \eta^*)\, dY_s - \int_0^T (\rho_s, \eta^*)\, (\rho_s, h)\, ds \ , \\ \displaystyle\int_0^T (\rho_s, \eta^*) \circ dY_s - \int_0^T (\rho_s, \eta^* h)\, ds \ , \end{cases}$$

where the non–adapted stochastic integrals are respectively a Skorokhod integral and a generalized Stratonovitch integral [6].

Conclusion

The advantage of smoothing over filtering is that the linear dependence on (β, ξ, η, χ) is made explicit: provided the underlying probability measure does not change, evaluating A for a different set of data (β, ξ, η, χ) will not require the computation of a new infinite–dimensional conditional density. In the filtering approach, one would have to solve another stochastic PDE, with a different "right–hand side".

On the other hand, from the computational point of view, solving the equation for the smoothing density requires not only the computation but also the storage of the filtering density, and is therefore more expensive. Moreover, in the filtering approach it is enough to integrate the unnormalized filtering density at final time T, whereas in the smoothing approach one has *(i)* at each time t, to integrate some functions involving (ξ, η, χ) against the normalized smoothing density, and *(ii)* to integrate the resulting processes over the interval $[0, T]$.

The next section will be devoted to applying these two approaches to the computation of quantities related to the direct likelihood function maximization and the EM algorithm.

4 Application to the MLE problem

4.1 Direct maximization of the likelihood function

It follows from (5) that the log-likelihood function $L(\theta)$ can be expressed as

$$L(\theta) = \log(p_T^\theta, 1)$$

with – see (17)

$$dp_t^\theta = \mathcal{L}_\theta^* p_t^\theta \, dt + h_\theta^* p_t^\theta r^{-1} \, dY_t \tag{26}$$

and

$$\mathcal{L}_\theta \stackrel{\Delta}{=} \frac{1}{2} \sum_{i,j=1}^{m} a^{ij}(\cdot) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{m} b_\theta^i(\cdot) \frac{\partial}{\partial x_i} \, .$$

It follows from (8) and (10) that $\nabla L(\theta)$ belongs to the class of conditional expectations considered in Section 3. The approach based on filtering (Theorem 3.2) gives

$$\nabla L(\theta) = \frac{(w_T^\theta, 1)}{(p_T^\theta, 1)}$$

with $(p_t^\theta \, : \, 0 \leq t \leq T)$ and $(w_t^\theta \, : \, 0 \leq t \leq T)$ given respectively by (26) and – see (22)

$$dw_t^\theta = \mathcal{L}_\theta^* w_t^\theta \, dt + h_\theta^* w_t^\theta r^{-1} \, dY_t + [\nabla h_\theta]^* p_t^\theta r^{-1} \, dY_t + \mathcal{J}_\theta^* p_t^\theta \, dt \, , \qquad w_0^\theta = \nabla p_0^\theta \, , \tag{27}$$

where $\mathcal{J}_\theta \phi \stackrel{\Delta}{=} [\nabla b_\theta]^* D\phi$.

Remark. Equation (27) is exactly what would be obtained by deriving formally equation (26) with respect to the parameter θ. This result was indeed obtained in [4], relying on the existence of a "robust" (i.e. continuous with respect to observation sample paths) version of Zakaï equation.

If θ is a p–dimensional parameter, then the gradient $(w_t^\theta \, : \, 0 \leq t \leq T)$ is a p–dimensional vector: each component of this vector actually solves a stochastic PDE which is coupled only with $(p_t^\theta \, : \, 0 \leq t \leq T)$ and with no other component; moreover the coupling occurs only through the "right–hand side" and each of these $(p+1)$ stochastic PDE's has the same dynamics. In other words, one has to solve the same stochastic PDE with $(p+1)$ different "right–hand side". Note that smoothing could provide a more efficient method to deal with such a problem.

4.2 The EM algorithm

It follows from (6) and (7) that the auxiliary function $Q(\theta, \theta')$ belongs to the class of conditional expectations considered in Section 3. The approach based on filtering (Theorem 3.2) gives

$$Q(\theta, \theta') = \frac{(w_T^{\theta\theta'}, 1)}{(p_T^{\theta'}, 1)}$$

with $(p_t^{\theta'} : 0 \le t \le T)$ and $(w_t^{\theta\theta'} : 0 \le t \le T)$ given respectively by (26) and – see (22)

$$\begin{aligned} dw_t^{\theta\theta'} &= \mathcal{L}_{\theta'}^* w_t^{\theta\theta'}\, dt + h_{\theta'}^* w_t^{\theta\theta'} r^{-1}\, dY_t + [h_\theta - h_{\theta'}]^* p_t^{\theta'} r^{-1}\, dY_t + \mathcal{J}_{\theta\theta'}^* p_t^{\theta'}\, dt \\ &\quad -\tfrac{1}{2}\left([b_\theta - b_{\theta'}]^* a^{-1}[b_\theta - b_{\theta'}] + [h_\theta - h_{\theta'}]^* r^{-1}[h_\theta - h_{\theta'}]\right) p_t^{\theta'}\, dt\ , \\ w_0^{\theta\theta'} &= p_0^{\theta'} \log \frac{p_0^\theta}{p_0^{\theta'}}\ , \end{aligned}$$

where $\mathcal{J}_{\theta\theta'}\phi \stackrel{\Delta}{=} [b_\theta - b_{\theta'}]^* D\phi$.

On the other hand, smoothing (Theorem 3.5) gives

$$\begin{aligned} Q(\theta,\theta') &= (\rho_0^{\theta'}, \log \frac{p_0^\theta}{p_0^{\theta'}}) + \int_0^T (\rho_s^{\theta'}, [b_\theta - b_{\theta'}]^* D\left(\log \frac{\rho_s^{\theta'}}{\pi_s^{\theta'}}\right)) ds \\ &\quad -\tfrac{1}{2}\int_0^T (\rho_s^{\theta'}, [b_\theta - b_{\theta'}]^* a^{-1}[b_\theta - b_{\theta'}] + [h_\theta - h_{\theta'}]^* r^{-1}[h_\theta - h_{\theta'}])\, ds + A'\ , \end{aligned} \tag{28}$$

$$A' = \begin{cases} \displaystyle\int_0^T (\rho_s^{\theta'}, [h_\theta - h_{\theta'}]^*) r^{-1}\, dY_s - \int_0^T (\rho_s^{\theta'}, [h_\theta - h_{\theta'}]^*) r^{-1} (\rho_s^{\theta'}, h_{\theta'})\, ds\ , \\ \displaystyle\int_0^T (\rho_s^{\theta'}, [h_\theta - h_{\theta'}]^*) r^{-1} \circ dY_s - \int_0^T (\rho_s^{\theta'}, [h_\theta - h_{\theta'}]^* r^{-1} h_{\theta'})\, ds\ , \end{cases} \tag{29}$$

where $(\pi_t^{\theta'} : 0 \le t \le T)$ and $(\rho_t^{\theta'} : 0 \le t \le T)$ are the normalized density of filtering and smoothing, computed from the unique solution $(p_t^{\theta'} : 0 \le t \le T)$ and $(v_t^{\theta'} : 0 \le t \le T)$ of (26) and – see (19)

$$dv_t^{\theta'} + \mathcal{L}_{\theta'} v_t^{\theta'}\, dt + h_{\theta'}^* v_t^{\theta'} r^{-1}\, dY_t = 0\ , \qquad v_T^{\theta'} \equiv 1\ , \tag{30}$$

respectively. Moreover, the non–adapted stochastic integrals in (29) are respectively a Skorokhod integral and a generalized Stratonovitch integral [6].

Remark. It is now possible to give a more precise meaning to the (E–step) and (M–step) of the algorithm. Indeed, θ' being fixed

3. (E–step) compute the normalized smoothing density $(\rho_t^{\theta'} : 0 \le t \le T)$ – this requires in particular to compute the normalized filtering density $(\pi_t^{\theta'} : 0 \le t \le T)$,
4. (M–step) maximize $Q(\cdot,\theta')$ – where for each $\theta \in \Theta$ the computation of $Q(\theta,\theta')$ requires according to (28) *(i)* at each time t, to integrate some functions depending on (θ,θ') against the normalized smoothing density $\rho_t^{\theta'}$, and *(ii)* to integrate the resulting processes over the interval $[0,T]$.

Remark. A partial answer can be given to the question **[M]** raised in the Introduction. Indeed

- the differentiability of $\theta \mapsto Q(\theta,\theta')$ relies in an obvious way on the existence of derivatives with respect to θ of $p_0^\theta(\cdot)$, $b_\theta(\cdot)$ and $h_\theta(\cdot)$,
- computing the corresponding derivatives, and maximizing $\theta \mapsto Q(\theta,\theta')$ will not involve the computation of any other infinite–dimensional conditional density.

Moreover, as was pointed out in [2], there are particular cases in which the M–step can be dealt with explicitely. This includes the case where

- $\log p_0^\theta(\cdot)$ depends quadratically on θ,
- $b_\theta(\cdot)$ and $h_\theta(\cdot)$ depend linearly on θ,

since $\theta \mapsto Q(\theta, \theta')$ becomes then a quadratic form.

It follows from (9) and (10) that $\nabla^{1,0} Q(\theta, \theta')$ belongs to the class of conditional expectations considered in Section 3. The approach based on filtering (Theorem 3.2) gives

$$\nabla^{1,0} Q(\theta, \theta') = \frac{(w_T^{\theta\theta'}, 1)}{(p_T^{\theta'}, 1)}$$

with $(p_t^{\theta'} : 0 \le t \le T)$ and $(w_t^{\theta\theta'} : 0 \le t \le T)$ given respectively by (26) and – see (22)

$$\begin{aligned} dw_t^{\theta\theta'} &= \mathcal{L}_{\theta'}^* w_t^{\theta\theta'}\, dt + h_{\theta'}^* w_t^{\theta\theta'} r^{-1}\, dY_t + [\nabla h_\theta]^* p_t^{\theta'} r^{-1}\, dY_t + \mathcal{J}_\theta^* p_t^{\theta'}\, dt \\ &\quad - \Big([\nabla b_\theta]^* a^{-1} [b_\theta - b_{\theta'}] + [\nabla h_\theta]^* r^{-1} [h_\theta - h_{\theta'}] \Big)\, p_t^{\theta'}\, dt \\ w_0^{\theta\theta'} &= \frac{p_0^{\theta'}}{p_0^\theta} \nabla p_0^\theta \,, \end{aligned}$$

where $\mathcal{J}_\theta \phi \triangleq [\nabla b_\theta]^* D\phi$.

Remark. Comparing with (27), one can check once again that

$$\nabla^{1,0} Q(\theta, \theta') \mid_{\theta=\theta'} = \nabla L(\theta') \,,$$

as expected.

As for the smoothing approach, one can use again the results of Section 3. Alternatively, one can directly differentiate with respect to θ the expression (28) for $Q(\theta, \theta')$, thus illustrating the point [**M**]. Indeed

$$\begin{aligned} \nabla^{1,0}(\theta, \theta') &= (\rho_0^{\theta'}, \frac{\nabla p_0^\theta}{p_0^\theta}) + \int_0^T (\rho_s^{\theta'}, [\nabla b_\theta]^*\, D \left(\log \frac{\rho_s^{\theta'}}{\pi_s^{\theta'}} \right)) \, ds \\ &\quad - \int_0^T (\rho_s^{\theta'}, [\nabla b_\theta]^* a^{-1} [b_\theta - b_{\theta'}] + [\nabla h_\theta]^* r^{-1} [h_\theta - h_{\theta'}])\, ds + A' \,, \\ A' &= \begin{cases} \displaystyle\int_0^T (\rho_s^{\theta'}, [\nabla h_\theta]^*) r^{-1}\, dY_s - \int_0^T (\rho_s^{\theta'}, [\nabla h_\theta]^*) r^{-1} (\rho_s^{\theta'}, h_{\theta'})\, ds \,, \\ \displaystyle\int_0^T (\rho_s^{\theta'}, [\nabla h_\theta]^*) r^{-1} \circ dY_s - \int_0^T (\rho_s^{\theta'}, [\nabla h_\theta]^*) r^{-1} h_{\theta'})\, ds \,, \end{cases} \end{aligned}$$

where the non–adapted stochastic integrals are respectively a Skorokhod integral and a generalized Stratonovitch integral [6].

5 Time–discretization, and relation with MLE of parameters in partially observed Markov chains

Before turning to the presentation of the numerical results, it is worth describing the approach that has been adopted to actually compute the expressions obtained for $L(\theta)$, $\nabla L(\theta)$ and $Q(\theta, \theta')$. From the results of the previous section, this should reduce in some sense to discretizing stochastic PDE's (26), (27) and (30).

However, instead of discretizing separately these stochastic PDE's and e.g. just plugging the resulting approximations into a discretized version of (28), a global approximation of the original continuous–time problem by a discrete–time problem will be presented. In particular

- the approximation $\overline{L}(\theta)$ to the log–likelihood function $L(\theta)$ of the continuous–time problem, will be interpreted as the log–likelihood function of the discrete–time problem,
- the approximation $\overline{Q}(\theta, \theta')$ to the auxiliary function $Q(\theta, \theta')$ of the continuous–time problem will be such that the fundamental relation (2) will hold for the discrete–time problem, i.e. $\overline{L}(\theta) - \overline{L}(\theta') \geq \overline{Q}(\theta, \theta')$.

Consider indeed the following discrete–time statistical model. Let first $(t_n : 0 \leq n \leq N)$ be a uniform partition of the interval $[0, T]$ with time–step Δt. Suppose that on a measurable space $(\Omega, \mathcal{F})$ are given

- a family $(\overline{P}_\theta : \theta \in \Theta)$ of probability measures,
- a discrete–time stochastic process $(\overline{X}_n : 0 \leq n \leq N)$ taking values in $\mathbf{R}^m$,
- a stochastic process $(Y_t : t \geq 0)$ taking values in $\mathbf{R}^d$,

such that under $\overline{P}_\theta$, $(\overline{X}_n : 0 \leq n \leq N)$ is a Markov chain with transition probabilities kernel

$$\Pi_\theta \stackrel{\Delta}{=} (I - \Delta t \mathcal{L}_\theta)^{-1} \tag{31}$$

and initial density p_0^θ, and this Markov chain is observed in continuous–time through

$$dY_t = h_\theta(\overline{X}_n)\, dt + d\overline{W}_t \;, \qquad t_n \leq t < t_{n+1} \;,$$

where $(\overline{W}_t : 0 \leq t \leq T)$ is a Wiener process with matrix covariance r, independent of the Markov chain $(\overline{X}_n : 0 \leq n \leq N)$.

Remark. Equivalently, one can consider that the Markov chain is observed through the discrete–time measurements

$$y_n \stackrel{\Delta}{=} \frac{\Delta Y_n}{\Delta t} = h_\theta(\overline{X}_n) + \overline{w}_n \qquad (\Delta Y_n \stackrel{\Delta}{=} Y_{t_{n+1}} - Y_{t_n}) \;,$$

where $(\overline{w}_n : 0 \leq n \leq N)$ is a Gaussian white noise sequence with matrix covariance $r \Delta t^{-1}$, independent of the Markov chain $(\overline{X}_n : 0 \leq n \leq N)$.

First, it follows from hypotheses $(H_1 - H_2)$ that $\forall x \in \mathbf{R}^m$, $(\Pi_\theta(x,\cdot) : \theta \in \Theta)$ are mutually absolutely continuous probability measures on $\mathbf{R}^m$. Define then

$$f_{\theta,\theta'}(x,y) \stackrel{\Delta}{=} \frac{\Pi_\theta(x,dy)}{\Pi_{\theta'}(x,dy)} ,$$

as the corresponding Radon–Nikodym derivative. Define next

$$\Psi_n^\theta(x) \stackrel{\Delta}{=} \exp\left\{h_\theta^*(x) r^{-1} \Delta Y_n - \tfrac{1}{2} h_\theta^*(x) r^{-1} h_\theta(x) \Delta t\right\} .$$

Then $(\overline{P}_\theta : \theta \in \Theta)$ are mutually absolutely continuous probability measures on $(\Omega, \mathcal{F})$ with Radon–Nikodym derivative

$$\overline{\Lambda}_{\theta,\theta'} \stackrel{\Delta}{=} \frac{d\overline{P}_\theta}{d\overline{P}_{\theta'}} = \frac{p_0^\theta}{p_0^{\theta'}}(\overline{X}_0) \prod_{i=0}^{N-1} f_{\theta,\theta'}(\overline{X}_i, \overline{X}_{i+1}) \prod_{i=0}^{N-1} \frac{\Psi_i^\theta}{\Psi_i^{\theta'}}(\overline{X}_i) .$$

Consider also the probability measure $\overline{P}_\theta^\dagger$ defined by

$$\overline{Z}^\theta \stackrel{\Delta}{=} \frac{d\overline{P}_\theta}{d\overline{P}_\theta^\dagger} = \prod_{i=0}^{N-1} \Psi_i^\theta(\overline{X}_i) ,$$

so that under $\overline{P}_\theta^\dagger$, $(Y_t : 0 \leq t \leq T)$ is a Wiener process independent of the Markov chain $(\overline{X}_n : 0 \leq n \leq N)$.

Let again $(\mathcal{Y}_t : 0 \leq t \leq T)$ denote the observation filtration. It turns out that the log–likelihood function for the estimation of the parameter θ is now defined by

$$\overline{L}(\theta) = \log \overline{\mathbf{E}}_\theta^\dagger(\overline{Z}^\theta \mid \mathcal{Y}_T) , \tag{32}$$

whereas the auxiliary function is defined by

$$\overline{Q}(\theta,\theta') = \overline{\mathbf{E}}_{\theta'}(\log \overline{\Lambda}_{\theta,\theta'} \mid \mathcal{Y}_T) = \frac{\overline{\mathbf{E}}_{\theta'}^\dagger(\log \overline{\Lambda}_{\theta,\theta'} \overline{Z}^{\theta'} \mid \mathcal{Y}_T)}{\overline{\mathbf{E}}_{\theta'}^\dagger(\overline{Z}^{\theta'} \mid \mathcal{Y}_T)} . \tag{33}$$

5.1 Direct maximization of the likelihood function

The idea is to find an equation for $(\overline{p}_n^\theta : 0 \leq n \leq N)$ defined by

$$(\overline{p}_n^\theta, \phi) \stackrel{\Delta}{=} \overline{\mathbf{E}}_\theta^\dagger(\phi(\overline{X}_n)\overline{Z}_n^\theta \mid \mathcal{Y}_{t_n}) ,$$

where

$$\overline{Z}_n^\theta \stackrel{\Delta}{=} \prod_{i=0}^{n-1} \Psi_i^\theta(\overline{X}_i) .$$

By definition

$$\begin{aligned}
(\overline{p}_{n+1}^\theta, \phi) &= \overline{\mathbf{E}}_\theta^\dagger(\phi(\overline{X}_{n+1})\overline{Z}_{n+1}^\theta \mid \mathcal{Y}_{t_{n+1}}) \\
&= \overline{\mathbf{E}}_\theta^\dagger(\phi(\overline{X}_{n+1})\Psi_n^\theta(\overline{X}_n)\overline{Z}_n^\theta \mid \mathcal{Y}_{t_{n+1}}) \\
&= \overline{\mathbf{E}}_\theta^\dagger(\Psi_n^\theta(\overline{X}_n^\theta)\,[\Pi_\theta\phi](\overline{X}_n)\,\overline{Z}_n^\theta \mid \mathcal{Y}_{t_{n+1}}) \\
&= (\overline{p}_n^\theta, \Psi_n^\theta(\Pi_\theta\phi)) ,
\end{aligned}$$

which results in the following equation

$$\overline{p}^{\theta}_{n+1} = \Pi^{*}_{\theta}(\Psi^{\theta}_{n}\overline{p}^{\theta}_{n}) \ , \qquad \overline{p}^{\theta}_{0} = p^{\theta}_{0} \ . \tag{34}$$

Using expression (31) for the transition probabilities kernel gives the following discretization scheme of Zakaï equation (26), which combines a Trotter–like product formula and a Euler implicit scheme

$$(I - \Delta t \mathcal{L}^{*}_{\theta})\,\overline{p}^{\theta}_{n+1} = \Psi^{\theta}_{n}\,\overline{p}^{\theta}_{n} \ , \qquad \overline{p}^{\theta}_{0} = p^{\theta}_{0} \ . \tag{35}$$

It follows from (32) that the log–likelihood function $L(\theta)$ is therefore approximated by

$$\overline{L}(\theta) = \log(\overline{p}^{\theta}_{N}, 1) \ . \tag{36}$$

To approximate the gradient $\nabla L(\theta)$, one could either

- directly discretize equation (27),
- derive the exact expression for the gradient of the approximated log–likelihood function $\overline{L}(\theta)$.

The second method is prefered, and gives

$$\nabla \overline{L}(\theta) = \frac{(\overline{w}^{\theta}_{N}, 1)}{(\overline{p}^{\theta}_{N}, 1)} \ ,$$

with – deriving equation (35) with respect to the parameter θ

$$(I - \Delta t \mathcal{L}^{*}_{\theta})\,\overline{w}^{\theta}_{n+1} = \Psi^{\theta}_{n}\overline{w}^{\theta}_{n} + \Delta t[\nabla \mathcal{L}^{*}_{\theta}]\,\overline{p}^{\theta}_{n+1} + [\nabla \Psi^{\theta}_{n}]\,\overline{p}^{\theta}_{n} \ , \qquad \overline{w}^{\theta}_{0} = \nabla p^{\theta}_{0} \ .$$

Remark. (normalization) To avoid numerical overflow one should rather solve, instead of (35), the normalized equations

$$\left.\begin{aligned} \overline{\pi}^{\theta}_{n+\frac{1}{2}} &= \Psi^{\theta}_{n}\,\overline{\pi}^{\theta}_{n}/l^{\theta}_{n+1} \\ (I - \Delta t \mathcal{L}^{*}_{\theta})\,\overline{\pi}^{\theta}_{n+1} &= \overline{\pi}^{\theta}_{n+\frac{1}{2}} \end{aligned}\right\} \qquad \overline{\pi}^{\theta}_{0} = p^{\theta}_{0} \ ,$$

where $l^{\theta}_{n+1} \triangleq (\overline{\pi}^{\theta}_{n}, \Psi^{\theta}_{n})$. It is easily seen that $\overline{p}^{\theta}_{n} = \gamma^{\theta}_{n}\overline{\pi}^{\theta}_{n}$ with $\gamma^{\theta}_{n} \triangleq l^{\theta}_{n} \cdot l^{\theta}_{n-1} \cdots l^{\theta}_{1}$ and $(\overline{p}^{\theta}_{n}, 1) = \gamma^{\theta}_{n}$ so that

$$\overline{L}(\theta) = \log \gamma^{\theta}_{N} = \sum_{i=1}^{N-1} \log\,(\overline{\pi}^{\theta}_{i}, \Psi^{\theta}_{i}) \ .$$

In the same way, defining $\overline{\overline{w}}^{\theta}_{n}$ by the relation $\overline{w}^{\theta}_{n} = \gamma^{\theta}_{n}\overline{\overline{w}}^{\theta}_{n}$ gives

$$\left.\begin{aligned} (I - \Delta t \mathcal{L}^{*}_{\theta})\,\overline{\overline{w}}^{\theta}_{n+\frac{1}{2}} &= \Psi^{\theta}_{n}\,\overline{\overline{w}}^{\theta}_{n} + \Delta t[\nabla \mathcal{L}^{*}_{\theta}]\,\overline{\pi}^{\theta}_{n+\frac{1}{2}} + [\nabla \Psi^{\theta}_{n}]\,\overline{\pi}^{\theta}_{n} \\ \overline{\overline{w}}^{\theta}_{n+1} &= \overline{\overline{w}}^{\theta}_{n+\frac{1}{2}} \cdot (l^{\theta}_{n+1})^{-1} \end{aligned}\right\} \qquad \overline{\overline{w}}^{\theta}_{0} = \nabla p^{\theta}_{0} \ .$$

Note that, although $\overline{w}^{\theta}_{n}$ is the gradient of $\overline{p}^{\theta}_{n}$, $\overline{\overline{w}}^{\theta}_{n}$ is *not* the gradient of $\overline{\pi}^{\theta}_{n}$. Actually $\overline{\overline{w}}^{\theta}_{n} = \overline{w}^{\theta}_{n}/(\overline{p}^{\theta}_{n}, 1)$ so that

$$\nabla \overline{L}(\theta) = (\overline{\overline{w}}^{\theta}_{N}, 1) \ .$$

5.2 The EM algorithm

Although it is rather straightforward, in the discrete–time case, to obtain the expression of the auxiliary function $\overline{Q}(\cdot,\cdot)$ in terms of nonlinear smoothing, it is nevertheless worth presenting a derivation that follows the same lines as in the continuous–time case. Indeed, there are two different methods – one based on nonlinear filtering, the other on nonlinear smoothing – for the computation of (33).

• *Filtering*

Define

$$\overline{\lambda}_n^{\theta,\theta'} = \log \frac{p_0^\theta}{p_0^{\theta'}}(\overline{X}_0) + \sum_{i=0}^{n-1} \log f_{\theta,\theta'}(\overline{X}_i, \overline{X}_{i+1}) + \sum_{i=0}^{n-1} \log \frac{\Psi_i^\theta}{\Psi_i^{\theta'}}(\overline{X}_i) \ .$$

The idea again is to find an equation for $(\overline{w}_n^{\theta,\theta'} : 0 \le n \le N)$ defined by

$$(\overline{w}_n^{\theta,\theta'}, \phi) \triangleq \overline{\mathbf{E}}_{\theta'}^\dagger(\phi(\overline{X}_n)\overline{\lambda}_n^{\theta,\theta'}\overline{Z}_n^{\theta'} \mid \mathcal{Y}_{t_n}) \ .$$

First

$$\overline{w}_0^{\theta,\theta'} = p_0^{\theta'} \log \frac{p_0^\theta}{p_0^{\theta'}} \ .$$

Next, by definition

$$\begin{aligned}
(\overline{w}_{n+1}^{\theta,\theta'}, \phi) &= \overline{\mathbf{E}}_{\theta'}^\dagger(\phi(\overline{X}_{n+1})\overline{\lambda}_{n+1}^{\theta,\theta'}\overline{Z}_{n+1}^{\theta'} \mid \mathcal{Y}_{t_{n+1}}) \\
&= \overline{\mathbf{E}}_{\theta'}^\dagger(\phi(\overline{X}_{n+1})\Psi_n^{\theta'}(\overline{X}_n)[\overline{\lambda}_n^{\theta,\theta'} + \log f_{\theta,\theta'}(\overline{X}_n, \overline{X}_{n+1}) + \log \frac{\Psi_n^\theta}{\Psi_n^{\theta'}}(\overline{X}_n)]\overline{Z}_n^{\theta'} \mid \mathcal{Y}_{t_{n+1}}) \\
&= \overline{\mathbf{E}}_{\theta'}^\dagger(\Psi_n^{\theta'}(\overline{X}_n)\,[\Pi_{\theta'}\phi](\overline{X}_n)\,\overline{\lambda}_n^{\theta,\theta'}\,\overline{Z}_n^{\theta'} \mid \mathcal{Y}_{t_{n+1}}) + \overline{\mathbf{E}}_{\theta'}^\dagger(\Psi_n^{\theta'}(\overline{X}_n)\,[\kappa_{\theta,\theta'}\phi](\overline{X}_n)\,\overline{Z}_n^{\theta'} \mid \mathcal{Y}_{t_{n+1}}) \\
&\quad + \overline{\mathbf{E}}_{\theta'}^\dagger(\Psi_n^{\theta'}(\overline{X}_n)\,\log \frac{\Psi_n^\theta}{\Psi_n^{\theta'}}(\overline{X}_n)\,[\Pi_{\theta'}\phi](\overline{X}_n)\,\overline{Z}_n^{\theta'} \mid \mathcal{Y}_{t_{n+1}}) \\
&= (\overline{w}_n^{\theta,\theta'}, \Psi_n^{\theta'}(\Pi_{\theta'}\phi)) + (\overline{p}_n^{\theta'}, \Psi_n^{\theta'}(\kappa_{\theta,\theta'}\phi)) + (\overline{p}_n^{\theta'}, \Psi_n^{\theta'} \log \frac{\Psi_n^\theta}{\Psi_n^{\theta'}}(\Pi_{\theta'}\phi)) \ ,
\end{aligned}$$

where the operator $\kappa_{\theta,\theta'}$ is defined by

$$(\kappa_{\theta,\theta'}\phi)(x) \triangleq \int \phi(y) \log f_{\theta,\theta'}(x,y)\, \Pi_{\theta'}(x, dy) \ . \tag{37}$$

Therefore, the resulting equation is

$$\overline{w}_{n+1}^{\theta,\theta'} = \Pi_{\theta'}^*(\Psi_n^{\theta'}\overline{w}_n^{\theta,\theta'}) + \kappa_{\theta,\theta'}^*(\Psi_n^{\theta'}\overline{p}_n^{\theta'}) + \Pi_{\theta'}^*(\Psi_n^{\theta'} \log \frac{\Psi_n^\theta}{\Psi_n^{\theta'}}\overline{p}_n^{\theta'}) \ , \qquad \overline{w}_0^{\theta,\theta'} = p_0^{\theta'} \log \frac{p_0^\theta}{p_0^{\theta'}} \ .$$

It follows from (33) that the auxiliary function $Q(\theta,\theta')$ is approximated by

$$\overline{Q}(\theta,\theta') = \frac{(\overline{w}_N^{\theta,\theta'}, 1)}{(\overline{p}_N^{\theta'}, 1)} \ . \tag{38}$$

• *Smoothing*

Introduce the backward equation – dual to (34)

$$\overline{v}_n^{\theta'} = \Psi_n^{\theta'}(\Pi_{\theta'}\overline{v}_{n+1}^{\theta'}) , \qquad \overline{v}_N^{\theta'} \equiv 1 . \tag{39}$$

Then

$$(\overline{w}_{n+1}^{\theta,\theta'}, \overline{v}_{n+1}^{\theta'}) = (\Pi_{\theta'}^*(\Psi_n^{\theta'}\overline{w}_n^{\theta,\theta'}), \overline{v}_{n+1}^{\theta'}) + (\kappa_{\theta,\theta'}^*(\Psi_n^{\theta'}\overline{p}_n^{\theta'}), \overline{v}_{n+1}^{\theta'}) + (\Pi_{\theta'}^*(\Psi_n^{\theta'} \log \frac{\Psi_n^{\theta}}{\Psi_n^{\theta'}}\overline{p}_n^{\theta'}), \overline{v}_{n+1}^{\theta'})$$

$$= (\overline{w}_n^{\theta,\theta'}, \overline{v}_n^{\theta'}) + (\Psi_n^{\theta'}\overline{p}_n^{\theta'}, [\kappa_{\theta,\theta'}\overline{v}_{n+1}^{\theta'}]) + (\overline{p}_n^{\theta'}\overline{v}_n^{\theta'}, \log \frac{\Psi_n^{\theta}}{\Psi_n^{\theta'}}) .$$

Introducing the unnormalized smoothing density $\overline{q}_i^{\theta'} = \overline{p}_i^{\theta'}\overline{v}_i^{\theta'}$, gives

$$(\overline{w}_N^{\theta,\theta'}, 1) = (\overline{w}_0^{\theta,\theta'}, \overline{v}_0^{\theta'}) + \sum_{i=0}^{N-1}[(\overline{w}_{i+1}^{\theta,\theta'}, \overline{v}_{i+1}^{\theta'}) - (\overline{w}_i^{\theta,\theta'}, \overline{v}_i^{\theta'})]$$

$$= (\overline{q}_0^{\theta'}, \log \frac{p_0^{\theta}}{p_0^{\theta'}}) + \sum_{i=0}^{N-1}(\Psi_i^{\theta'}\overline{p}_i^{\theta'}, [\kappa_{\theta,\theta'}\overline{v}_{i+1}^{\theta'}]) + \sum_{i=0}^{N-1}(\overline{q}_i^{\theta'}, \log \frac{\Psi_n^{\theta}}{\Psi_n^{\theta'}})$$

$$= (\overline{q}_0^{\theta'}, \log \frac{p_0^{\theta}}{p_0^{\theta'}}) + \sum_{i=0}^{N-1}(\overline{p}_{i+\frac{1}{2}}^{\theta'}, [\kappa_{\theta,\theta'}\overline{v}_{i+1}^{\theta'}])$$

$$+ \sum_{i=0}^{N-1}(\overline{q}_i^{\theta'}, [h_\theta - h_{\theta'}]^* r^{-1}(\Delta Y_i - h_{\theta'}\Delta t)) - \tfrac{1}{2}\sum_{i=0}^{N-1}(\overline{q}_i^{\theta'}, [h_\theta - h_{\theta'}]^* r^{-1}[h_\theta - h_{\theta'}])\Delta t ,$$

where in the last expression $\overline{p}_{i+\frac{1}{2}}^{\theta'} \triangleq \Psi_i^{\theta'}\overline{p}_i^{\theta'}$, and the identity

$$\log \frac{\Psi_i^{\theta}}{\Psi_i^{\theta'}} = [h_\theta - h_{\theta'}]^* r^{-1}(\Delta Y_i - h_{\theta'}\Delta t) - \tfrac{1}{2}[h_\theta - h_{\theta'}]^* r^{-1}[h_\theta - h_{\theta'}]\Delta t$$

has been used.

Remark. (normalization) Here again one should rather solve, instead of (39), the normalized equations

$$\left.\begin{aligned} \overline{\overline{v}}_{n+\frac{1}{2}}^{\theta'} &= \Pi_{\theta'}\overline{\overline{v}}_{n+1}^{\theta'} \\ \overline{\overline{v}}_n^{\theta'} &= \Psi_n^{\theta'}\overline{\overline{v}}_{n+\frac{1}{2}}^{\theta'}/j_n^{\theta'} \end{aligned}\right\} \qquad \overline{\overline{v}}_N^{\theta'} \equiv 1 ,$$

where $j_n^{\theta'}$ is chosen in such a way that $(\overline{\pi}_n^{\theta'}, \overline{\overline{v}}_n^{\theta'}) = 1$, which gives $j_n^{\theta'} = (\overline{\pi}_n^{\theta'}, \Psi_n^{\theta'}\overline{\overline{v}}_{n+\frac{1}{2}}^{\theta'})$. It is then easily seen that $j_n^{\theta'} = l_{n+1}^{\theta'}$, and that $\overline{v}_n^{\theta'} = \delta_n^{\theta'}\overline{\overline{v}}_n^{\theta'}$ with $\delta_n^{\theta'} \triangleq j_n^{\theta'} \cdot j_{n+1}^{\theta'} \cdots j_N^{\theta'}$. Moreover, the normalized smoothing density is given by $\overline{p}_n^{\theta'} \triangleq \overline{\pi}_n^{\theta'}\overline{\overline{v}}_n^{\theta'}$.

Remark. In terms of normalized conditional densities

$$\overline{Q}(\theta,\theta') = (\overline{\rho}_0^{\theta'}, \log \frac{p_0^\theta}{p_0^{\theta'}}) + \sum_{i=0}^{N-1} (\overline{\pi}_{i+\frac{1}{2}}^{\theta'}, [\kappa_{\theta,\theta'} \overline{\overline{v}}_{i+1}^{\theta'}])$$

$$+ \sum_{i=0}^{N-1} (\overline{\rho}_i^{\theta'}, [h_\theta - h_{\theta'}]^* r^{-1} (\Delta Y_i - h_{\theta'} \Delta t)) - \frac{1}{2} \sum_{i=0}^{N-1} (\overline{\rho}_i^{\theta'}, [h_\theta - h_{\theta'}]^* r^{-1} [h_\theta - h_{\theta'}]) \Delta t$$

to be compared with (28), (29).

Remark. It is now possible to give a more precise meaning to the (E–step) and (M–step) of the algorithm. Indeed, θ' being fixed

3. (E–step) compute the normalized smoothing density $(\overline{\rho}_n^{\theta'} : 0 \leq n \leq N)$ – this requires in particular to compute the normalized filtering density $(\overline{\pi}_n^{\theta'} : 0 \leq n \leq N)$,
4. (M–step) maximize $\overline{Q}(\cdot, \theta')$ – where for each $\theta \in \Theta$ the computation of $\overline{Q}(\theta, \theta')$ requires *(i)* at each time n, to integrate some functions depending on (θ, θ') against the normalized smoothing density $\overline{\rho}_n^{\theta'}$, and *(ii)* to sum the resulting discrete–time processes from $n = 0$ to $n = N - 1$.

Remark. With the time–discretization introduced above, the numerical implementation (including discretization with respect to the space variable) of the EM algorithm requires in the M–step, the explicit evaluation of the transition probabilities kernel $\Pi_\theta = (I - \Delta t \mathcal{L}_\theta)^{-1}$. On the other hand, the numerical implementation of the direct maximization algorithm requires only the solution of linear equations with operator $(I - \Delta t \mathcal{L}_\theta^*)$, a much faster task.

Remark. There are some similarity between the discrete–time version of the EM algorithm and the statistical estimation of probabilistic functions of Markov processes. This theory has been introduced in [1], and has found interesting applications in acoustic speech recognition [5]. Indeed, assume that observations are generated according to a hidden Markov model (HMM): to each possible state x of the a non–observed Markov chain defined by its initial probability p_0 and its transition probabilities kernel Π, is associated a probability function $B(x, \cdot)$ which describes the conditional law of the observation given that the chain is in state x. Such a model will be denoted by $\mathcal{M} = (p_0, \Pi, B)$. Then (under the additional assumption that both the Markov chain and the observation sequence take values in finite sets), the maximum likelihood estimation of the parameters of the hidden Markov model $\mathcal{M}$ is achieved by an iterative procedure involving *reestimation formulas* [1,5], which are obtained from the explicit maximization of an auxiliary function $\mathcal{Q}(\mathcal{M}, \mathcal{M}')$.

Consider now the parametric model described above. It is possible to turn it into a parametric hidden Markov model $\mathcal{M}_\theta = (p_0^\theta, \Pi_\theta, B_\theta)$ with

$$B_\theta(x,y) = (2\pi)^{-\frac{d}{2}} (\det r)^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}[h_\theta(x) - y]^* r^{-1} [h_\theta(x) - y] \Delta t\right\} .$$

In particular $B_\theta(x, y_n) \propto \Psi_n^\theta(x)$. Then it is easily seen that the auxiliary function defined in (33) is such that $\overline{Q}(\theta, \theta') = \mathcal{Q}(\mathcal{M}_\theta, \mathcal{M}_{\theta'})$. Moreover, equations (34) and (39) – which are known

as Baum's forward and backward equations [1,5] – play a central role in the theory of statistical estimation of hidden Markov models.

6 Numerical example

The continouous–time model is described by

$$dX_t = -\theta_2 X_t\,dt + \theta_3 \frac{X_t}{1+X_t^2}\,dt + a^{1/2}\,dW_t\ , \qquad X_0 \sim \mathcal{N}(\theta_1, \Sigma)\ , \tag{40}$$

$$dY_t = \theta_4 \arctan(\frac{X_t}{\theta_4})\,dt + r^{1/2}\,d\overline{W}_t\ , \tag{41}$$

and the unkown parameter is $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$. The noises covariances in the problem are Σ, a and r, and can be associated with the parameters θ_1, (θ_2, θ_3) and θ_4 respectively.

Although the unknown parameter is actually four–dimensional, results will be presented for the estimation of one component of θ at a time, and the influence of the "associated" noise covariance will be investigated.

For each of the cases presented below, the log–likelihood function has been maximized in order to find the MLE, either using the direct approach or the EM algorithm based on nonlinear smoothing. To achieve the direct maximization, one can rely on existing minimization routines from a scientific library, e.g. `e04jbf` from NAG which uses a quasi–Newton algorithm and does not require the user to provide a routine for the computation of the gradient. On the other hand, the M–step of the EM algorithm can either

- be solved explicitely when applicable, e.g. when the auxiliary function depends quadratically on the parameter to be estimated,
- rely on routines from a scientific library.

Two figures are given for each of the cases considered. On the first figure, the following objects can be found

- *in solid line:* the log–likelihood function $\overline{L}(\cdot)$ vs. the free parameter,
- *in dashed line:* iterations of the quasi–Newton algorithm for the direct maximization of the log–likelihood function $\overline{L}(\cdot)$, i.e. straight lines connecting successive points

  $$A_0, A_1, \cdots, A_n, \cdots\ ,$$

  defined by

  $$A_n \triangleq (\widehat{\theta}_n, \overline{L}(\widehat{\theta}_n))\ .$$

On the second figure, the following objects can be found

- *in solid line:* the log–likelihood function $\overline{L}(\cdot)$ vs. the free parameter,
- *in dotted lines:* the auxiliary functions corresponding to successive estimates, i.e. functions $\overline{L}_n(\cdot) \triangleq \overline{Q}(\cdot, \widehat{\theta}_n) + \overline{L}(\widehat{\theta}_n)$, vs. the free parameter,

- *in dashed lines:* iterations of the EM algorithm, i.e. straight lines connecting successive points $A_0, B_0, A_1, B_1, \cdots, A_n, B_n, \cdots$ defined by

$$A_n \triangleq (\widehat{\theta}_n, \overline{L}(\widehat{\theta}_n)) ,$$
$$B_n \triangleq (\widehat{\theta}_{n+1}, \overline{L}_n(\widehat{\theta}_{n+1})) .$$

Remark. In the example introduced above, although the auxiliary function $Q(\theta, \theta')$ of the continuous–time model depends quadratically on the parameters θ_1, θ_2 and θ_3, the discrete–time approximation $\overline{Q}(\theta, \theta')$ depends quadratically on θ_1 only. This can be seen on the expression of the operator $\kappa_{\theta,\theta'}$ – see (37).

Description of cases study

In all these cases, the "true" value of the parameter – i.e. the value used for simulating sample paths of the observation process – is $(\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (1.0, 0.25, 5.0, 2.0)$.

The time interval is $[0, T]$ with $T = 10.0$ and time–step $\Delta t = 0.1$. Observation process sample paths are simulated in the following way. First, simple Euler time–discretization scheme (equivalent on this particular example to Milshtein scheme) is used to simulate the signal process (40)

$$x_{n+1} = x_n + [-\theta_2 x_n + \theta_3 \frac{x_n}{1 + x_n^2}]\Delta t + w_n ,$$

with $x_0 \sim \mathcal{N}(\theta_1, \Sigma)$ and $(w_n : 0 \leq n \leq N)$ a Gaussian white noise sequence with covariance matrix $a\Delta t$. Next, discrete measurements are generated by

$$y_n = \theta_4 \arctan(\frac{x_n}{\theta_4}) + \overline{w}_n ,$$

with $(\overline{w}_n : 0 \leq n \leq N)$ a Gaussian white noise sequence with matrix covariance $r\Delta t^{-1}$, independent of $(w_n : 0 \leq n \leq N)$.

These discrete measurements are used to solve equations (34) and (39), and therefore to compute the approximations $\overline{L}(\theta)$ and $\overline{Q}(\theta, \theta')$ defined by (32) and (33) respectively.

- *Estimation of* θ_1

Fixed parameters: $(\theta_2, \theta_3, \theta_4) = (\theta_2^*, \theta_3^*, \theta_4^*)$.
Noises variances: $a = 1.0$, $r = 1.0$, and $\Sigma = 1.0$ (Case I – fig. 1 and 2) or $\Sigma = 0.01$ (Case II – fig. 3 and 4).
In Case I the EM algorithm has converged after 11 iterations, whereas in Case II it has not converged after 200 iterations. Therefore, only the 12 first iterations are shown on fig. 4.

- *Estimation of* θ_3

Fixed parameters: $(\theta_1, \theta_2, \theta_4) = (\theta_1^*, \theta_2^*, \theta_4^*)$.
Noises variances: $\Sigma = 1.0$, $r = 1.0$, and $a = 1.0$ (Case III – fig. 5 and 6) or $a = 0.01$ (Case IV –

fig. 7 and 8).
In Case III the EM algorithm has converged after 5 iterations, whereas in Case IV it has not converged after 200 iterations. Therefore, only the 12 first iterations are shown on fig. 8.

- *Estimation of* θ_4

Fixed parameters: $(\theta_1, \theta_2, \theta_3) = (\theta_1^*, \theta_2^*, \theta_3^*)$.
Noises variances: $\Sigma = 1.0$, $a = 1.0$, and $r = 1.0$ (Case V – fig. 9 and 10) or $r = 0.01$ (Case VI – fig. 11 and 12).
In Case V the EM algorithm has converged after 9 iterations, whereas in Case VI it has converged after 27 iterations.

The reason why the EM algorithm is so slowly convergent when noise covariances are small – Case II, IV and VI – is that the log–likelihood function is then approximated from below by a set of very sharp auxiliary functions: this situation does not allow to update significantly enough the current estimate at each M–step. Actually, this can be seen directly from (6), (7) – or equivalently from (28), (29). Assume for instance that both $p_0^\theta(\cdot)$ and $b_\theta(\cdot)$ are independent of θ, and that the observation noise covariance r is small. Then every auxiliary function $Q(\cdot, \theta')$ will certainly be very sharp. It should be stressed that in such cases, the slow variation of the estimate should not be interpreted as an indication that the algorithm has already achieved convergence, as one would possibly conclude.

7 Conclusion

The direct maximization of the log–likelihood function has been compared with the EM algorithm, for the MLE of parameters in partially observed diffusion processes. Some formulas given in [2] have been clarified, and it has been shown that smoothing is necesary to make the EM algorithm approach efficient. On the other hand, formulas have been given in terms of filtering stochastic PDE's for the computation of the original log–likelihood function and its gradient.

It has been shown that

- **[E]** the E–step in the EM algorithm is certainly slower than the direct computation of the log–likelihood function, since it involves nonlinear smoothing instead of nonlinear filtering.
- **[M]** the computation of the auxiliary function $Q(\theta, \theta')$ in the M–step of the EM algorithm, θ' being fixed, requires *(i)* at each time t, to integrate some functions depending on (θ, θ') against a normalized smoothing density depending only on θ', and *(ii)* to integrate the resulting processes over the interval $[0, T]$. This gives another evidence that the EM algorithm is more complicated than the direct approach as far as computations are concerned. On the other hand, the maximization of the auxiliary function is generally simple to deal with.
- **[EM]** the EM algorithm converges very slowly whenever some noise covariances associated with the parameters to be estimated are small.

However, the EM algorithm should provide an interesting approach for non–parametric estimation in the context of partially observed diffusion processes, i.e. non–parametric estimation of the initial density, the drift and the observation function. This form of the EM algorithm is used indeed in the context of finite–space Markov chains with finite–state observations (hidden Markov models), and leads to well–known reestimation formulas, which are of practical use e.g. in acoustic speech recognition.

References

[1] L.E. BAUM, An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, in: *Inequalities III (Los Angeles-1969)* (ed. O.Shisha) 1–8, Academic Press (1971).

[2] A. DEMBO and O. ZEITOUNI, Parameter estimation of partially observed continuous-time stochastic processes via the EM algorithm, *Stochastic Processes and Applications* **23** (1) 91–113 (1986).

[3] A.P. DEMPSTER, N.M. LAIRD and D.B. RUBIN, Maximum likelihood estimation from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B* **39** (1) 1–38 (1977).

[4] F. LE GLAND, *Estimation de paramètres dans les processus stochastiques en observation incomplète,* Thèse de Docteur–Ingénieur, Université de Paris IX–Dauphine (1981).

[5] S.E. LEVINSON, L.R. RABINER and M.M. SONDHI, An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition, *Bell System Technical Journal* **62** (4) 1035–1074 (1983).

[6] D. NUALART and E. PARDOUX, Stochastic calculus with anticipating integrands, *to appear in Probability Theory and Related Fields.*

[7] D. OCONE, Stochastic calculus of variations for stochastic partial differential equations, *to appear in Journal of Functional Analysis.*

[8] E. PARDOUX, Stochastic PDEs and filtering of diffusion processes, *Stochastics* **3** (2) 127–167 (1979).

[9] E. PARDOUX, Equations du lissage non-linéaire, in: *Filtering and Control of Random Processes (Paris-1983)* (eds. H.Korezlioglu, G.Mazziotto and J.Szpirglas) 206–218, Springer–Verlag (LNCIS-61) (1984).

[10] E. PARDOUX and Ph. PROTTER, A two-sided stochastic integral and its calculus, *Probability Theory and Related Fields* **76** (1) 15–49 (1987).

[11] E. PARDOUX, Two-sided stochastic calculus for SPDEs, in: *Stochastic PDEs and Applications (Trento-1985)* (eds. G.DaPrato and L.Tubaro) 200–207, Springer–Verlag (LNM-1236) (1987).

[12] A.S. SZNITMAN, Martingales dépendant d'un paramètre: une formule d'Itô, *Z. Wahrscheinlichkeitstheorie* **60** (1) 41–70 (1982).

[13] C.F.J. WU, On the convergence properties of the EM algorithm, *Annals of Statistics* **11** (1) 95–103 (1983).

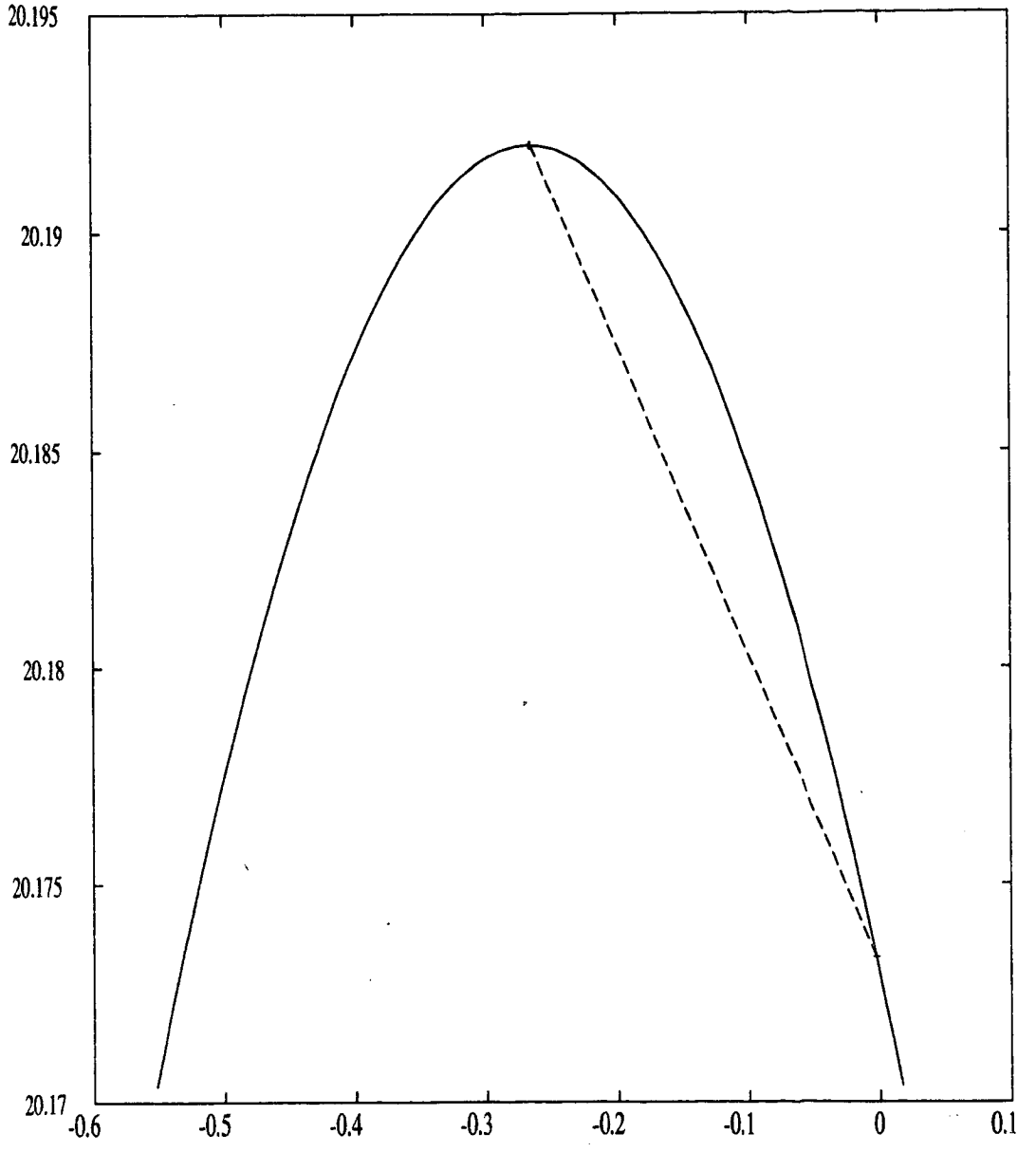

Figure 1: Case I – Direct maximization

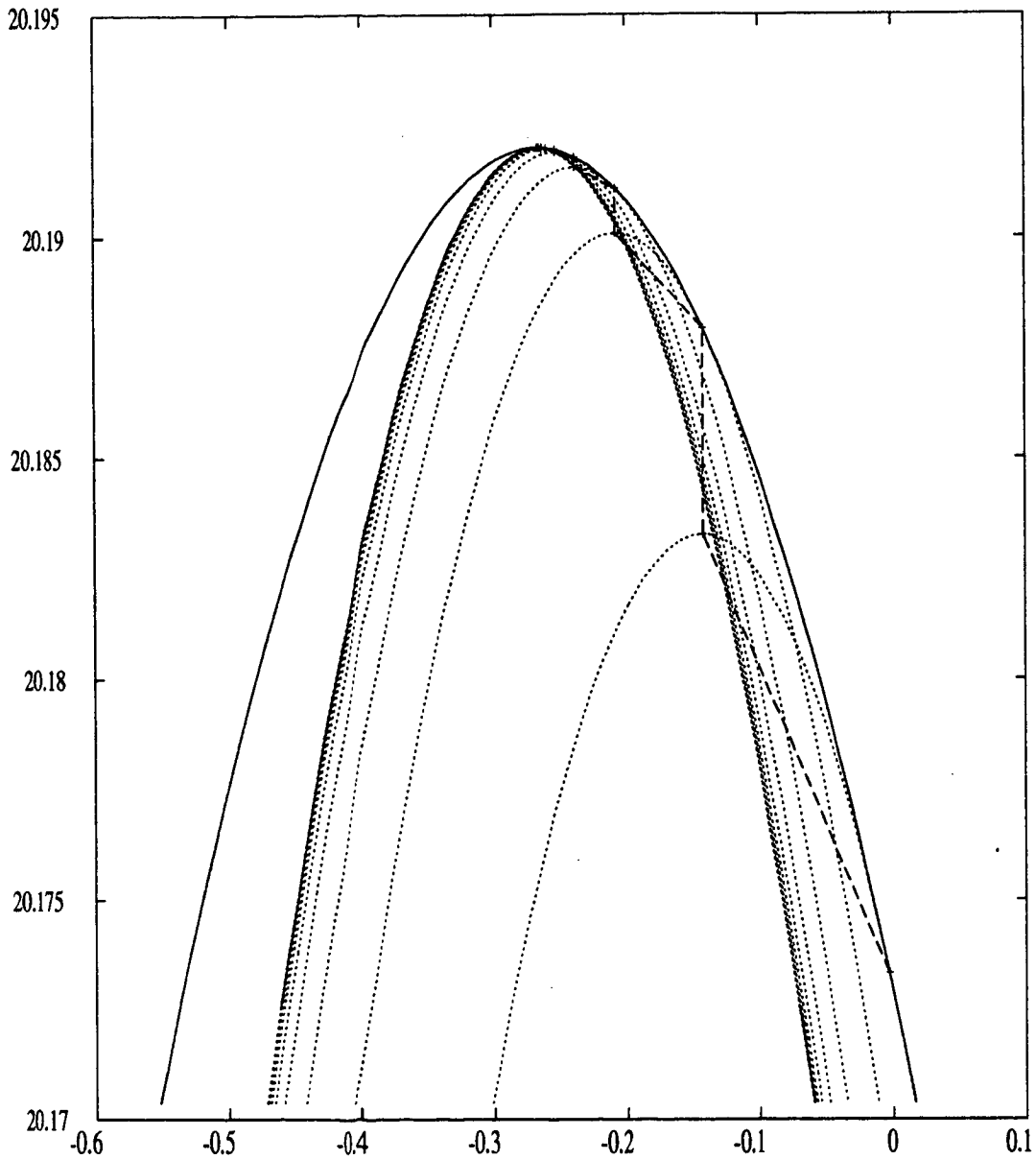

Figure 2: Case I – EM algorithm

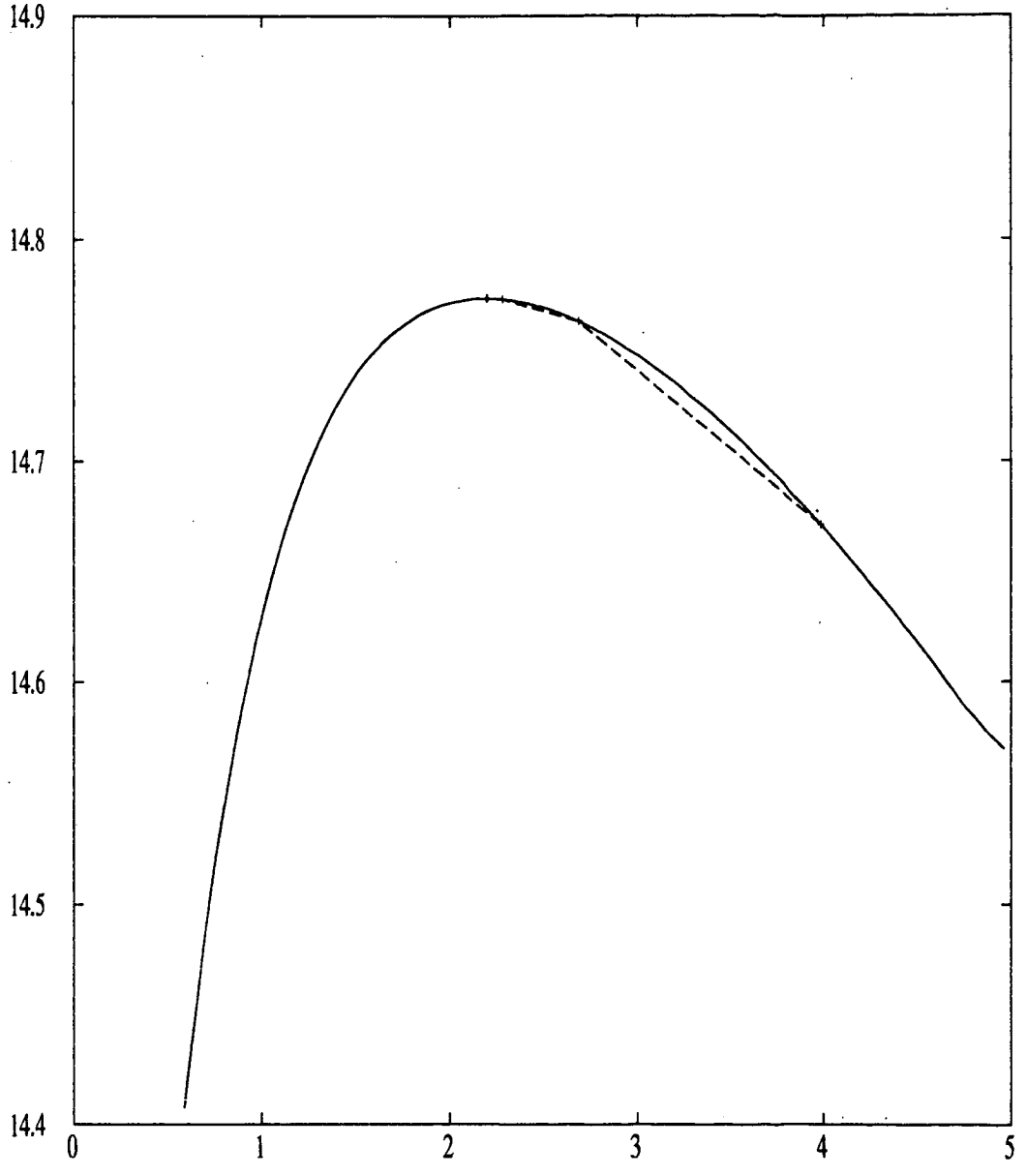

Figure 3: Case II – Direct maximization

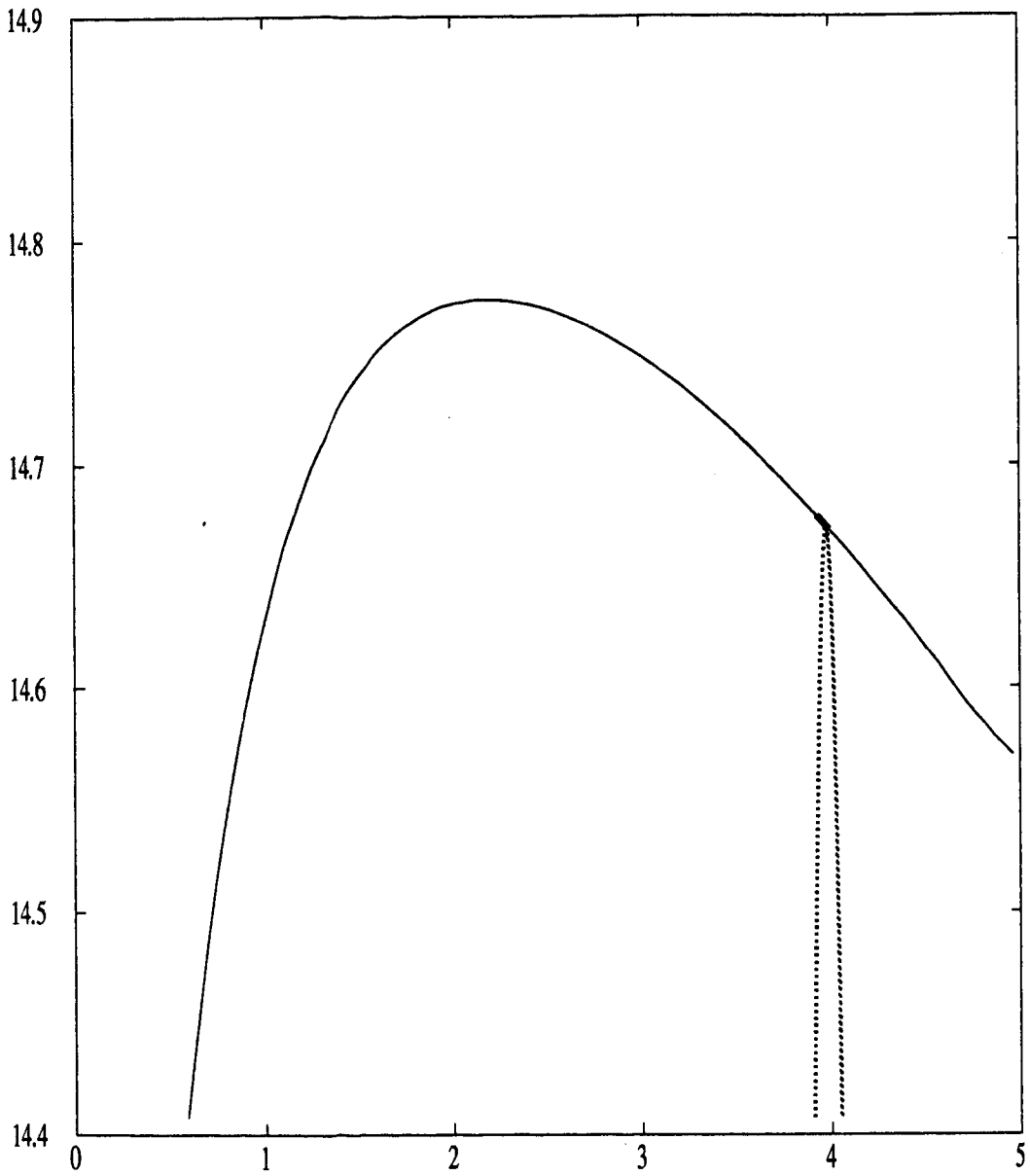

Figure 4: Case II – EM algorithm

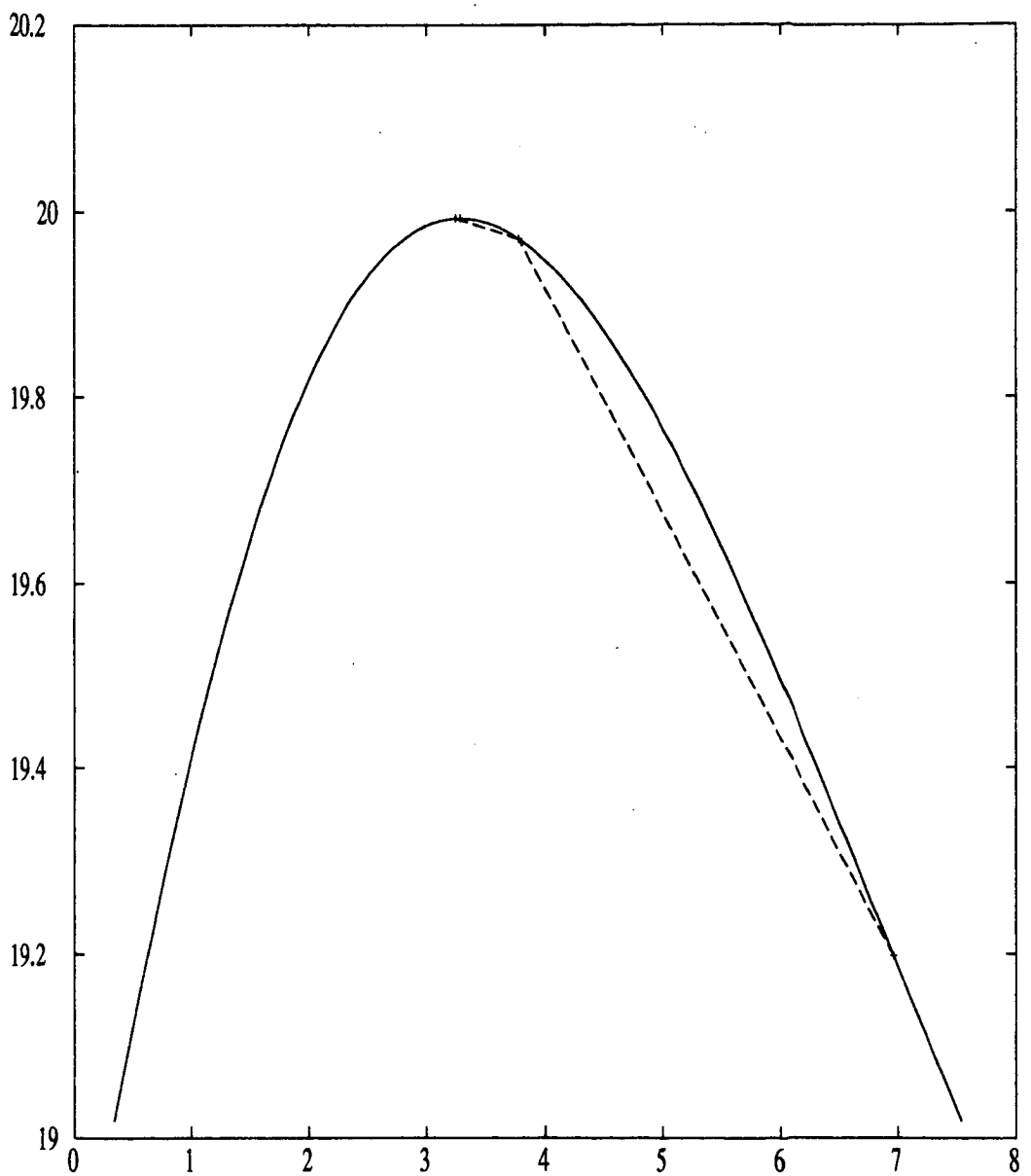

Figure 5: Case III – Direct maximization

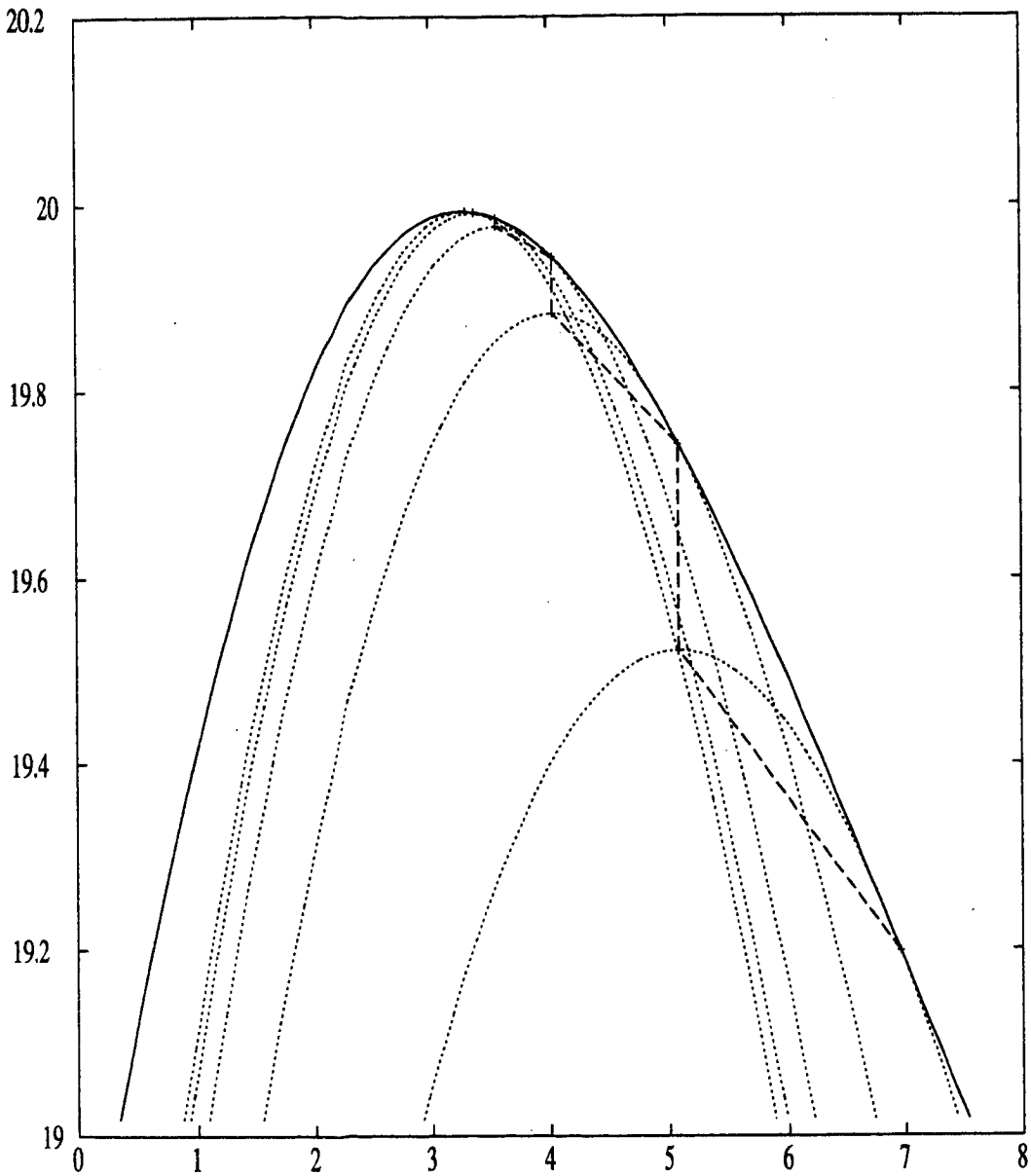

Figure 6: Case III – EM algorithm

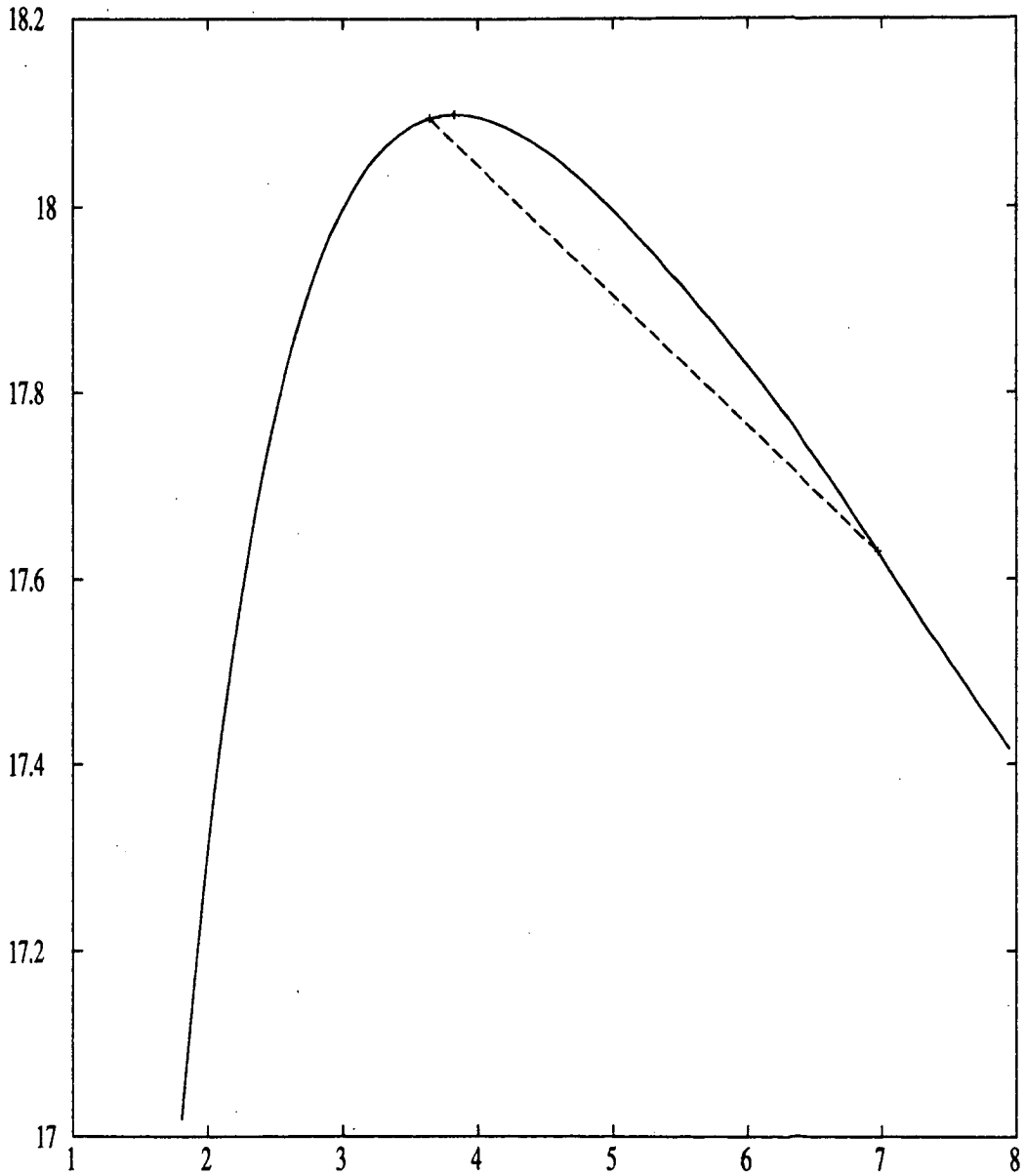

Figure 7: Case IV – Direct maximization

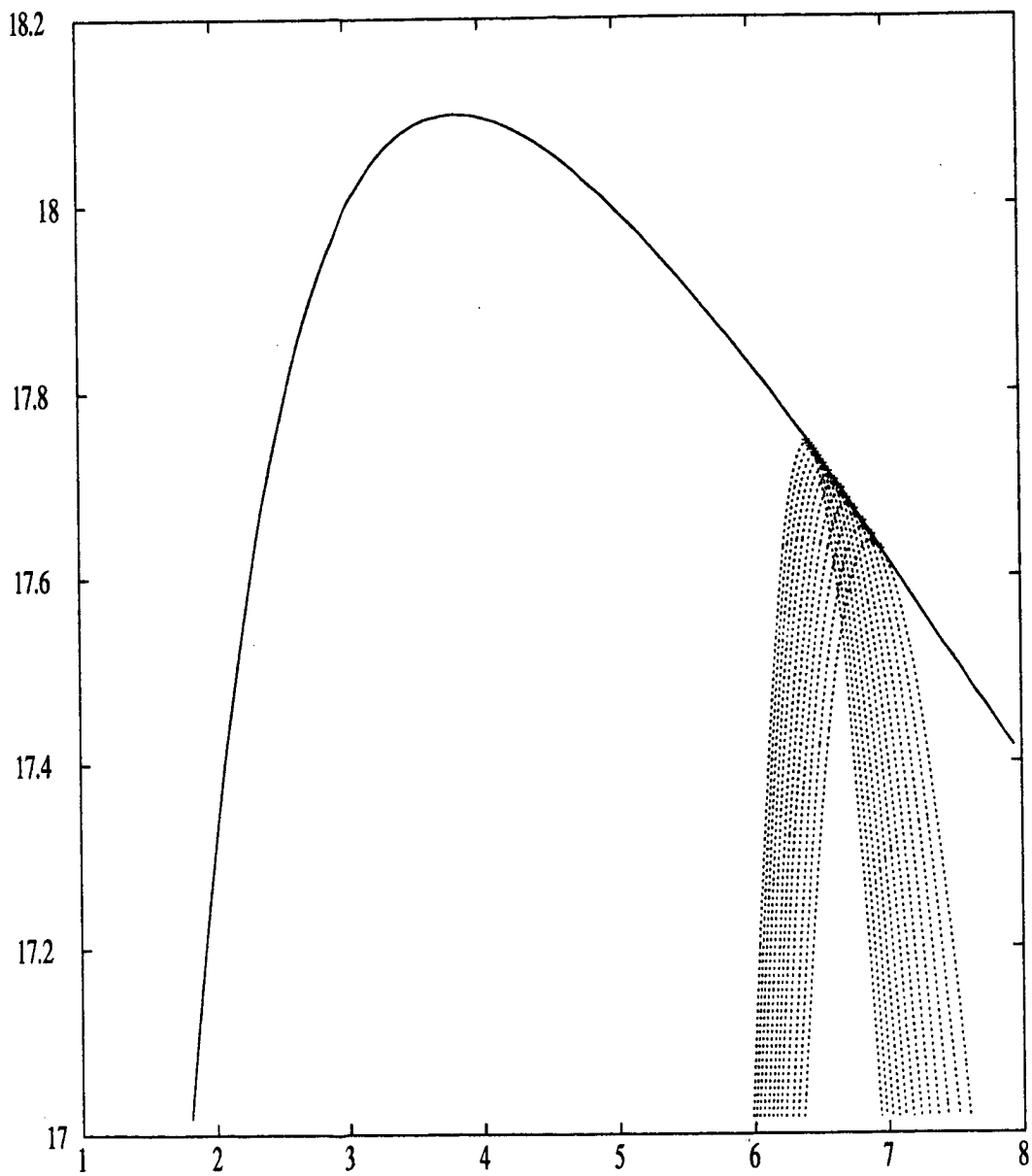

Figure 8: Case IV – EM algorithm

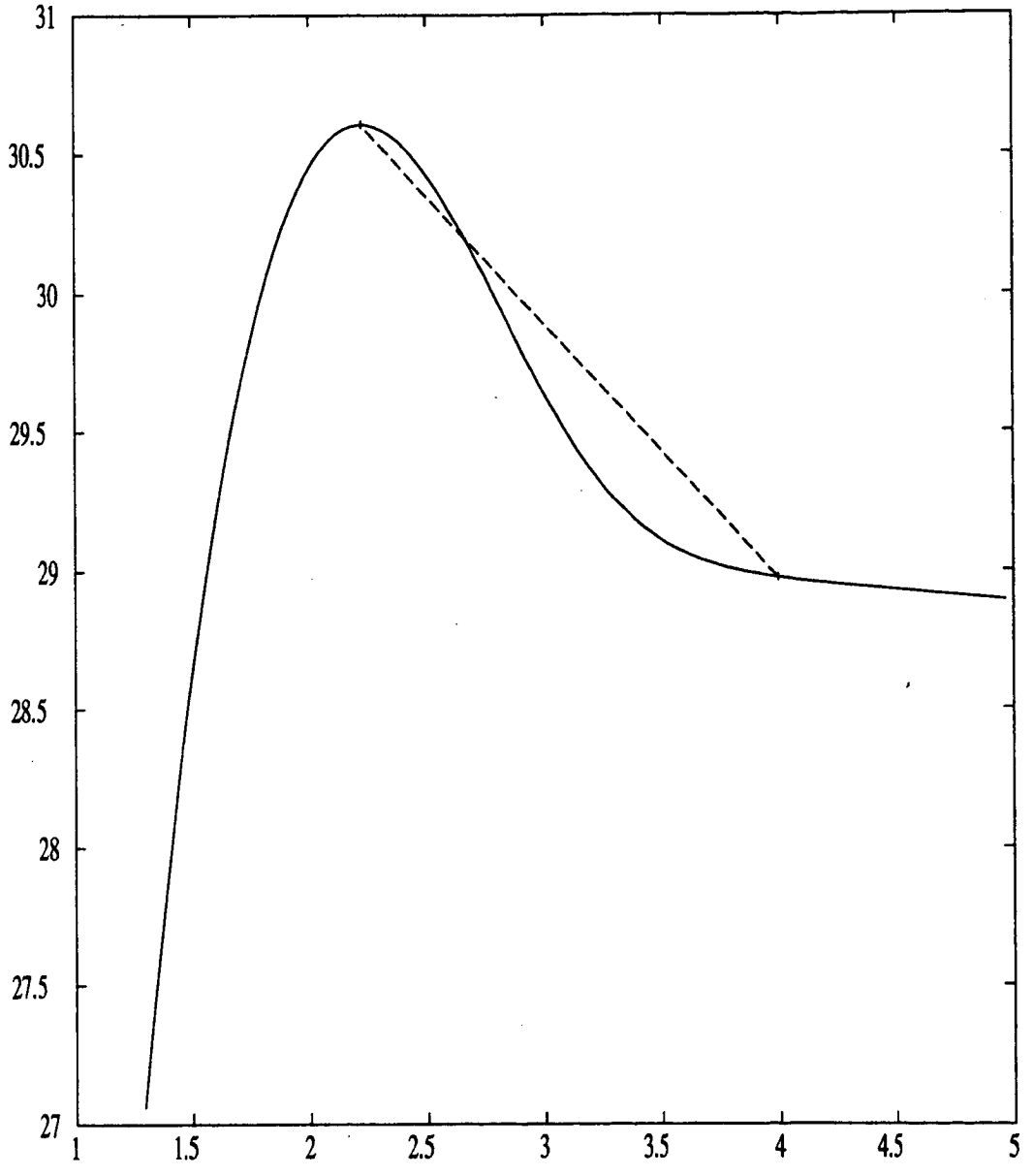

Figure 9: Case V – Direct maximization

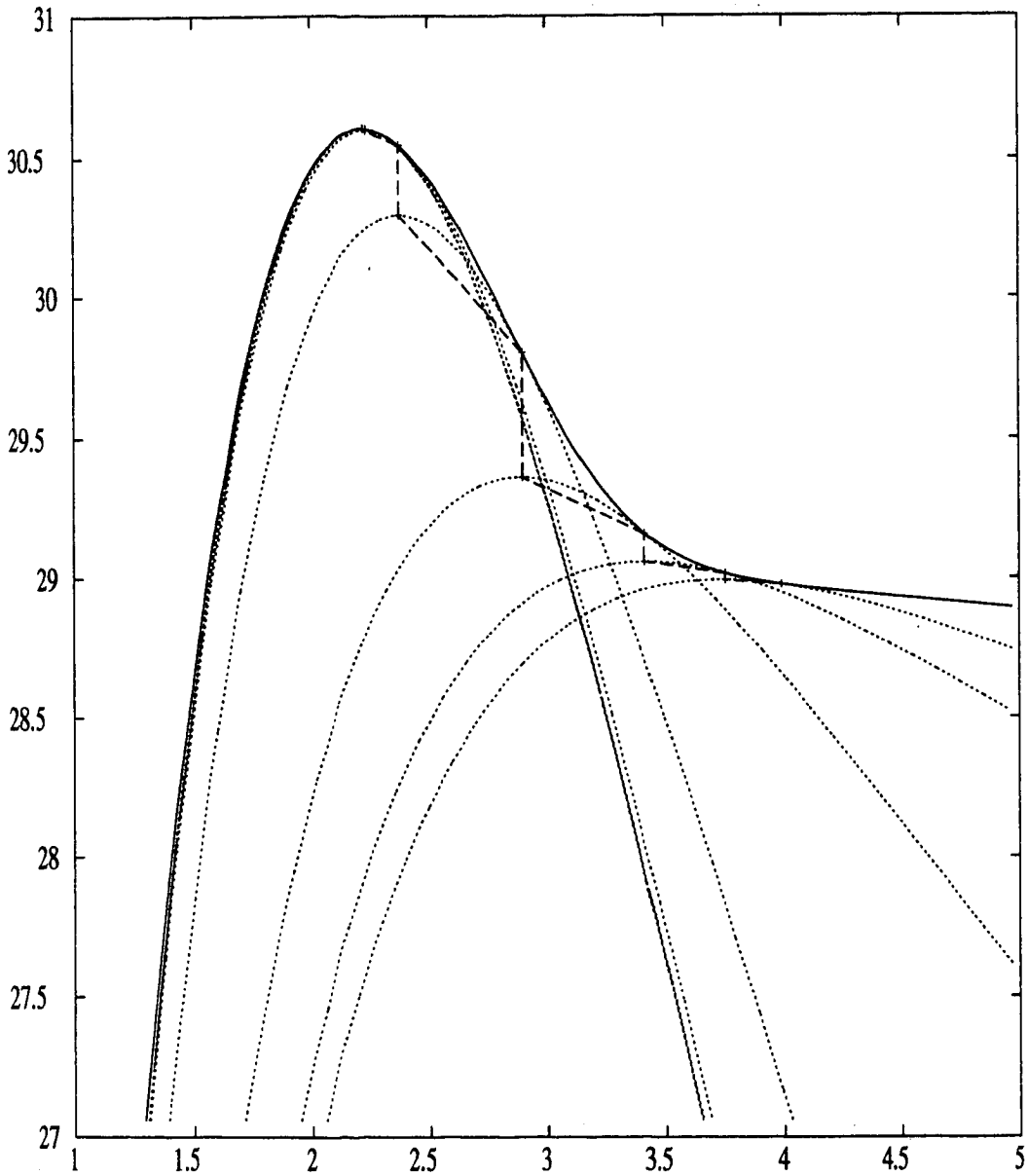

Figure 10: Case V – EM algorithm

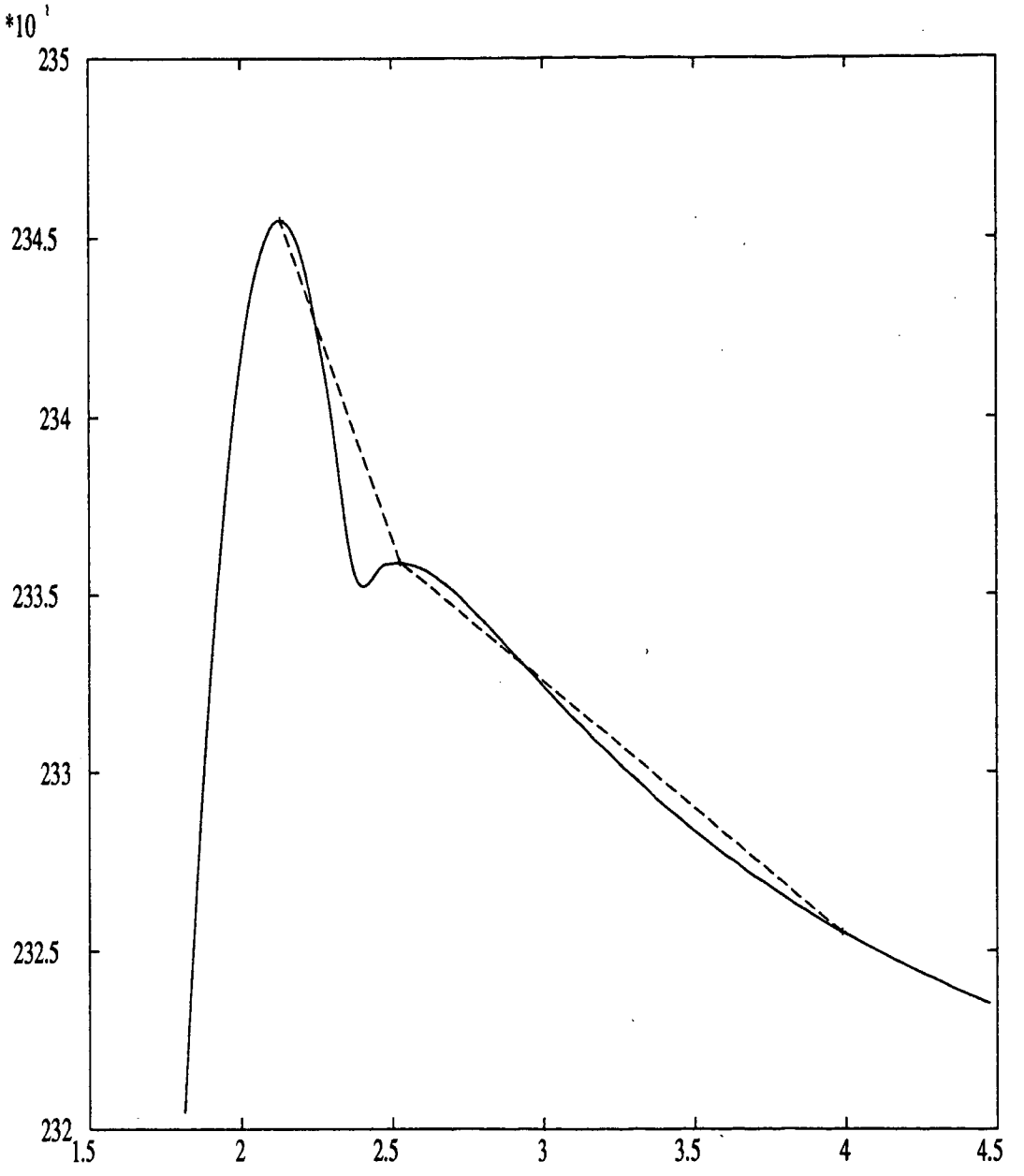

Figure 11: Case VI – Direct maximization

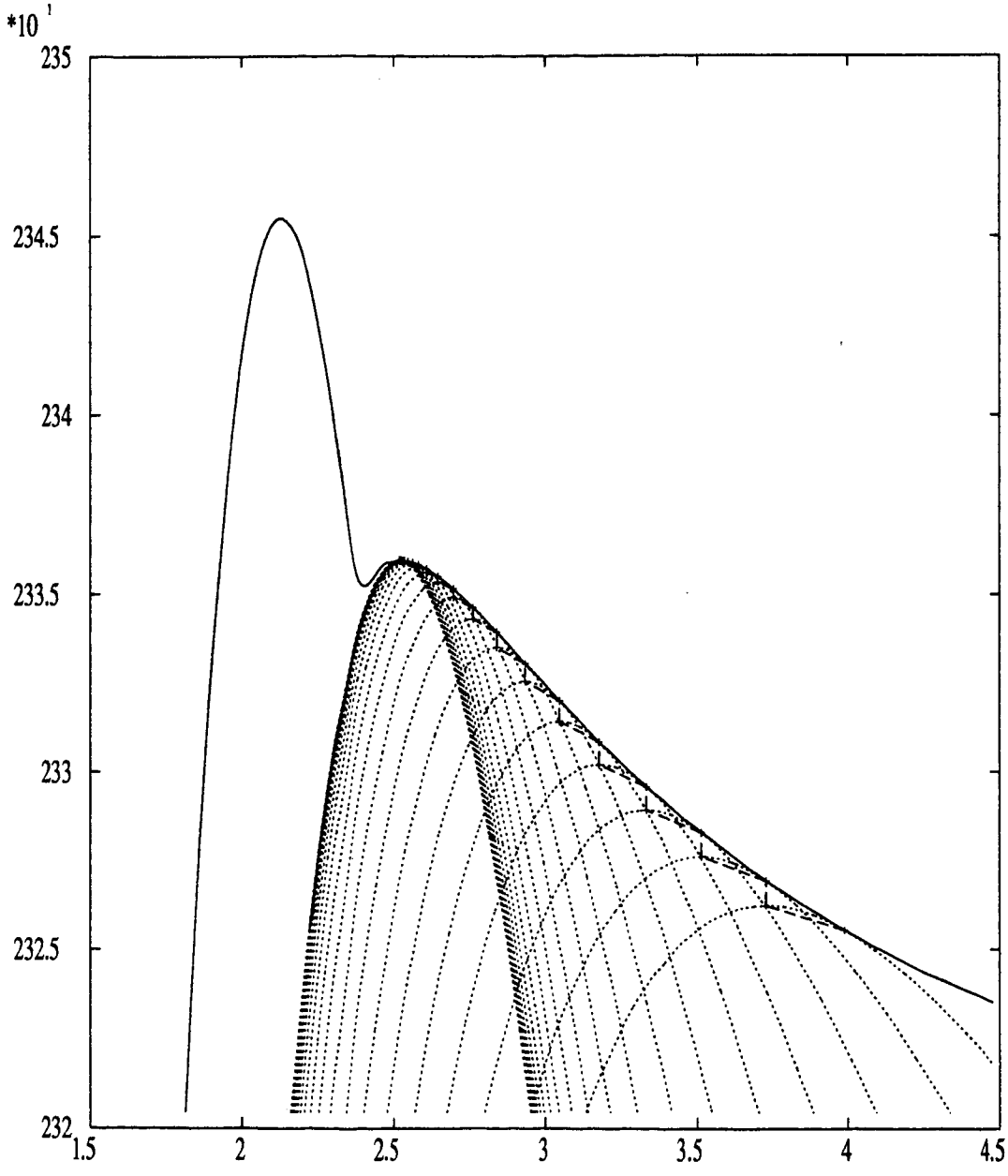

Figure 12: Case VI – EM algorithm

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique