Unsupervised Deconvolution of Sparse Spike Trains Using Stochastic Approximation

Frédéric Champagnat, Yves Goussard, *Member, IEEE,* and Jérôme Idier

***Abstract*— This paper presents an unsupervised method for restoration of sparse spike trains. These signals are modeled as random Bernoulli–Gaussian processes, and their unsupervised restoration requires (i) estimation of the hyperparameters that control the stochastic models of the input and noise signals and (ii) deconvolution of the pulse process. Classically, the problem is solved iteratively using a maximum generalized likelihood approach despite questionable statistical properties.**

The contribution of the article is threefold. First, we present a new "core algorithm" for supervised deconvolution of spike trains, which exhibits enhanced numerical efficiency and reduced memory requirements. Second, we propose an original implementation of a hyperparameter estimation procedure that is based upon a stochastic version of the expectation-maximization (EM) algorithm. This procedure utilizes the same core algorithm as the supervised deconvolution method. Third, Monte Carlo simulations show that the proposed unsupervised restoration method exhibits satisfactory theoretical and practical behaviors and that, in addition, good global numerical efficiency is achieved.

I. Introduction

This paper deals with restoration of sparse spike trains distorted by a linear system and corrupted by additive white noise. Such a problem occurs in a variety of areas such as geophysics, communications, medical imaging, nondestructive evaluation (NDE), radar, etc. In order to obtain an adequate solution, the spiky nature of the unknown signal must be accounted for. In a stochastic framework, this can be done by modeling the pulse train as a Bernoulli–Gaussian (BG) process and by using a Bayesian estimator for the restoration. This approach produced interesting results and several recursive [1]–[4] and iterative [5]–[8] methods are now available.

However, as is often in Bayesian estimation, the parameters of the probability distributions which control the problem, also referred to as the *hyperparameters*, must be specified somehow. In addition, in many applications, the linear system that distorts the pulse train is not known precisely and must also be estimated. As far as the latter problem is concerned, several techniques based upon higherorder statistics (see, e.g., [9]–[11]) or distance between probability distributions [12]–[15] have been developed in the past few years; they provide an adequate answer, as long as the data sample is large enough.

On the other hand, the problem of hyperparameter estimation is not solved appropriately at the moment. In this paper, attention is focused on this problem, while the linear system is assumed to be known. In the sequel, signal and hyperparameter estimation will be referred to as *unsupervised* deconvolution. A commonly used technique consists of jointly estimating the spike train, the hyperparameters, and possibly the linear system through maximization of a single *generalized likelihood* (GL) defined as the probability distribution of all stochastic quantities conditionally on all deterministic parameters. From a practical standpoint, this maximum generalized likelihood (MGL) approach is appealing as suboptimal maximization of the generalized likelihood can generally be implemented easily. This explains why it has been applied to several problems such as signal and image processing [16], [17], automatic control [18], [19], and pattern recognition [20]. For deconvolution of BG signals, Mendel and coworkers [5], [21], [22] followed by others [23] used the approach to jointly determine the pulse signal, a parametric model of the linear system and the hyperparameters. The resulting methods were easy to implement and produced interesting results.

However, the asymptotic behavior of MGL estimators is very questionable. In general, they do not fulfill the property of consistency [24], [25]. But a more serious and seldom reported issue is that MGL estimates may not exist [24]. This indicates that incautious use of the MGL approach is risky, and that development of well-behaved hyperparameter estimation techniques is highly desirable.

The contribution of this paper is threefold: first, we propose a new *core algorithm* that may be used as the basic element of several iterative BG deconvolution methods. The new algorithm presents a low numerical complexity and small memory requirements, thereby extending the applicability of BG deconvolution to large data samples.

Second, we present a hyperparameter estimation method based upon a true maximum likelihood (ML) estimator, which guarantees a satisfactory theoretical behavior of the estimates. The major difficulty lies in the computation and maximization of the likelihood function. Here, the ML estimator is implemented by means of an expectation-maximization (EM) algorithm. Two factors contribute to achieving fast convergence: (i) adequate choice of the auxiliary variable of the EM algorithm; (ii) utilization of a stochastic version of the EM algorithm (SEM algorithm [26]). In the context of BG

Manuscript received July 13, 1994; revised May 27, 1996. This work was supported, in part, by the Direction des Recherches et Études Techniques under Contract 901501/A000/DRET/DS/SR, and by the Natural Sciences and Engineering Research Council of Canada under Research Grant OGP0138417. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhi Ding.

F. Champagnat and J. Idier are with the Laboratoire des Signaux et Systèmes, École Supérieure d'Électricité, Plateau de Moulon, 91192 Gif-sur-Yvette, Cédex, France (e-mail: champagnat@lss.supelec.fr).

Y. Goussard is with the École Polytechnique, Institut de Génie Biomédical, Montréal, Québec H3C 3A7, Canada.

Publisher Item Identifier S 1053-587X(96)09055-1.

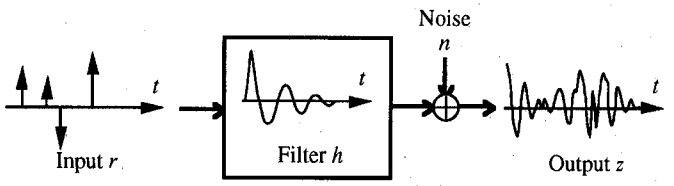


Fig. 1. Schematic representation of the phenomena under study.

deconvolution, the SEM algorithm was formerly proposed by Lavielle [27], [28] but our implementation uses a smaller set of auxiliary variables in order to yield better convergence rates. In addition, implementation of the SEM algorithm makes use of the same core algorithm that is used for restoration of the pulse train. This results in a consistent algorithmic framework and methods with a low numerical complexity for both signal restoration and hyperparameter estimation.

Third, evaluation and comparison of SEM and of an alternative MGL hyperparameter estimation method is carried out using Monte Carlo simulations. This allows us to assess the cost/performance ratio of the different techniques precisely and to clarify their respective ranges of application.

The paper is organized as follows. A mathematical formulation of the problem is given in Section II. Then, the core algorithm and the corresponding BG signal restoration methods are presented in Section III. Section IV contains background results on EM and SEM algorithms. Section V is devoted to hyperparameter estimation using the SEM algorithm. The Monte Carlo simulation results are presented in Section VI and the conclusions of this study are drawn in Section VII.

II. Problem Formulation

A schematic representation of the phenomena under study is given in Fig. 1. The observed signal is the noise-corrupted convolution product of the unknown pulse signal and of the impulse response of the linear system. As all signals are assumed to be discrete-time, the input-output equation can be written in the following matrix form:

$$\mathbf{z} = H\mathbf{r} + \mathbf{n} \tag{1}$$

where vectors $\mathbf{z}$, $\mathbf{r}$ and $\mathbf{n}$ contain the samples of the observed signal, of the input signal and of the observation noise, respectively. $(P \times N)$ matrix H is made up of shifted samples of the impulse response of the linear system.

Here, the observation noise is assumed to be a zero-mean Gaussian process with variance r_n. The unknown pulse train is modeled as a BG process, which is made up of two parts: An *unobserved* part Q, that controls the occurence of a pulse and an *observed* part R that represents the amplitude of the pulse process at each time-sample. Q is an independent identically distributed (i.i.d.) Bernoulli process with parameter λ, and R is a white zero-mean Gaussian process with variance Qr_x. Therefore, a BG process X can be formally defined as a sequence of independent random variables (RV's) X_k such that

$$\begin{aligned} &X_k \triangleq (Q_k, R_k) \\ &Q_k\text{: binary RV } \begin{cases} \Pr\{q_k = 1\} = \lambda \\ \Pr\{q_k = 0\} = 1 - \lambda \end{cases} \\ &R_k\text{: zero-mean Gaussian RV with variance } q_k r_x. \end{aligned} \tag{2}$$

The above definition and the assumption on the observation noise show that all probability distributions associated with the problem are controlled by $\boldsymbol{\theta} = \{\lambda, r_x, r_n\}$; these three quantities represent the hyperparameters of the deconvolution problem.

Let us first consider supervised deconvolution. As shown by (2), restoration of the pulse process requires two operations: *detection* of position variables Q_k and *estimation* of amplitude variables R_k. Straightforward derivation of a MAP estimator of $\mathbf{x} = \{\mathbf{q}, \mathbf{r}\}$ yields the maximization of the following *joint* likelihood:

$$p(\mathbf{q}, \mathbf{r} \mid \mathbf{z}; \boldsymbol{\theta}) \propto p(\mathbf{z} \mid \mathbf{r}; \boldsymbol{\theta}) p(\mathbf{r} \mid \mathbf{q}; \boldsymbol{\theta}) \Pr(\mathbf{q}; \boldsymbol{\theta}) \tag{3}$$

where vectors $\mathbf{q}$ and $\mathbf{r}$ contain the samples of the position and amplitude variables, respectively. This approach, in which detection and estimation are carried out jointly, has been reported to produce a high number of false detections [21]–[23]. It is preferable to use a sequential approach where detection is performed first through maximization of the posterior marginal likelihood of $\mathbf{q}$

$$\Pr(\mathbf{q} \mid \mathbf{z}; \boldsymbol{\theta}) \propto p(\mathbf{z} \mid \mathbf{q}; \boldsymbol{\theta}) \Pr(\mathbf{q}; \boldsymbol{\theta}). \tag{4}$$

After the estimate $\hat{\mathbf{q}}$ of the Bernoulli sequence has been obtained, the amplitude variables are determined through maximization of the posterior likelihood of $\mathbf{r}$ conditionally to $\mathbf{z}$ and $\mathbf{q}$

$$p(\mathbf{r} \mid \mathbf{z}, \hat{\mathbf{q}}; \boldsymbol{\theta}) \propto p(\mathbf{z} \mid \mathbf{r}; \boldsymbol{\theta}) p(\mathbf{r} \mid \hat{\mathbf{q}}; \boldsymbol{\theta}). \tag{5}$$

As input-output (1) is linear, and since the conditional distributions of $\mathbf{n}$ and $\mathbf{r}$ are Gaussian, determination of $\mathbf{r}$ is a classical MAP estimation problem under linear and Gaussian assumptions. The solution can be expressed in closed-form and is classically given by

$$\hat{\mathbf{r}} = \Pi H' B^{-1} \mathbf{z}, \tag{6}$$

$$B \triangleq H \Pi H' + r_n I \tag{7}$$

where Π denotes the covariance matrix of the prior distribution $p(\mathbf{r} \mid \mathbf{q}; \boldsymbol{\theta})$. From (2), we get

$$\Pi = r_x Q \quad \text{with} \quad Q \triangleq \operatorname{diag}\{q_k\}. \tag{8}$$

The expression of the detection criterion can be derived easily. Using (1), (2), and the Gaussian distribution of the observation noise, (4) can be written as

$$\begin{aligned} J_{\boldsymbol{\theta}}(\mathbf{q}) &\triangleq p(\mathbf{z} \mid \mathbf{q}; \boldsymbol{\theta}) \Pr(\mathbf{q}; \boldsymbol{\theta}) \\ &= \frac{1}{(2\pi)^{P/2} \sqrt{|B|}} \exp\left\{ -\frac{\mathbf{z}' B^{-1} \mathbf{z}}{2} \right\} \lambda^{N_e} (1 - \lambda)^{N - N_e} \end{aligned} \tag{9}$$

where N_e denote the number of nonzero samples of $\mathbf{q}$. The major difficulty lies in the maximization of $J_{\boldsymbol{\theta}}(\mathbf{q})$, because

of the discrete nature of the Bernoulli variable. Exact maximization of J_θ would require (9) to be evaluated for all possible configurations of $\mathbf{q}$, and such a task is intractable for signals of realistic sizes. In practice, only a small part of all possible configurations is explored. Thus, the efficiency of the detection algorithm is a function of the proportion of explored configurations and of the exploration strategy. These points are discussed in Section III.

Let us now turn to unsupervised deconvolution. Here too, the discrete nature of the Bernoulli sequence generates difficulties. The ML estimate of θ is defined by

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \triangleq \arg\max_{\theta\in\Theta} p(\mathbf{z};\boldsymbol{\theta}), \quad \text{where } \Theta = [0,1]\times[0,+\infty)^2. \tag{10}$$

The likelihood function $p(\mathbf{z};\boldsymbol{\theta})$ cannot be evaluated directly, but must be computed through projection of joint conditional distribution $p(\mathbf{z},\mathbf{x};\boldsymbol{\theta})$. Using (2) and Bayes rule, we obtain

$$\begin{aligned} p(\mathbf{z};\boldsymbol{\theta}) &= \int_x p(\mathbf{z},\mathbf{x};\boldsymbol{\theta})\,d\mathbf{x} \\ &= \sum_q \int_{\mathbf{r}} p(\mathbf{z},\mathbf{r}\mid \mathbf{q};\boldsymbol{\theta})\mathrm{Pr}(\mathbf{q};\boldsymbol{\theta})d\mathbf{r} \\ &= \sum_q p(\mathbf{z}\mid \mathbf{q};\boldsymbol{\theta})\mathrm{Pr}(\mathbf{q};\boldsymbol{\theta}). \end{aligned} \tag{11}$$

Equation (11) shows that for a given value of the hyperparameters, evaluation of the likelihood requires the summation with respect to (w.r.t.) $\mathbf{q}$ of a function, which is identical to the detection criterion given in (4). Therefore, direct evaluation of the likelihood—not to mention its maximization—cannot be carried out in practice. This difficulty will be overcome through derivation of a stochastic extension to the EM algorithm, as explained in Sections IV and V.

III. Supervised Deconvolution

A. General Considerations

Maximization of detection criterion (9) is performed along lines similar to those presented in [5], [6], and [23]: These techniques consist of defining the notion of *neighboring sequences*, and of maximizing the detection criterion along a series of neighboring sequences. The overall efficiency of such a scheme relies on three main factors: (i) the nature of the neighborhoods; (ii) the exploration strategy of the current neighborhood; and (iii) the availability of a numerically efficient *core algorithm*, i.e., a set of formulas relating the criteria of two neighboring sequences. When the exploration strategy of the current neighborhood always results in increasing the criterion at each step, the algorithm converges in a finite (hopefully small) number of iterations, as $\mathbf{q}$ spans a finite set. Practically, such a scheme may provide interesting results only when the criterion is *well-behaved*, i.e., when a local optimum close to the global one can be reached from any initial point through a series of neighboring sequences with increasing criterion values.

Kormylo and Mendel's classical *single most likely replacement* (SMLR) algorithm [21], [5] can be interpreted in this framework as follows: A neighborhood is made up of all sequences that differ from the current one by one sample, and the selection strategy consists of choosing as the next current sequence the one that maximizes the criterion over the entire neighborhood of the current sequence. This strategy exhibits a satisfactory practical behavior [22]; moreover, when associated with a finite impulse response (FIR) model of the filter (whereas Mendel and coworkers used an ARMA model), it yields a very simple algorithmic structure [7] and avoids the use of a *realization* procedure [21] to identify the ARMA parameters before running the deconvolution algorithm. Conversely, this algorithm requires an $O(N^2)$ memory load, a potentially penalizing factor in applications like NDE. In Section III-B, we propose an original core algorithm suited for an SMLR exploration strategy. This algorithm presents a reduced memory requirement and is faster in many practical situations. In addition, the core algorithm is at the heart of the Gibbs sampler used in the "S-step" of the unsupervised deconvolution procedure of Section V.

Improved performance of SMLR-type methods may be expected if the proportion of explored configurations is increased. This can be achieved by expanding the size of the neighborhoods, and techniques like SSS and SSS-SMLR detectors [6] rely on this principle. The core algorithm described in Section III-B can be easily extended from first-order neighborhoods to second-order ones defined as follows: Two Bernoulli sequences are (k-order) neighbors when they differ at most at k *consecutive* samples. The performance of a second-order extension and of several selection strategies was investigated in [29]. The conclusions are that the slightly better estimation results produced by second-order methods are not worth extra computational load, and that SMLR provides the best tradeoff between result accuracy and computational cost. Consequently, throughout the rest of the paper, we shall only deal with first-order methods, which will be associated with an SMLR exploration strategy in the case of supervised deconvolution.

B. Core Algorithm

The core algorithm links the criterion values associated with two (first-order) neighboring sequences. We start with the following logarithmic expression for criterion: J

$$\begin{aligned} L(\mathbf{q};\boldsymbol{\theta}) &\triangleq 2\ln J_\theta(\mathbf{q}) + P\ln 2\pi \\ &= -\mathbf{z}'B^{-1}\mathbf{z} - \ln|B| + 2N_e\ln\lambda \\ &\quad + 2(N-N_e)\ln(1-\lambda) \end{aligned} \tag{12}$$

because it lends itself nicely to algebraic manipulations and tends to prevent overflows. Moreover we introduce ratio $\mu \triangleq r_x/r_n$ and normalized matrix $\tilde{B} \triangleq B/r_n$. In the sequel tilded quantities depend on μ only, and the main part of the algorithm can be expressed in terms of those normalized quantities.

Given any initial sequence $\mathbf{q}_0$, let $\mathbf{q}_k$, $k \in [1,N]$ denote the sequence differing from $\mathbf{q}_0$ only at site k. We seek a relationship between $L(\mathbf{q}_k;\boldsymbol{\theta})$ and $L(\mathbf{q}_0;\boldsymbol{\theta})$. Let $\mathbf{v}_k$ be the N-vector whose coordinates are 0 except for the kth one, which is equal to 1. Subscript k (resp. 0) will refer to any quantity related to $\mathbf{q}_k$ (resp. $\mathbf{q}_0$). From (7) we get

$$\tilde{B}_k = \tilde{B}_0 + \varepsilon_k\mu H\mathbf{v}_k\mathbf{v}_k'H' \tag{13}$$

where ε_k takes the value 1 (resp. -1) when a 1 is added to (resp. removed from) sequence $\mathbf{q}_0$. Define auxiliary quantities

$$\mathbf{h}_k \triangleq H\mathbf{v}_k \quad \text{and} \quad \tilde{\rho}_k \triangleq \varepsilon_k + \mu \mathbf{h}_k' \tilde{B}_0^{-1} \mathbf{h}_k. \tag{14}$$

Applying the matrix inversion lemma to (13) yields

$$\tilde{B}_k^{-1} = \tilde{B}_0^{-1} - \mu \tilde{B}_0^{-1} \mathbf{h}_k \tilde{\rho}_k^{-1} \mathbf{h}_k' \tilde{B}_0^{-1} \tag{15}$$

$$\mathbf{z}' \tilde{B}_k^{-1} \mathbf{z} = \mathbf{z}' \tilde{B}_0^{-1} \mathbf{z} - \mu (\mathbf{z}' \tilde{B}_0^{-1} \mathbf{h}_k)^2 \tilde{\rho}_k^{-1}. \tag{16}$$

It may also be shown from (13) (cf. [30]) that

$$|\tilde{B}_k| = |\tilde{B}_0| \varepsilon_k \tilde{\rho}_k. \tag{17}$$

These expressions enable us to compute $L(\mathbf{q}_k; \boldsymbol{\theta})$ knowing $L(\mathbf{q}_0; \boldsymbol{\theta})$, and to update $\tilde{B}_0^{-1}$ for the next iteration [7]. They provide the key to SMLR-like algorithms. Actually, the algorithm derived in [7] makes use of auxiliary quantity $\tilde{A} \triangleq H' \tilde{B}^{-1} H$ instead of $\tilde{B}_0^{-1}$ in order to yield a faster algorithm. The associated numerical cost is $O(N^2)$ multiplications per iteration. The major drawback of this structure is the memory requirement associated with the storage of $N \times N$ matrix $\tilde{A}$.

Here, we propose to use another auxiliary quantity of smaller dimension in order to relieve this drawback: Inspection of (7) reveals that term $H\Pi H'$ is not full rank. More precisely, by defining $G \triangleq HD$ where D denotes the $N \times N_e$ matrix made of the nonzero columns in Q, (7) can be rewritten as

$$\tilde{B} = G\mu G' + I, \quad \text{therefore } \tilde{B}^{-1} = I - \mu G \tilde{C}^{-1} G' \tag{18}$$

where $\tilde{C} \triangleq \mu G' G + I$ is a $N_e \times N_e$ matrix. We are now able to express the right-hand side of (16) in terms of $\tilde{C}_0^{-1}$

$$\mathbf{z}' \tilde{B}_0^{-1} \mathbf{h}_k = \mathbf{z}' \mathbf{h}_k - \mu \mathbf{z}' G_0 \tilde{C}_0^{-1} G_0' \mathbf{h}_k \tag{19}$$

$$\mathbf{h}_k' \tilde{B}_0^{-1} \mathbf{h}_k = \mathbf{h}_k' \mathbf{h}_k - \mu \mathbf{h}_k' G_0 \tilde{C}_0^{-1} G_0' \mathbf{h}_k \tag{20}$$

where quantities $\mathbf{z}'\mathbf{h}_k$, $\mathbf{h}_k'\mathbf{h}_k$, $\mathbf{z}'G_0$ and $G_0'\mathbf{h}_k$ can be extracted from matrix $H'H$ and vector $H'\mathbf{z}$, both of which may be computed only once when the procedure is initialized. Hence, iterative computation of $L(\mathbf{q}; \boldsymbol{\theta})$ may be implemented using matrix $\tilde{C}_0^{-1}$ only.

In order for the algorithm to proceed, updated equations for matrix $\tilde{C}^{-1}$ must be derived. Note that the size of $\tilde{C}^{-1}$ varies whenever a pulse is added to or removed from current sequence $\mathbf{q}_0$. Hereafter, we only deal with the addition of a pulse, the other case being a straightforward consequence. Addition of a pulse at site k amounts to a recursion on $\tilde{C}$

$$\tilde{C}_k = \begin{bmatrix} \tilde{C}_0 & \mu G_0' \mathbf{h}_k \\ \mu \mathbf{h}_k' G_0 & 1 + \mu \mathbf{h}_k' \mathbf{h}_k \end{bmatrix}. \tag{21}$$

Then, invoking the inversion lemma for block matrices

$$\tilde{C}_k^{-1} = \begin{bmatrix} \tilde{C}_0^{-1} + \tilde{\mathbf{b}} \tilde{\rho}_k \tilde{\mathbf{b}}' & \tilde{\mathbf{b}} \\ \tilde{\mathbf{b}}' & \tilde{\rho}_k^{-1} \end{bmatrix}$$

where

$$\tilde{\mathbf{b}} \triangleq -\mu \tilde{\rho}_k^{-1} \tilde{C}_0^{-1} G_0' \mathbf{h}_k. \tag{22}$$

Finally, the core algorithm can be summarized as follows.

- **Computation of criterion values**

$$\tilde{\rho}_k = \varepsilon_k + \mu \mathbf{h}_k' \mathbf{h}_k - \mu^2 \mathbf{h}_k' G_0 \tilde{C}_0^{-1} G_0' \mathbf{h}_k \tag{23}$$

$$\mathbf{z}' \tilde{B}_0^{-1} \mathbf{h}_k = \mathbf{z}' \mathbf{h}_k - \mu \mathbf{z}' G_0 \tilde{C}_0^{-1} G_0' \mathbf{h}_k \tag{24}$$

$$L(\mathbf{q}_k; \boldsymbol{\theta}) - L(\mathbf{q}_0; \boldsymbol{\theta}) = \mu \tilde{\rho}_k^{-1} r_n^{-1} (\mathbf{z}' \tilde{B}_0^{-1} \mathbf{h}_k)^2 - \ln(\varepsilon_k \tilde{\rho}_k) - 2\varepsilon_k \ln(1/\lambda - 1) \tag{25}$$

- **Update of** $\tilde{C}^{-1}$ (when $\varepsilon_k = 1$)

$$\tilde{\mathbf{b}} = -\mu \tilde{\rho}_k^{-1} \tilde{C}_0^{-1} G_0' \mathbf{h}_k, \tag{26}$$

$$\tilde{C}_k^{-1} = \begin{bmatrix} \tilde{C}_0^{-1} + \tilde{\mathbf{b}} \tilde{\rho}_k \tilde{\mathbf{b}}' & \tilde{\mathbf{b}} \\ \tilde{\mathbf{b}}' & \tilde{\rho}_k^{-1} \end{bmatrix}. \tag{27}$$

Provided that $H'H$ and $H'z$ have been computed and stored during the initialization, a single evaluation of $L(\mathbf{q}_k; \boldsymbol{\theta}) - L(\mathbf{q}_0; \boldsymbol{\theta})$ and an update of $\tilde{C}^{-1}$ both require $O(N_e^2)$ multiplications. In comparison, the corresponding orders were $O(1)$ and $O(N^2)$ for the former algorithm derived under identical assumptions [7]. This well balanced computational load between exploration and updating enables the efficient implementation of different local (either deterministic or stochastic) selection strategies. Our practical experience showed that it runs faster when associated with a SMLR selection strategy. But the most interesting feature is the drastic storage reduction implied by this core algorithm. This feature allows the processing of signals of arbitrary length provided N_e remains reasonable.

IV. EM and SEM Algorithms

We consider the problem of ML estimation of a parameter $\boldsymbol{\theta}$ given the data $\mathbf{z}$. The estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is defined by

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \triangleq \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\theta}} p(\mathbf{z}; \boldsymbol{\theta}). \tag{28}$$

Dempster *et al.* [31] proposed a rather general scheme for maximizing a likelihood function: The EM algorithm. It is an iterative procedure that increases the likelihood, but does not guarantee convergence toward the ML estimate. A set of conditions required to ensure convergence of the EM algorithm toward the ML estimate can be found in [31] and [32]. The EM algorithm has been successfully applied to various problems in tomography and image processing in situations where direct maximization of the likelihood is too difficult [33]–[35]. Depending on the complexity of the addressed problem, the implementation of the EM algorithm may be very difficult and even impossible. Several algorithms derived independently, such as the Baum–Welsh reestimation formulas [36] for hidden Markov chains, can be viewed as special cases of the EM formalism as introduced by Dempster *et al.*

A. EM Algorithm

The EM algorithm relies on the introduction of an *auxiliary variable* $\mathbf{y}$ which, roughly speaking, makes the likelihood $p(\mathbf{z}, \mathbf{y}; \boldsymbol{\theta})$ easier to compute. Classically, $\{\mathbf{z}, \mathbf{y}\}$ is referred to as the complete data set. Let $\boldsymbol{\theta}_0$ denote the current estimate

of the parameter, and define

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}_0) \triangleq \int p(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta}_0) \ln p(\mathbf{z},\mathbf{y};\boldsymbol{\theta}) d\mathbf{y},$$
$$D(\boldsymbol{\theta}\|\boldsymbol{\theta}_0) \triangleq \int p(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta}_0) \ln \frac{p(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta}_0)}{p(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta})} d\mathbf{y}. \tag{29}$$

$D(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)$ is a Kullback–Leibler distance, which is known to be nonnegative [37]. Definitions (29) and a logarithmic form of Bayes rule yield

$$\ln p(\mathbf{z};\boldsymbol{\theta}) - \ln p(\mathbf{z};\boldsymbol{\theta}_0) = Q(\boldsymbol{\theta},\boldsymbol{\theta}_0) - Q(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0) + D(\boldsymbol{\theta} \,\|\, \boldsymbol{\theta}_0). \tag{30}$$

Since $D(\boldsymbol{\theta} \,\|\, \boldsymbol{\theta}_0)$ is nonnegative, any value of $\boldsymbol{\theta}$ such that $Q(\boldsymbol{\theta},\boldsymbol{\theta}_0) > Q(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)$ increases the likelihood. The rationale underlying the EM algorithm consists of increasing the likelihood at each iteration by choosing the value of $\boldsymbol{\theta}$ that maximizes $Q(\boldsymbol{\theta},\boldsymbol{\theta}_0)$. Thus, one iteration of the algorithm is usually defined by the two steps [31] that follow:

$$\begin{aligned} &\textit{Expectation } (E) && \text{Compute } Q(\boldsymbol{\theta},\boldsymbol{\theta}_i), \text{function of } \boldsymbol{\theta}, \\ &\textit{Maximization } (M) && \hat{\boldsymbol{\theta}}_{i+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_i). \end{aligned}$$

Convergence rates are linear and depend on the comparative information content of the complete data and the observed data about $\boldsymbol{\theta}$: less-informative complete data sets lead to improved convergence rate [31], [38]. Conversely, less-informative data sets can yield intractable computations in the E-step as well as in the M-step. In fact, the choice of the complete data set governs the tradeoff between convergence rate and complexity.

EM algorithms are interesting in situations where an auxiliary variable $\mathbf{y}$ can be chosen such that $Q(\boldsymbol{\theta},\boldsymbol{\theta}_0)$ can be maximized more easily than $p(\mathbf{z};\boldsymbol{\theta})$. This is the case in problems of mixtures, missing or censored data, and in positron emission tomography [35]. But for unsupervised BG deconvolution, no choice of auxiliary quantity allows implementation of EM. Therefore an alternative approach must be sought. Here we adopt a SEM algorithm in which Q is approximated in a stochastic manner.

B. SEM as a Stochastic Approximation of Q

Since Q cannot be computed nor maximized, we are looking for some approximation to this quantity. This kind of problem occurs in settings not restricted to BG deconvolution [39], [40]. As the definition of Q involves an expectation operator, we can resort to stochastic approximation techniques and replace the expectation by averages of K samples $(\mathbf{Y}_k, \quad k = 1 \cdots K)$ drawn from $p(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta}_0)$ [39], [40], [41]

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}_0) = E[\ln p(\mathbf{z},\mathbf{Y};\boldsymbol{\theta}) \mid \mathbf{z};\boldsymbol{\theta}_0] \approx \frac{1}{K}\sum_{k=1}^{K} \ln p(\mathbf{z},\mathbf{Y}_k;\boldsymbol{\theta}). \tag{31}$$

In the sequel we only consider the case $K = 1$, which results in simpler computations: This choice corresponds to the SEM algorithm, that was used by Celeux and Diebolt [41] for estimation of mixture parameters in order to speed-up the convergence of EM algorithms. One iteration of the SEM algorithm is defined by the following two steps:

$$\begin{aligned} &\textit{Sampling } (S) && \text{sample } \mathbf{Y}_i \text{ from } p(\mathbf{y} \mid \mathbf{z};\hat{\boldsymbol{\theta}}_i) \qquad (32) \\ &\textit{Maximization } (M) && \hat{\boldsymbol{\theta}}_{i+1} = \arg\max \ln p(\mathbf{z},\mathbf{Y}_i;\boldsymbol{\theta}). \qquad (33) \end{aligned}$$

This process generates an homogeneous Markov chain $(\hat{\boldsymbol{\theta}}_i)$.

When the goal of the estimation procedure is to determine the proportion λ^* of a two-component mixture of known densities, Celeux and Diebolt [42], [43] established several properties of the chain, the most important of which are given hereafter, as follows:

1) The chain is ergodic and converges to its steady-state distribution ψ_P (where P denotes the size of $\mathbf{z}$);
2) Let Λ_P and λ_P respectively denote a random variable with distribution ψ_P, and the ML estimate associated with the sample of size P. As P goes to infinity $\sqrt{P}(\Lambda_P - \lambda_P)$ converges in distribution to a zero-mean Gaussian distribution.

The practical significance of the above properties is the following: A realization of the SEM chain can be decomposed into a mean value and some residual additive noise. When P grows to infinity the mean value converges to the true proportion while the variance of the residual noise decreases to zero at rate $O(1/P)$. This mean value is determined practically using averages over I successive samples $\sum_1^I \hat{\boldsymbol{\theta}}_i/I$.

These results have been formally established in the case of the above mentioned mixture problem only. However, Celeux and Diebolt conjecture that their results hold in more general cases [42] and the numerical experiments presented in Section VI tend to support their conjecture. In the next section we present the main features of our implementation of SEM applied to unsupervised BG deconvolution.

V. Unsupervised Deconvolution Using a SEM Approach

For SEM as well for EM algorithms, the first issue to be dealt with is the choice of auxiliary variable $\mathbf{y}$. In [27] and [28], Lavielle described a SEM approach to unsupervised deconvolution of BG processes in which $\mathbf{y} = \{\mathbf{q},\mathbf{r}\}$. Such a choice yields an extremely simple M-step; this allowed Lavielle to estimate both filter coefficients and hyperparameters when the signal-to-noise ratio is high ($\geq$50 dB). However, no systematic study of the hyperparameter estimates is provided in this work.

Convergence of SEM to its steady-state distribution is governed by the underlying EM algorithm. As mentioned in Section IV-A, convergence of the procedure can be sped up if the information content of the complete data set is reduced. Such a reduction can be obtained by including only one of the components of $\mathbf{y}$ in the complete data set. As the amplitude variable $\mathbf{r}$ is defined conditionally to the location variable $\mathbf{q}$ (see (2)), $\mathbf{q}$ is the natural choice for building the complete data set. This choice results in the desired reduction of information content but, on the other hand, turns the M-step into a nonlinear optimization problem. One last difficulty should be mentioned: in the S-step, the probability distribution $\Pr(\mathbf{q} \mid \mathbf{z};\boldsymbol{\theta})$ must be evaluated for all possible configurations of $\mathbf{q}$ each time a sample of the auxiliary variables $\mathbf{Q}$ is drawn, i.e., at each iteration of the algorithm. As already underlined in Sections II and III, such computations cannot

be implemented in practice. The techniques developed to overcome the difficulties of the M- and S-steps are presented hereafter.

A. M-Step

The optimization problem is defined by (33) with $\mathbf{Y} = \mathbf{Q}$ and $\boldsymbol{\theta} = \{\lambda, r_x, r_n\}$. Therefore, the criterion to be maximized is identical to marginal detection criterion (12) used in supervised deconvolution and it takes the following form:

$$L(\mathbf{Q};\boldsymbol{\theta}) = L_{\mathbf{Q}}^{(1)}(r_x, r_n) + 2L_{\mathbf{Q}}^{(2)}(\lambda)$$

where

$$L_{\mathbf{Q}}^{(1)}(r_x, r_n) \triangleq -\mathbf{z}' B^{-1}\mathbf{z} - \ln|B|$$

and

$$L_{\mathbf{Q}}^{(2)}(\lambda) \triangleq N_e \ln\lambda + (N - N_e)\ln(1-\lambda).$$

SEM iteration index i is omitted in order to simplify the notations. Optimization must be performed with respect to $\boldsymbol{\theta}$. Maximization of L can be decoupled into two independent optimization problems

$$\hat{\lambda} = \arg\max_{\lambda} L_{\mathbf{Q}}^{(2)}, \tag{34}$$

$$\{\hat{r}_x, \hat{r}_n\} = \arg\max_{r_x, r_n} L_{\mathbf{Q}}^{(1)}(r_x, r_n). \tag{35}$$

Solution to (34) is immediately given by $\hat{\lambda} = N_e/N$. Determination of $\{\hat{r}_x, \hat{r}_n\}$ is not as straightforward. In order to make the computations easier, we use the normalized quantities $\mu = r_x/r_n$ and $\tilde{B} = B/r_n$ introduced in Section III-B. Let $\tilde{L}_{\mathbf{Q}}^{(1)}(\mu, r_n)$ be defined by

$$\begin{aligned}\tilde{L}_{\mathbf{Q}}^{(1)}(\mu, r_n) &\triangleq L_{\mathbf{Q}}^{(1)}(\mu r_n, r_n) \\ &= -\ln|\tilde{B}| - P\ln r_n - \mathbf{z}'\tilde{B}^{-1}\mathbf{z}/r_n. \end{aligned}\tag{36}$$

Maximization of $\tilde{L}_{\mathbf{Q}}^{(1)}(\mu, r_n)$ with respect to r_n immediately yields

$$\hat{r}_n(\mu) = \mathbf{z}'\tilde{B}^{-1}\mathbf{z}/P \tag{37}$$

and substituting the above expression into (36) allows us to transform the original two-dimensional optimization problem (35) into the following one-dimensional maximization problem:

$$\hat{\mu} = \arg\max \tilde{L}(\mu), \quad \tilde{L}(\mu) \triangleq -P\ln \mathbf{z}'\tilde{B}^{-1}\mathbf{z} - \ln|\tilde{B}|. \tag{38}$$

After $\hat{\mu}$ is determined, $\hat{r}_n$ is computed from $\hat{\mu}$ through (37). The goal is now to find the value of μ that maximizes $\tilde{L}(\mu)$. Note that existence of such a maximum in $[0, +\infty)$ is guaranteed, since $\tilde{L}$ is a continuous function and it can be shown that

$$\lim_{\mu\to 0}\tilde{L}(\mu) = -P\ln \mathbf{z}'\mathbf{z} \quad \text{and} \quad \lim_{\mu\to+\infty}\tilde{L}(\mu) = -\infty.$$

In order to reduce the actual dimension of the problem from N to N_e, we express $\tilde{B} = \mu GG' + I$ in terms of $\tilde{C} = \mu G'G + I$ as in Section III-B, and we get

$$\tilde{L}(\mu) = -P\ln(\mathbf{z}'\mathbf{z} - \mu\mathbf{z}'G\tilde{C}^{-1}G'\mathbf{z}) - \ln|\tilde{C}|. \tag{39}$$

Computation of the first and second derivatives of $\tilde{L}(\mu)$ can be carried out easily after diagonalization of $(N_e \times N_e)$ matrix $G'G$. Then, $\hat{\mu}$ can be determined using any second-order local descent algorithm. Finally, the algorithm for estimation of $\{r_x, r_n\}$ can be summarized as follows.

- Diagonalization of $G'G$ and computation of the first and second derivatives of $\tilde{L}(\mu)$.
- Determination of $\hat{\mu}$ using a second-order descent algorithm.
- Computation of $\hat{r}_n$ using (37) and (18).
- Computation of $\hat{r}_x = \hat{\mu}\hat{r}_n$.

This algorithm is easy to implement and presents a moderate computational complexity, as the dimension of the matrices involved in the computation of the derivatives of $\tilde{L}$ is reduced from N to N_e.

B. S-Step

Here again, SEM iteration index i is omitted in order to simplify the notations. The S-step is defined by

$$\textit{Sampling } (S) \quad \text{sample } \mathbf{Q} \text{ from } \Pr(\mathbf{q} \mid \mathbf{z};\boldsymbol{\theta}). \tag{40}$$

Direct sampling of $\mathbf{Q}$ according to (40) would require $\Pr(\mathbf{q} \mid \mathbf{z};\boldsymbol{\theta})$ to be evaluated for all possible configurations of $\mathbf{q}$. As underlined in Sections II and V, such a task is intractable in practice, and an indirect approach must be utilized.

The problem of sampling a random vector $\mathbf{Y}$ with many possible discrete states from a given distribution $\pi(\mathbf{Y})$ is frequently encountered in the area of Markov-based image processing. A classical solution, referred to as *stochastic relaxation*, consists of generating a homogeneous and reversible Markov chain $(\mathbf{Y}_j)_{j\geq 0}$ which converges in distribution to $\pi(\mathbf{Y})$. The desired sample is taken from $(\mathbf{Y}_j)_{j\geq 0}$ after the steady-state is reached.

With stochastic relaxation algorithms, only a small number of components of $\mathbf{Y}_j$ may be modified during the transition from state j to state $j+1$. Let S_j and $\mathbf{Y}_{S_j}$ denote the set of components that may be modified and the corresponding subvector of $\mathbf{Y}_j$, respectively. Define $\bar{\mathbf{Y}}_{S_j} \triangleq \mathbf{Y}_j \setminus \mathbf{Y}_{S_j}$. One iteration of the algorithm is made up of two steps: i) selection of site S_j and ii) sampling of $\mathbf{Y}_{S_j}$ from a transition probability distribution, which is a function of conditional probability $\pi(\mathbf{Y}_{S_j} \mid \bar{\mathbf{Y}}_{S_j})$. Within this framework, several algorithms may be derived; they differ mainly by the site selection technique, which may be either deterministic or probabilistic, and by the specific relationship between transition probabilities and $\pi(\mathbf{Y}_{S_j} \mid \bar{\mathbf{Y}}_{S_j})$. Convergence of several of these procedures is investigated in [44].

From a practical standpoint, these algorithms are useful only if conditional probabilities $\pi(\mathbf{Y}_{S_j} \mid \bar{\mathbf{Y}}_{S_j})$ can be evaluated easily. Such a situation prevails when π is the distribution of a Markov field, but also in our case, where $\pi(\mathbf{q}) = \Pr(\mathbf{q} \mid \mathbf{z};\boldsymbol{\theta})$. Indeed, the core algorithm presented in Section III-B provides a simple way of evaluating the variation of $\ln\Pr(\mathbf{q} \mid \mathbf{z};\boldsymbol{\theta})$ when a component q_j of $\mathbf{q}$ is modified. Computation of conditional probabilities $\Pr(q_j \mid \mathbf{z}, \bar{\mathbf{q}}_j;\boldsymbol{\theta})$ follows easily.

The stochastic relaxation algorithm must be selected so as to minimize the amount of computations required to reach a steady state.

According to the results presented in [44], a Gibbs sampler with a predefined (i.e., deterministic) ordering of the sites is best suited to our problem. At this point, the parameters that remain to be fixed are the number of components of $\mathbf{q}$ that may be modified at each iteration, the exact ordering of the sites and the initial state of the chain. Our implementation of SEM makes use of a cyclical sweep with sites made up of one sample only. An extension that use sites made up of two adjacent samples of $\mathbf{q}$ in order to speed up the convergence of the chain has been developed and tested in [29]. The estimation results obtained with this second-order extension where statistically indistinguishable from those presented in Section VI.

Finally, in order to achieve fast convergence, the initial value of the chain must be selected among the likely states of the steady-state distribution. Supervised deconvolution algorithms presented in Section III aim at finding a value of $\mathbf{q}$ that maximizes the steady-state distribution. We therefore use the SMLR algorithm (see Section III-A) to initialize the Gibbs sampler.

A synthetic view of the overall unsupervised deconvolution procedure is given in Fig. 2. In this block-diagram, most of the computational time is spent in the S-step by the Gibbs sampler. In our experiments, sampling was achieved with five cyclical sweeps, i.e., $5N$ iterations of the Gibbs sampler. Each iteration consists essentially of evaluating likelihood increments as in (25). Each evaluation requires $O(N_e^2)$ (say, $O(\lambda^2 N^2)$ on the average) multiplications. As I successive samples of the SEM chain $(\hat{\boldsymbol{\theta}}_i)$ are needed to form the final estimate $\hat{\boldsymbol{\theta}}$, the whole procedure requires about $5I\lambda^2 N^3$ multiplications. Note that I does not have to increase as N grows. Therefore, the overall numerical complexity is $O(\lambda^2 N^3)$, which is reasonable for usual values of λ and N (up to about 500 for implementation on usual workstations).

C. MGL Estimation

In this section, we briefly review the MGL method that will be compared to the SEM approach in the simulations of Section VI.

Due to implementation convenience, the most popular MGL approach [21]–[23] relies on maximization of *joint* likelihood $p(\mathbf{z}, \mathbf{r}, \mathbf{q}; \boldsymbol{\theta}) = p(\mathbf{z} \mid \mathbf{r}; \boldsymbol{\theta}) p(\mathbf{r} \mid \mathbf{q}; \boldsymbol{\theta}) \Pr(\mathbf{q}; \boldsymbol{\theta})$ w.r.t. $\mathbf{r}$, $\mathbf{q}$ and $\boldsymbol{\theta}$. Such an approach presents two major drawbacks: (i) It lacks reliability, since a large number of false alarms is generally obtained; (ii) It is not bounded above on the range of its natural parameter set [24], and the search domain for $\boldsymbol{\theta}$ must therefore be restricted in order for an estimate to be defined. On the other hand, it is acknowledged in [21], [22] that maximization of *marginal* likelihood $p(\mathbf{z}, \mathbf{q}; \boldsymbol{\theta}) = p(\mathbf{z} \mid \mathbf{q}; \boldsymbol{\theta}) \Pr(\mathbf{q}; \boldsymbol{\theta})$ (see (4)) w.r.t. $\mathbf{q}$ and $\boldsymbol{\theta}$ is far more reliable although more computationally demanding. In addition, it can be shown that

- when $\mathbf{q}$ is held constant, $p(\mathbf{z}, \mathbf{q}; \boldsymbol{\theta})$ is bounded above: by (9) and (12) we have $L(\mathbf{q}; \boldsymbol{\theta}) \propto \ln p(\mathbf{z}, \mathbf{q}; \boldsymbol{\theta})$, and $L(\mathbf{q}; \boldsymbol{\theta})$ is bounded above w.r.t. $\boldsymbol{\theta}$ (see Section V-A);

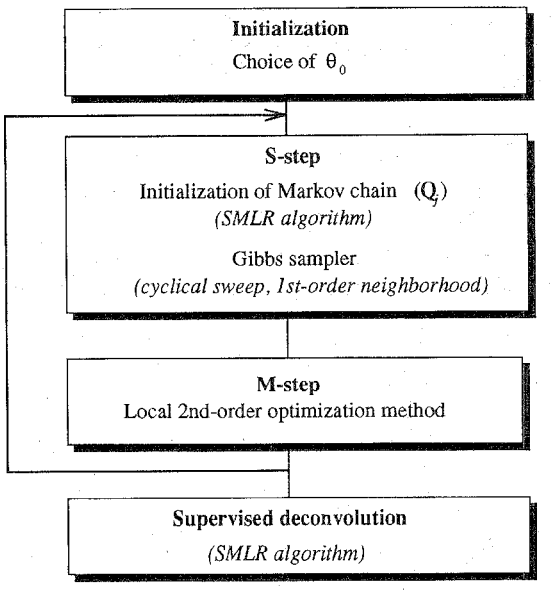


Fig. 2. Schematic representation of the unsupervised deconvolution procedure.

- since $\mathbf{q}$ spans a finite set, the marginal likelihood is bounded above w.r.t. both $\mathbf{q}$ and $\boldsymbol{\theta}$.

Therefore, we adopted the marginal criterion to perform MGL estimation (this approach is referred to as *Kormylo's maximum likelihood deconvolution* [22]).

For computing the MGL estimates, we use the classical suboptimal alternating technique

$$\hat{\mathbf{q}}_i = \arg \max_{\mathbf{q} \in \{0,1\}^N} p(\mathbf{z}, \mathbf{q}; \hat{\boldsymbol{\theta}}_i) \tag{41}$$

$$\hat{\boldsymbol{\theta}}_{i+1} = \arg \max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{z}, \hat{\mathbf{q}}_i; \boldsymbol{\theta}). \tag{42}$$

This procedure is often referred to as a *block-component method* [22], and converges in a finite number of iterations as it generates a sequence $(\hat{\mathbf{q}}_i)_{i \geq 1}$ that increases the function $p(\mathbf{z}, \mathbf{q}; \hat{\boldsymbol{\theta}}(\mathbf{q}))$, which spans a finite set.

Comparison of (42) and (33) reveals that the second step of this iterative procedure is identical to the M-step of the SEM algorithm. This step is performed as explained in Section V-A. Step (41) in the iterative procedure corresponds to maximization of the marginal likelihood used for detection of the Bernoulli sequence in supervised deconvolution (see (4)). As shown in Section III, exact maximization of the criterion is not feasible. Following the discussion in Section III-A, we chose the SMLR algorithm to (suboptimally) perform this task.

VI. Numerical Experiments

In this section, we compare the results of unsupervised deconvolution applied to a set of synthetic data corresponding to the single parameter set

$$\lambda^* = 0.05, \quad r_x^* = 1 \quad \text{and} \quad r_n^* = 0.005.$$

Such a choice corresponds to an usual value for λ^* in the context of BG deconvolution, and to a standard signal-to-noise ratio $(\lambda r_x r_h / r_n)$ of 10 dB, provided the energy of the filter, r_h, equals 1.

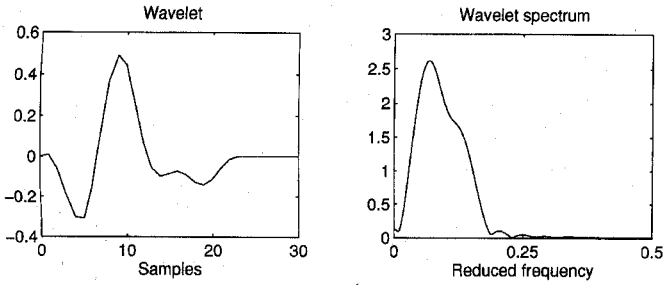


Fig. 3. Impulse response and power spectrum of the linear system used for the simulations.

A first set of 1000 realizations of 200-sample long BG input signals was drawn. Each signal was convolved with the wavelet, i.e., the impulse response, depicted in Fig. 3. This wavelet had been separately identified on actual seismic data. Its spectrum is poor but realistic, contrarily to the classical Kramer wavelet used in many simulations [21]. The wavelet energy was normalized to one. Then a realization of Gaussian white noise with variance r_n^* was added to each convolved trace. This first set of "short" signals was processed by supervised and unsupervised deconvolution algorithms.

Then a second set of "long" signals was generated in order to investigate the asymptotic behavior of hyperparameters estimates in unsupervised deconvolution algorithms. It consisted of 400 signals each made up of 1600 samples. Whereas the processing of signals of size $N = 200$ requires a moderate computation time, it is not so for signals of size $N = 1600$. According to the complexity analysis presented in Section V-B, as λ^* was kept constant while the size of each trace was multiplied by a factor of eight, the overall computational load was multiplied by a factor $8^3 \approx 500$. This corresponds to a huge increase of the computational burden.

Our first goal was to study the statistical behavior of SEM and MGL. We investigated the bias (B), variance (V) and mean-square error (MSE) of the hyperparameter estimates obtained with the four algorithms. Among these quantities, the MSE is the only one that defines a distance measure between true and estimated parameters. Therefore, MSE is viewed as the most significant one.

As these quantities do not admit closed-form expressions, Monte Carlo simulations were employed. Let K denote the number of signals in a given data set, and let $\hat{\boldsymbol{\theta}}(\mathbf{Z}_k)$ be the hyperparameter estimate obtained with signal k. B, V and MSE in one set were determined using sample averages as follows:

$$B = \frac{1}{K}\sum_{k=1}^{K}(\hat{\boldsymbol{\theta}}(\mathbf{Z}_k) - \boldsymbol{\theta}^*), \text{MSE} = \frac{1}{K}\sum_{k=1}^{K}(\hat{\boldsymbol{\theta}}(\mathbf{Z}_k) - \boldsymbol{\theta}^*)^2$$

and

$$V = \text{MSE} - B^2.$$

Then we define the normalized quantities

$$\tilde{B} \triangleq B/\boldsymbol{\theta}^*, \quad \tilde{S} \triangleq \sqrt{V}/\boldsymbol{\theta}^* \quad \text{and} \quad \tilde{E} \triangleq \sqrt{\text{MSE}}/\boldsymbol{\theta}^*.$$

Quantities $\tilde{B}$, $\tilde{S}$ and $\tilde{E}$ can be interpreted as relative error measures that we wish as small as possible.

TABLE I
NORMALIZED MEAN SQUARE ERROR OF THE ESTIMATES

		λ	r_x	r_n
Short signals	MGL	0.34	0.73	0.14
	SEM	0.34	0.6	0.12
Long signals	MGL	0.21	0.28	0.067
	SEM	0.14	0.18	0.041

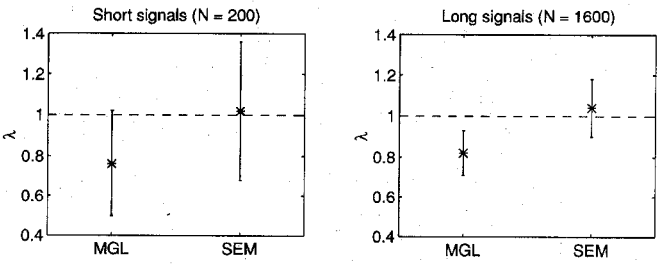


Fig. 4. Normalized mean and standard deviation of the estimates ($\lambda^* = 0.05$).

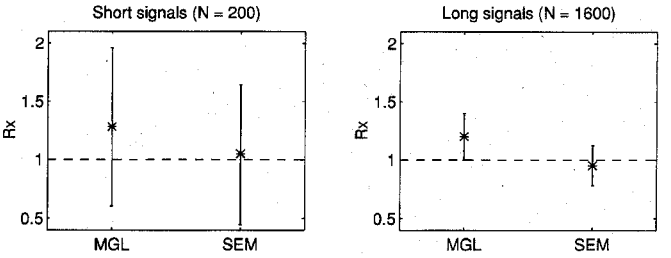


Fig. 5. Normalized mean and standard deviation of the estimates ($r_x^* = 1$).

The algorithms where initialized with $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}^0 \triangleq (\lambda^0, r_x^0, r_n^0)$, where

$$\lambda^0 = 0.01, \quad r_x^0 = 4 \quad \text{and} \quad r_n^0 = 0.01.$$

Recall that in the case of SEM estimation (cf. Section IV-B), a Markov chain $(\tilde{\boldsymbol{\theta}}_i)$ is generated. Then, the final hyperpameter estimate is taken as the average over a number I of successive samples of the chain. This number must be fixed. After preliminary tests we chose $I = 20$ samples. This empirical rule is qualified later in this section.

$\tilde{B}$, $\tilde{S}$ and $\tilde{E}$ were evaluated for each hyperparameter. The different values of $\tilde{E}$ for each hyperparameter are presented in Table I. $\tilde{B}$ and $\tilde{S}$ are represented graphically in Figs. 4–6. Each figure corresponds to one hyperparameter. Parameter r_n is always better estimated than r_x and λ as testified by MSE values relative to this parameter.

In terms of MSE, SEM is superior to MGL. SEM and MGL present similar variances, and the larger MSE of MGL is essentially produced by a higher bias. From this first analysis, we may conclude that from an estimation standpoint, SEM should be preferred to MGL. The asymptotic behavior of the estimates, which may be inferred from the comparison of results obtained with short and long data sets, confirms these first conclusions. As the number of samples is multiplied by a factor of eight, a reduction in variance of approximatively the same factor is expected according to classical asymptotic

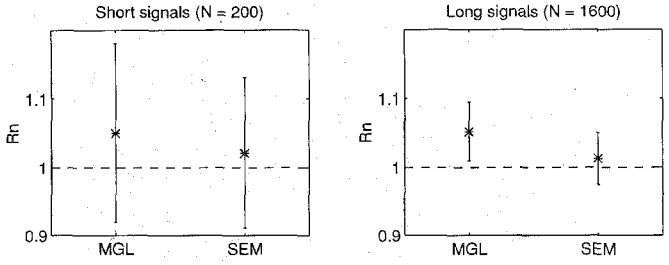


Fig. 6. Normalized mean and standard deviation of the estimates ($r_n^* = 0.005$).

TABLE II
SCORE $S(i)$ AND MEAN EXECUTION TIME OF SUPERVISED AND UNSUPERVISED DECONVOLUTION ALGORITHMS

	SMLR	MGL	SEM
Score	0.52	1.62	0.61
Mean time (s.)	0.12	0.46	13

theory. Such a reduction is indeed observed. For the SEM method, the MSE decreases by a factor of nine, as the bias has a negligible influence in the MSE. For MGL, however, the reduction of the MSE is smaller as the bias of the estimates remains essentially unchanged. This is consistent with the formerly reported nonconsistency of MGL estimates. Therefore, it appears that SEM yields the best asymptotic behavior, which is in agreement with the conjecture of Celeux and Diebolt [42], [43].

Our second goal was to investigate the degradation of the detection performance when we go from a supervised method to an unsupervised one. When θ^* is known an estimated Bernoulli sequence $\mathbf{q}^s$ may be obtained using a supervised method such as SMLR (see Section III). When θ^* is not known, an unsupervised method, such as MGL or SEM, must be employed to compute an estimate $\hat{\theta}$ and a Bernoulli sequence $\mathbf{q}^u$. We want to assess how close $\mathbf{q}^u$ is from $\mathbf{q}^s$. Let $\mathbf{q}_k^{(i)}$ be the detection sequence estimated for signal k using method $i \in \{\text{SMLR}, \text{MGL}, \text{SEM}\}$. We defined a performance index $L_k^{(i)}$ as

$$L_k^{(i)} \triangleq L(\mathbf{q}_k^{(i)}, \theta^*) \quad \text{where } i \in \{\text{SMLR}, \text{MGL}, \text{SEM}\}.$$

Then figures of merit $S(i)$, $i \in \{\text{SMLR}, \text{MGL}, \text{SEM}\}$ were computed according to

$$S(i) = \sum_{k=1}^{K} \left(\hat{L}_k - L_k^{(i)}\right)/\hat{L}_k \quad \text{where } \hat{L}_k = \arg\max_i L_k^{(i)}. \tag{43}$$

The above expressions show that an efficient algorithm will obtain a low score (0 if it is always better than the other ones), and that the figure of merit increases as efficiency decreases. The scores obtained by the three algorithms together with the mean execution time on HP715-type workstations are presented in Table II for the set of short signals. They show that, on the average, SEM performance is very close to the supervised result, and it is significantly better than MGL results.

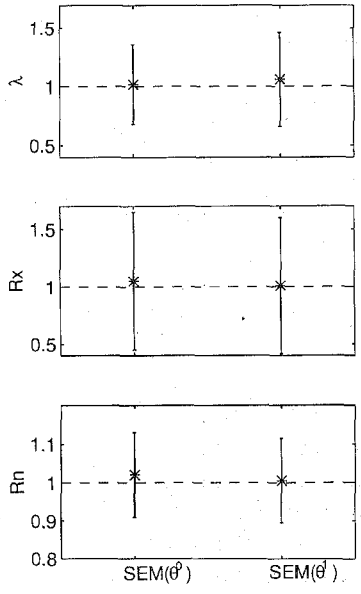


Fig. 7. Normalized mean and standard deviation of the estimates for two initialization choices.

Our third goal was to investigate initialization issues. This is a potential problem since the SEM algorithm is designed to mimick the EM algorithm, and the latter generally offers no guarantee of global convergence. Here again, no specific theoretical result may be obtained due to the intricate nature of the likelihood (11). Therefore, we carried out another set of experiments with a different initial value $\hat{\theta}_0 = \theta^1 \triangleq (\lambda^1, r_x^1, r_n^1)$

$$\lambda^1 = 0.2, \quad r_x^1 = 0.15 \quad \text{and} \quad r_n^1 = 0.03.$$

We observed that, on the average, this initialization caused slower convergence of the SEM chain to its equilibrium, so that a larger number ($I = 30$) of SEM samples was required, and the first ten samples where discarded from the computation of the estimated parameter. Comparison of normalized bias and standard deviation for both initializations with SEM algorithms is shown in Fig. 7. It appears that estimation of r_x and r_n is not sensitive to initialization. As regards the estimation of λ, initialization with θ^1 yields a slight increase of variance. Therefore, our feeling about the method is that initialization is not a critical issue provided that the number of samples drawn from the SEM chain has been properly set. A probably more flexible technique would be to design simple tests to adapt I to the currently processed data.

VII. CONCLUSION

This paper addressed the problem of hyperparameter estimation in the context of BG deconvolution. Our main goal was to offer an alternative to classical MGL approaches, because these techniques do not yield consistent estimates, and may not define any estimates at all. However, implementation of estimators with better statistical behavior is often difficult. Here, a maximum likelihood estimator was adopted, and its implementation was carried out using a stochastic approximation of the EM algorithm. Similar techniques were

formerly proposed in the same context by Lavielle [27], [28] and Goussard [45]. Our first contribution is the derivation of a new core algorithm for BG supervised and unsupervised deconvolution. With this new structure, memory requirements are reduced drastically so that signals of arbitrary length may be processed, provided that the number of spikes remains moderate. For unsupervised deconvolution, the core algorithm is one key element that makes the SEM approach practical and efficient for signals of realistic sizes, another one being an adequate choice of the complete data set.

Finally, assessment of the performances of the proposed methods was performed using large data sets and Monte Carlo simulations. For unsupervised deconvolution, the statistical behavior of the SEM estimator was studied, and comparisons were performed with an MGL method.

The results show the superiority of the SEM over the MGL approach, at least for the class of tested signals. On the other hand, SEM proves to be more computationally demanding due to the sampling part of the method, but this load is rather small if the number of spikes in the signal is moderate. In this situation, SEM will be preferred since it features smaller bias than and comparable variance to MGL. However, when the actual number of spikes in the signal is very large, both methods become computationally intensive.

The comparatively good statistical properties of the SEM methods suggest the following further extensions of this work.

- *Extension of the technique to blind deconvolution where the filter as well as hyperparameters must be identified*—this extension requires a reconsideration of the M-step only, the S-step being left unchanged.
- *Introduction of a relaxation between successive samples of the SEM Markov chain*—then almost sure convergence toward maximum likelihood might be foreseen when the relaxation step decreases to zero. Therefore, such an approach should yield stronger convergence results than those available for SEM in its present form.

REFERENCES

[1] A. K. Mahalanabis, S. Prasad, and K. P. Mohandas, "Recursive decision-directed estimation of reflection coefficients for seismic data deconvolution," *Automatica*, vol. 18, pp. 721–726, 1982.

[2] G. B. Giannakis, J. M. Mendel, and X. Zhao, "A fast prediction-error detector for estimating sparse-spike sequences," *IEEE Trans. Geosci. Remote Sensing*, vol. 27, pp. 344–351, 1989.

[3] Y. Goussard and G. Demoment, "Recursive deconvolution of Bernoulli-Gaussian processes using a MA representation," *IEEE Trans. Geosci. Remote Sensing*, vol. 27, pp. 384–394, 1989.

[4] J. Idier and Y. Goussard, "Stack algorithm for recursive deconvolution of Bernoulli-Gaussian processes," *IEEE Trans. Geosci. Remote Sensing*, vol. 28, pp. 975–978, 1990.

[5] J. Kormylo and J. M. Mendel, "Maximum-likelihood detection and estimation of Bernoulli-Gaussian processes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 482–488, 1982.

[6] C. Y. Chi and J. M. Mendel, "Improved maximum-likelihood detection and estimation of Bernoulli-Gaussian processes," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 429–435, 1984.

[7] Y. Goussard, G. Demoment, and J. Idier, "A new algorithm for iterative deconvolution of sparse spike trains," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Albuquerque, NM, 1990, pp. 1547–1550.

[8] M. Lavielle, "Bayesian deconvolution of Bernoulli-Gaussian processes," *Signal Processing*, vol. 33, pp. 67–79, 1993.

[9] *IEEE Trans. Automat. Contr.*, vol. 35, 1990 (Spec. Issue Higher Order Stat. Syst. Theory Signal Processing).

[10] *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, 1990 (Spec. Issue Higher Order Stat. Syst. Theory Signal Processing).

[11] J. M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications," *Proc. IEEE*, vol. 79, pp. 278–305, 1991.

[12] A. T. Walden, "Non-Gaussian reflectivity, entropy, and deconvolution," *Geophys.*, vol. 50, pp. 2862–2888, 1985.

[13] E. Gassiat, *Déconvolution Aveugle*, Ph.D, dissertation, Univ. de Paris-Sud, Centre d'Orsay, France, 1988.

[14] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, pp. 349–369, 1989.

[15] S. Bellini and F. Rocca, "Asymptotically efficient blind deconvolution," *Signal Processing*, vol. 20, pp. 193–209, 1990.

[16] P. Devijver and M. Dekesel, "Champs aléatoires de Pickard et modélisation d'images digitales," *Traitement du Signal*, vol. 5, pp. 131–150, 1988.

[17] S. Lakshmanan and H. Derin, "Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 799–813, 1989.

[18] Y. Bar-Shalom, "Optimal simultaneous state estimation and parameter identification in linear discrete-time systems," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 308–319, 1972.

[19] W. L. Tsang, J. D. Glover, and R. E. Bach, "Identifiability of unknown covariance matrices for some special cases of linear, time-invariant, discrete-time dynamic systems," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 970–974, 1981.

[20] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[21] J. M. Mendel, *Optimal Seismic Deconvolution*. New York: Academic, 1983.

[22] J. Goutsias and J. M. Mendel, "Maximum-likelihood deconvolution: An optimization theory perspective," *Geophys.*, vol. 51, pp. 1206–1220, 1986.

[23] Y. Goussard, "Déconvolution de processus aléatoires non-Gaussiens par maximisation de vraisemblances," Ph.D. dissertation, Univ. de Paris-Sud, Centre d'Orsay, France, 1989.

[24] E. Gassiat, F. Monfront, and Y. Goussard, "On simultaneous signal estimation and parameter identification using a generalized likelihood approach," *IEEE Trans. Inform. Theory*, vol. 38, pp. 157–162, 1992.

[25] F. Champagnat and J. Idier, "An alternative to standard maximum likelihood for Gaussian mixtures," in *Proc. IEEE ICASSP*, Detroit, MI, 1995, pp. 2020–2023.

[26] G. Celeux and J. Diebolt, "A probabilistic teacher algorithm for iterative maximum likelihood estimation," in *Classification and Related Methods of Data Analysis*. Amsterdam: Elsevier, North-Holand, pp. 617–623, 1987.

[27] M. Lavielle, "Déconvolution 2-D et détection de ruptures: Applications en géophysique," Ph.D. dissertation, Univ. de Paris-Sud, Centre d'Orsay, France, 1990.

[28] M. Lavielle, "A stochastic algorithm for parametric and nonparametric estimation in the case of incomplete data," *Signal Processing*, vol. 42, pp. 3–17, 1995.

[29] F. Champagnat, Y. Goussard, and J. Idier, "Unsupervised Bernoulli-Gaussian deconvolution," Tech. Rep. LSS, vol. # GPI-94/01, 1994.

[30] P. Lascaux and R. Théodor, *Analyze Numérique Appliquée à l'Art de l'Ingénieur*. Paris, France: Masson, 1986.

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B*, vol. 39, pp. 1–38, 1977.

[32] R. Redner and H. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, pp. 195–239, 1984.

[33] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 477–489, 1988.

[34] M. Segal, E. Weinstein, and B. Musicus, "Estimate-maximize algorithms for multichannel time delay and signal estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 1–16, 1991.

[35] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imaging*, vol. MI-1, pp. 113–122, 1982.

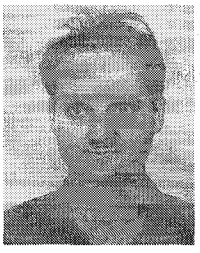
[36] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE Acoust., Speech, Signal Processing*, pp. 4–16, Jan. 1986.

[37] D. Dacunha-Castelle and M. Duflo, *Probabilités et Statistiques–Tome 1: Problèmes à temps fixe*, 2nd ed. Paris: Masson, 1990.

[38] J. A. Fessler and A. O. Hero, "Complete data spaces and generalized EM algorithms," in *Proc. IEEE Int. Conf. ASSP*, Minneapolis, MN, 1993, pp. IV 1–4.

[39] B. Chalmond, "An iterative Gibbsian technique for reconstruction of M-ary images," *Pattern Recog.*, vol. 22, pp. 747–761, 1989.

[40] G. C. G. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *J. Amer. Stat. Assoc.*, vol. 85, pp. 699–704, 1990.
[41] G. Celeux and J. Diebolt, "L'algorithme SEM: Un algorithme d'apprentissage probabiliste pour la reconnaissance de densités," *Revue Stat. Appl.*, vol. 34, pp. 35–52, 1986.
[42] ——, "Reconnaissance de mélange de densité et classification," Tech. Rep. 349, INRIA, 1984.
[43] ——, "Asymptotic properties of a stochastic EM algorithm for estimating proportions," Tech. Rep. 1591, INRIA, 1992.
[44] J. Goutsias, "A theoretical analysis of Monte Carlo algorithms for the simulation of Gibbs random field images," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1618–1628, 1991.
[45] Y. Goussard, "Blind deconvolution of sparse spike trains using stochastic optimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1992, vol. 4, pp. 593–596.

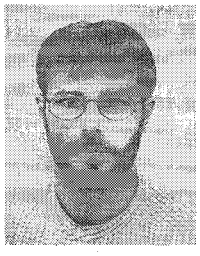


Frédéric Champagnat was born in Dakar, Senegal, in 1966. He graduated from the École Nationale Supérieure de Techniques Avancées in 1989, and received the Ph.D. degree in physics from the Université de Paris-Sud, Orsay, France, in 1993.

In 1994, he joined the Biomedical Engineering Institute of the École Polytechnique, Montreal, Canada, in a postdoctoral position. Since December 1995, he has been with the Laboratoire des Signaux et Systèmes, Gif-sur-Yvette, France. His main research interests are in probabalistic models and algorithms for inverse problems arising in signal and image processing.

Yves Goussard (M'89) was born in Paris, France, in 1957. He graduated from the École National Supèrieure de Techniques Avancées in 1980, and he received the Doc. Ing. and Ph.D. degrees from the Université de Paris-Sud, Orsay, France, in 1983 and 1989, respectively.

From 1983 to 1985, he was a visiting scholar at the Electrical Engineering and Computer Science Department of the University of California, Berkeley. In 1985, he was appointed a Chargé de Recherche at CNRS, Gif-sur-Yvette, France, and in 1992 he joined the Biomedical Engineering Institute of the École Polytechnique, Montreal, Canada, where is is now an Associate Professor. During the academic year 1990–1991, he was on sabbatical leave at the Department of Electrical Engineering Systems, University of Southern California, Los Angeles. After some work on nonlinear system identification and modeling, his interests moved toward ill-posed problems in signal and image processing with application to biological systems.



Jérôme Idier was born in France in 1966. He received the Dipl. degree in electrical engineering from the École Supérieure d'Électricité in 1988 and the Ph.D. degree in physics from the Université de Paris-Sud, Orsay, France, in 1991.

PSince 1991, he has been with the Centre National de la Recherche, assigned to the Laboratoire des Signaux et Systèmes. His major scientific interests are in probabilistic approaches to inverse problems for signal and image processing.