Generalized Marginal Likelihood for Gaussian mixtures

F. Champagnat and J. Idier
Laboratoire des signaux et systèmes
SUPÉLEC
Plateau de Moulon, 91192 GIF-SUR-YVETTE Cédex, France

June 14, 1994

Abstract

The dominant approach in Bernoulli-Gaussian myopic deconvolution consists in the joint maximization of a single *Generalized Likelihood* with respect to the input signal and the hyperparameters. The aim of this correspondence is to assess the theoretical properties of a related *Generalized Marginal Likelihood* criterion in a simplified framework where the filter is reduced to identity. Then the output is a mixture of Gaussian populations. Under a single reasonable assumption we prove that the maximum generalized marginal likelihood estimator always converge asymptotically. Then numerical experiments show that this estimator can perform better than *Maximum Likelihood* (ML) in the finite sample case, moreover asymptotic estimates are significant although biased.

1 Introduction

The problem of the restoration of spiky sequences distorted by a linear system and additive noise arises in areas such as seismic exploration [1], non-destructive evaluation and biomedical engineering [2].

When the filter is assumed known, the *ill-posed* nature of the induced deconvolution problem may be coped in a Bayesian framework using prior information about the expected structure of the input in the form of prior probability models. To allow for a better flexibility, a family of parametered priors is used instead of a unique one, and the set of those parameters as well as those of the noise distribution are often referred as to *hyperparameters*. The problem of the estimation of the input given hyperparameters is often referred as to simple deconvolution.

A recurrent problem in the context of Bayesian estimation is the practical problem of hyperparameter identification. A rather general methodology for this identification task relies on maximization of the likelihood of the parameters, i.e. the probability of the data given the parameters. In many circumstances, a maximum likelihood (ML) approach guarantees consistency and asymptotic efficiency of the related estimators [3].

In many contexts such a scheme cannot be implemented, due to the fact that the likelihood cannot be computed nor maximized in practice, and alternative schemes are to be seeked for. The aim of this correspondence is to assess, in a special case, the asymptotic properties of one alternative methodology based on a Generalized Likelihood (GL), i.e. the probability of the data and the input given the hyperparameters, to be maximized relatively either to the input and hyperparameters. Such methods and variants have been successfully implemented in the various areas of control [4] and image processing [5]. Despite their questionable theoretical properties [6] [7], maximum GL (MGL) techniques are backed by their empirical efficiency.

More specifically, the study presented here is restricted to the case of Bernoulli-Gaussian prior models for the input, and Gaussian white stationary process for the noise. The use of a BG process corresponds to an explicit model of the spiky nature of the input. BG models may be seen as discrete-time composite processes, the first part which is denoted q, modeling the time location for a spike, and the second, denoted r, its amplitude. Up to now, GL maximization has been the dominant method in BG deconvolution problems [8] [9] mainly because of its practicability. Let z denote the observed data, then GL estimation corresponds to the maximization of the joint likelihood $p(z, r, q, \theta)$ in r, q and θ. Gassiat *et al.* [10] presented a theoretical study of that estimator when the filter is reduced to identity. Their results established the poor behavior of such a criterion, and the inability to ensure the existence and the stability of corresponding estimates. Conversely, when estimates exist they may exhibit a small bias.

Based on the remark that joint criterions are not used anymore in the practice of BG simple deconvolution, the following study considers the finite sample and asymptotic behavior associated to a Generalized *Marginal* Likelihood (GML) criterion $p(z, q, \theta)$ in q and θ, where amplitudes of the spikes have been "integrated out". Compared to marginal likelihood methods, joint likelihood detection and estimation demonstrate too many false alarms unless these are penalized in an *ad-hoc* manner (e.g. see [8] on the Chi-t deconvolution method).

In order to be able to carry out mathematical derivations, we had to limit ourselves, following [10], to the case with no distortion. Then the output signal is reduced to the mixture of two zero-mean univariate Gaussian distributions. The estimation of the parameters governing a Gaussian mixture is a yet well documented area [11][12], for which many consistent estimators as ML are already available [13][14]. Our purpose is not to propose another competing estimator in that area but to assess, in a particular case some properties of GML methods. Moreover, such techniques are readily implemented in the general case of a filtered mixture , and may be competitive with more sophisticated ones based on stochastic approximations of ML approaches [15, Chapter 4].

The conclusions of this study are more balanced than those drawn by Gassiat *et al.* on GL criterion: first, we prove the existence of a global maximum for the GML both in the finite sample and asymptotic case. Second, the corresponding estimates possess a scale invariance property, and they asymptotically recover the true power of the signal. We prove the convergence of finite sample estimates toward the global maximum of the asymptotic GML under the reasonable assumption of uniqueness of this maximum. Finite sample and asymtotic maximum GML (MGML) estimates cannot be derived in closed-form. However computation of finite sample MGML estimates can be performed exactly in an efficient manner. A presented Monte Carlo experiment shows that MGML estimation can exhibit smaller bias and mean square error than ML estimation. Furthermore, the associated computationnal load can be neglected compared to ML estimation. The price paid is the loss of consitency for the MGML estimator and a further numerical experiment shows that the asymtotic bias ranges from moderate to large, depending on the amount of noise and the intensity of the pulse process.

2 Problem statement-Finite sample properties

In this section we study the case of a N-sized sample, then corresponding asymptotic expressions will be derived in Section 3.

2.1 Problem formulation

In the absence of distortion, the input-output equation reduces to a spike process corrupted by an additive noise $\mathbf{z} = \mathbf{r} + \mathbf{n}$. The noise $\mathbf{n}$ is assumed to be zero-mean white stationary of variance r_n, and independent from $\mathbf{r}$. The latter is modeled as a sample of a BG process $\mathbf{X} = (\mathbf{Q}, \mathbf{R})$ of the parameters λ, r_x and defined as follows: $\mathbf{Q}$ and $\mathbf{R}$ consist of independent random variables (RV) Q_k and R_k $(1 \leq k \leq N)$;

$$Q_k \text{ is a Bernoulli RV: } P(Q_k = 1) = \lambda \tag{1}$$

$$(R_k \mid Q_k = q) \text{ is a zero-mean Gaussian RV of variance } qr_x.$$

Let $\theta = (\lambda, r_x, r_n)$ denote the vector of hyperparameters that control the different probability distributions, and $\theta^* = (\lambda^*, r_x^*, r_n^*)$ denote the "true" parameters assumed to belong to $\Delta \stackrel{\Delta}{=}]0,1[\times]0,+\infty[^2$. Then the joint maximization of $p(\mathbf{z}, \mathbf{q}; \theta)$ with respect to (w.r.t.) $\mathbf{q}$ and θ amounts to minimization of:

$$L_N(\mathbf{q}, \lambda, r_x, r_n) \stackrel{\Delta}{=} L_N^{(1)}(\mathbf{q}, r_x, r_n) - 2L_N^{(2)}(\mathbf{q}, \lambda),$$

$$\text{where } \; L_N^{(1)}(\mathbf{q}, r_x, r_n) \stackrel{\Delta}{=} N_e \ln(r_x + r_n) + (N - N_e) \ln r_n + \frac{\Sigma_1}{r_x + r_n} + \frac{\Sigma_0}{r_n},$$

$$\text{and: } \; L_N^{(2)}(\mathbf{q}, \lambda) \stackrel{\Delta}{=} N_e \ln \lambda + (N - N_e) \ln(1 - \lambda)\,.$$

Finally N_e, Σ_0 and Σ_1 are defined by:

$$N_e = \sum_{k=1}^{N} q(k)\ , \ \ \Sigma_0 = \sum_{k=1}^{N} (1 - q(k)) z^2(k) \ \text{ and } \ \Sigma_1 = \sum_{k=1}^{N} q(k) z^2(k).$$

For further notationnal convenience we introduce the quantity $\hat{r}_z \stackrel{\Delta}{=} \sum_{k=1}^{N} z^2(k)/N$.

We seek a minimum of this function as $\mathbf{q}$ spans $\{0,1\}^N$, and θ spans Δ. First, we keep $\mathbf{q}$ fixed and optimize on θ. As $\mathbf{q}$ spans a finite set we will conclude to the existence of a global minimum for L_N. Then we will optimize among $\mathbf{q}$s having same N_e. We will thus reduce the original optimization problem to a minimization among at most N distinct values. This result is the basis for an exact numerical optimization scheme for L_N that consists merely in N evaluation of a simple function.

Before proceeding on the optimization of L_N we underline the connections between the MGML and ML methodologies.

2.2 Connections with ML estimation

The GML estimate $(\hat{\theta}, \hat{\mathbf{q}})$ is defined by:

$$(\hat{\theta}, \hat{\mathbf{q}}) = \arg\max_{\theta, \mathbf{q}} p(\mathbf{z}, \mathbf{q}; \theta),$$

then $\hat{\theta}$ can be put in the form:

$$\hat{\theta} = \arg\max_{\theta} \left\{ \max_{\mathbf{q}} p(\mathbf{z}, \mathbf{q}; \theta) \right\}.$$

Let $f(z; r)$ denote the density of a univariate zero-mean Gaussian RV of variance r, then it can be easily shown that:

$$\hat{\theta} = \arg\max_{\theta} \left\{ \prod_{i=1}^{N} \max\left\{ \lambda f(z_i, r_x + r_n), (1 - \lambda) f(z_i, r_n) \right\} \right\},$$

whereas ML estimation yields:

$$\tilde{\theta} = \arg\max_{\theta} \left\{ \prod_{i=1}^{N} \left(\lambda f(z_i, r_x + r_n) + (1-\lambda) f(z_i, r_n) \right) \right\}.$$

Thus, the GML criterion may be viewed as an approximation of the likelihood. In the next sections, it will be shown that this approximation yields much algebraic simplifications at the expense of the loss of consistency. However, the Monte Carlo experiments of Section 4.1 demonstrate that the MGML estimator can yield smaller bias and MSE than ML in the finite sample case.

2.3 Minimization of $L_N(\mathbf{q}, \theta)$ w.r.t. θ

First, we treat the boundary cases $\mathbf{q} = \mathbf{0}$ and $\mathbf{q} = \mathbf{1}$. $\mathbf{q} = \mathbf{0}$ leads to the estimates $\hat{\lambda}(\mathbf{0}) = 0$ (at the boundary of the domain), $\hat{r}_n(\mathbf{0}) = \hat{r}_z$ and

$$L_N(\mathbf{0}, \hat{\lambda}(\mathbf{0}), r_x, \hat{r}_n(\mathbf{0})) = N + N \ln \hat{r}_z.$$

The dual case $\mathbf{q} = \mathbf{1}$ leads to $\hat{\lambda}(\mathbf{1}) = 1$: signal and noise are indistinguishable, $r_x + r_n$ is estimated through $\hat{r}_z$ and

$$L_N(\mathbf{1}, \hat{\lambda}(\mathbf{1}), r_x, \hat{r}_z - r_x) = N + N \ln \hat{r}_z.$$

These criterions are not sensitive to r_x. This means that whenever $\mathbf{q} = \mathbf{0}$ or $\mathbf{q} = \mathbf{1}$, the output is white stationary Gaussian and we cannot tell what is signal or what is noise nor if there is a signal indeed.

We will now discuss the general case. The maximization of $L_N^{(2)}$ w.r.t. λ is straightforward:

$$\hat{\lambda}(\mathbf{q}) = \frac{N_e}{N} \quad \text{and} \quad L_N^{(2)}(\mathbf{q}, \hat{\lambda}(\mathbf{q})) = N_e \ln N_e + (N - N_e) \ln (N - N_e) - N \ln N. \tag{2}$$

Before proceeding we had rather define $\mu = (r_x + r_n)/r_n$ and swap dependent variables $(r_x, r_n) \leftrightarrow (\mu, r_n)$ in $L_N^{(1)}$. Then μ spans $]1, +\infty[$. Let us hold μ fixed and optimize w.r.t. r_n:

$$\hat{r}_n(\mathbf{q}, \mu) = \frac{1}{N} \left(\mu^{-1} \Sigma_1 + \Sigma_0 \right) \tag{3}$$

$$\text{and} \quad L_N^{(1)}(\mathbf{q}, \mu, \hat{r}_n(\mathbf{q}, \mu)) = N \ln \left(\mu^{-1} \Sigma_1 + \Sigma_0 \right) + N_e \ln \mu + N - N \ln N.$$

Finally, let us optimize w.r.t. μ on $]1, +\infty[$:

$$\hat{\mu}(\mathbf{q}) = \max \left\{ \frac{(N - N_e)\Sigma_1}{N_e \Sigma_0}, 1 \right\} \tag{4}$$

$$L_N^{(1)}(\mathbf{q}) = \begin{cases} N \ln (\hat{r}_z) + N & \text{if } (N - N_e)\Sigma_1 \leq N_e \Sigma_0 \\ \\ N_e \ln (\Sigma_1 / N_e) + (N - N_e) \ln (\Sigma_0 / (N - N_e)) + N & \text{otherwise} \end{cases} \tag{5}$$

Thus we reduced the minimization of $L_N(\mathbf{q},\theta)$ to the minimization of $L_N(\mathbf{q},\hat{\theta}(\mathbf{q}))$, which spans a set of at most 2^N distinct values. This result ensures the existence of a finite minimum for L_N, possibly at the boundaries of the set spanned by the hyperparameters. This property is not shared by the joint criterion studied in [10].

2.4 Minimization of $L_N(\mathbf{q},\hat{\theta}(\mathbf{q}))$ w.r.t. q

2.4.1 N_e held fixed

We proceed the optimization among the set $\{\mathbf{q},\mathbf{1}'\mathbf{q}=N_e\}$ (N_e held fixed). This will enable us to devise an optimization scheme for L_N that can be implemented in a simple way in order to compute the corresponding parameters.

To see this, let us sort the data $z(k), k=1..N$ in descending order of $z^2(k)$. At first glance, Σ_1 spans a finite set of values between $s_{min}=\sum_{N-N_e+1}^{N} z^2(k)$ and $s_{max}=\sum_1^{N_e} z^2(k)$, and the function $N_e \ln s + (N-N_e)\ln(\mathbf{z}'\mathbf{z}-s)$ is strictly increasing from s_{min} to $\hat{s}=N_e\hat{r}_z$, then strictly decreasing from $\hat{s}$ to s_{max}. It seems that we have to choose between two potential minima unless we recall that the expression for $L_N^{(1)}$ is valid for the **q**s such that $(N-N_e)\Sigma_1 > N_e\Sigma_0$, or equivalently $\Sigma_1>\hat{s}$. Therefore, when N_e is held fixed, the minimum is obtained while setting to 1 variables $q(k)$ corresponding to the N_e largest values $z^2(k)$. This conclusion is tightly linked to the well known following fact: when θ is held fixed, the optimization of L_N is a mere threshold test depending on θ.

2.4.2 Optimization w.r.t. N_e

Finally, let L'_N be a function defined on $\{0,1,\ldots,N\}$ by

$$L'_N(N_e)=N_e\ln\frac{\sum_{k=1}^{N_e} z^2(k)}{N_e^3}+(N-N_e)\ln\frac{\sum_{k=N_e+1}^{N} z^2(k)}{(N-N_e)^3}\,, \tag{6}$$

because $L'_N(N_e) = L_N([1\ldots10\ldots0]) - N - 2N\ln N$, the global minimization of L_N is equivalent to the minimization of L'_N w.r.t. N_e.

A closed form expression for the minimum of this function could not be derived. Nevertheless, the computation of MGML hyperparameter estimates associated to one signal sample is extremely simple: it mainly requires the numerical evaluation of $L'_N(N_e)$ for $N_e\in\{0,1,\ldots,N-1\}$. The Monte Carlo study of Section 4.1 makes use of repeated minimizations of L'_N in order to efficiently compute the MGML estimates.

The next section is devoted to an asymptotic study for the estimator.

3 Asymptotic behavior

For each N, we are guaranteed at least a GML estimate denoted $\hat{\theta}_N$. This section is devoted to the limiting behavior of the series $(\hat{\theta}_N)$. By the means of a slight modification of Theorem 1 and Lemma found in [10], we prove that the convergence of $(\hat{\theta}_N)$ is linked to the existence and the uniqueness of a global minimum of an auxiliary function depending on a unique threshold variable. Then we prove the existence of such a minimum and we supply numerical experiments in order to support the conjecture of uniqueness and to study the bias of related GML estimates.

3.1 Limiting expressions

We first derive an asymptotic expression for equation (6). A straightforward derivation is not obvious and, following [10], we tackle this difficulty through the introduction of an auxiliary function of an explicit threshold variable $T \in [0, +\infty[$ defined by:

$$L''_N(T) = \lambda_N(T) \ln \frac{\sigma_N(T)}{\lambda_N^3(T)} + \bar{\lambda}_N(T) \ln \frac{\bar{\sigma}_N(T)}{\bar{\lambda}_N^3(T)} \tag{7}$$

$$\text{where } \lambda_N(T) = \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\{z^2(k) \geq T\}}, \ \ \sigma_N(T) = \frac{1}{N} \sum_{k=1}^{N} z^2(k) \mathbf{1}_{\{z^2(k) \geq T\}}, \tag{8}$$

$$\lambda_N(T) + \bar{\lambda}_N(T) = 1 \ \text{ and } \ \sigma_N(T) + \bar{\sigma}_N(T) = \hat{r}_z.$$

Because $N L''_N(T) = L'_N(N \lambda_N(T)) + 2N \ln N$, the optimization of L''_N w.r.t. T is equivalent to the optimization of L'_N w.r.t. N_e, but an asymptotic expression for L''_N is more easily derived.

Let Z denote a random variable distributed as $Z(1)$ for instance, after the strong law of large numbers, we obtain almost surely (a.s.):

$$\lim_{N \to \infty} \lambda_N(T) \overset{a.s.}{=} E\left[\mathbf{1}_{\{Z^2 \geq T\}}\right] \tag{9}$$

$$\lim_{N \to \infty} \sigma_N(T) \overset{a.s.}{=} E\left[Z^2 \mathbf{1}_{\{Z^2 \geq T\}}\right]. \tag{10}$$

$$\text{Thus } L''_N \text{ converges a.s. toward } L''_\infty(T) = \lambda_\infty(T) \ln \frac{\sigma_\infty(T)}{\lambda_\infty^3(T)} + \bar{\lambda}_\infty(T) \ln \frac{\bar{\sigma}_\infty(T)}{\bar{\lambda}_\infty^3(T)} \tag{11}$$

where the quantities subscripted by ∞ are limiting values corresponding to the quantities subscripted by N.

Theorem

Let $(\hat{\theta}_N)$ be any series of GML estimates and $(\hat{T}_N)$ denote the associated threshold series. Assume L''_∞ has a unique minimum $\hat{T}$, then $\lim_{N \to \infty} \hat{T}_N \overset{a.s.}{=} \hat{T}$ and $\lim_{N \to \infty} \hat{\theta}_N \overset{a.s.}{=} \hat{\theta}$ where:

$$\hat{\lambda} = \lambda_\infty(\hat{T}) \ , \ \hat{r}_n = \frac{\bar{\sigma}_\infty(\hat{T})}{\bar{\lambda}_\infty(\hat{T})} \ , \ \hat{r}_x = \frac{\sigma_\infty(\hat{T})}{\lambda_\infty(\hat{T})} - \frac{\bar{\sigma}_\infty(\hat{T})}{\bar{\lambda}_\infty(\hat{T})} \text{ and } \hat{\theta} \in \Delta. \tag{12}$$

This theorem relies on the following lemma:

Lemma

Let (f_n) be a series of real monotonous random equations defined on $[0, +\infty]$ and assume that $f_n(s)$ converges a.s. towards a function $f(s)$ continuous on $[0, +\infty]$. Then let (s_n) be any series such that $\lim_{n \to \infty} s_n = s_\infty$ exists in $[0, +\infty]$, $f_n(s_n)$ converges a.s. toward $f(s_\infty)$.

Proof of the lemma will not be given here, for it consists in a minor adaptation of the similar lemma in [10] required to handle infinity. Let $\hat{T}^+ \triangleq \limsup \hat{T}_N$, possibly infinite, then there is a strictly increasing function ψ on $\mathbb{N}$ such that $\lim_{N\to\infty} \hat{T}_{\psi(N)} = \hat{T}^+$. We have $L''_{\psi(N)}(\hat{T}_{\psi(N)}) \leq L''_{\psi(N)}(T)$,$\forall T \in [0,+\infty]$, applying the lemma to series $(\hat{T}_{\psi(N)})$ and to the monotonous random functions $\lambda_{\psi(N)}(.)$, $\bar{\lambda}_{\psi(N)}(.)$, $\sigma_{\psi(N)}(.)$ and $\bar{\sigma}_{\psi(N)}(.)$ we obtain $L''_\infty(\hat{T}^+) \leq L''_\infty(T)$, $\forall T \in [0,+\infty]$. Because L''_∞ is assumed to have a unique global minimum, $\hat{T}^+ = \hat{T}$. It can be proved similarly that $\liminf_N \hat{T}_N = \hat{T}$, so that $\lim_{N\to\infty} \hat{T}_N = \hat{T}$. Associated expressions for $\hat{\theta}$ follow from the limiting form of finite sample estimates (2)(3)(4).

3.2 Existence of a global minimum in Δ

As L''_∞ is a function of θ^*, we may write $L''_\infty(T,\theta^*)$. L''_∞ is easily expressed in terms of error functions. Careful computations show that:

$$L''_\infty(T,\theta^*) = \ln r_z^* - \frac{(1-\lambda^*)r_x^*}{(r_x^* + r_n^*)r_z^*}\sigma_\infty(T)(1+o(1)) \;\text{ when }\; T \to +\infty\ , \tag{13}$$

$$\lim_{T\to 0} L''_\infty(T,\theta^*) = \ln r_z^* \;\text{ where }\; r_z^* = \lambda^* r_x^* + r_n^*. \tag{14}$$

$\sigma_\infty(T)$ is positive strictly decreasing from r_z^* to zero, then $\lim_{T\to+\infty} L''_\infty(T,\theta^*) = \ln r_z^*$. As $L''_\infty(T,\theta^*)$ is a continuous function of T, and admits finite limits on its boundaries $L''_\infty(T,\theta^*)$ is a bounded function of T. Moreover Equation (13) means that $L''_\infty(T,\theta^*) < \ln r_z^*$ for T large enough. This inequality guarantees the existence of a global minimum $\hat{T} \in]0,+\infty[$. Up to date, no proof for the uniqueness has been found because of the tedious analytical expression for the derivative of L''_∞. However, practical studies of L''_∞ for values of θ^* scattered over Δ support the assumption of uniqueness.

Further study of the asymptotic bias can be performed numerically only. Nevertheless, some properties valid for all θ^* may be shown that are useful for the numerical study. First, it may be easily checked that $\hat{\lambda}\hat{r}_x + \hat{r}_n = \lambda^* r_x^* + r_n^* = E\left[Z^2\right]$. Moreover we have a kind of homogeneity relationship:

$$L''_\infty(T,\lambda^*,r_x^*,r_n^*) = L''_\infty(\alpha T,\lambda^*,\alpha r_x^*,\alpha r_n^*) - \ln\alpha \ \ ,\forall \alpha > 0$$

which enables us to assess a scale invariance property for the estimators:

$$\begin{aligned}
\hat{\lambda}(\lambda^*,\alpha r_x^*,\alpha r_n^*) &= \hat{\lambda}(\lambda^*,r_x^*,r_n^*) \\
\hat{r}_x(\lambda^*,\alpha r_x^*,\alpha r_n^*) &= \alpha\hat{r}_x(\lambda^*,r_x^*,r_n^*) \\
\hat{r}_n(\lambda^*,\alpha r_x^*,\alpha r_n^*) &= \alpha\hat{r}_n(\lambda^*,r_x^*,r_n^*)
\end{aligned}$$

The estimates obtained in [10] do not satisfy these properties.

4 Numerical experiments

4.1 Finite sample ML and MGML estimates

10000 independant samples where drawn from a single univariate Gaussian mixture of parameters $\lambda^* = 0.1$, $r_x^* = 100$ and $r_n^* = 1$. This corresponds to a "signal-to-noise ratio" (SNR) of 10dB where the SNR is defined as $10\log(\lambda^* r_x^*/r_n^*)$. The ratio and the tested values are standard in the context of BG deconvolution. The samples were gathered by N in order to study the statistical behavior (bias and mean square error (MSE)) of estimates based on samples of size N. N spans $\{5, 6, \ldots, 70\}$, at $N = 70$ the ML and MGML estimators exibit an asymtotic behavior.

MGML estimates are easily computed using N computations of function $L'_N(N_e), N_e = 1 \ldots N-1$ (see Section 2). Computation of corresponding ML estimates required much more effort.

Previous work on ML estimation for two-component Gaussian mixtures [12][16][13] reports to main issues: first, the undboundedness of the likelihood and second, the existence spurious local maximas.

The first problem does not occur here, because we do not try to estimate the means of the Gaussians, they are assumed zero in this study and the likelihood considered here is truely bounded above.

In order to deal with the second problem and to get more reliable ML estimates we proceed in two steps. First we compute the log-likelihood on a grid spanning the parameter space, then we start an EM procedure [12][11] until the norm of the gradient decreases under 10^{-8}, where we may consider the gradient vanishes. A further computation of the Hessian is performed for the obtained estimates in order to insure that at least a local maximum has been attained.

The first step of search for a maximum on a coarse grid over Δ has been restricted to a search on $(\lambda, r_n) \in [0,1] \times [0, \hat{r}_z]$ by the means of the following identity:

$$\tilde{\lambda}\tilde{r}_x + \tilde{r}_n = \hat{r}_z,$$

where tilded quqntities reffer to ML estimates. That identity derives from expressions of the derivative of the log-likelihood which vanishes at a maximum. To our knowledge this identity has neither been pointed out nor used explicitly in previous work, but it is implicit in the EM algorithm. Making use of this identity the search can be restricted to the surface described by $\lambda r_x + r_n = \hat{r}_z$ which contains the ML estimate. Then the likelihood is expressed in terms of the two dependent variables $\lambda \in [0,1]$ and $r_n \in [0, \hat{r}_z]$.

Figure 1 (resp. Figure 2) summarize the results relative to mean estimates (top figure) and MSE (bottom figure) for λ^* (resp. r_n^*) expressed as a function of N the size of the considered sample. Concerning the parameter λ the MGML performs better than ML until $N = 50$ in terms of bias an MSE. Then the asymtotic behaviour of ML begins to take over MGML in terms of bias. These conclusions are almost identical for parameter r_n, except for the MSE of very small samples ($N = 5 \ldots 10$) where MGML is much higher than ML.

It should be pointed out, however, that the relative MSE for parameter r_n is always smaller than for the other parameters, that means that r_n^* is always better estimated than λ^* and r_x^*. To a certain extend these results are consistent with previous reports of empirical success of Generalized Likelihood approaches.

4.2 Asymptotic MGML estimates

The domain Δ of θ^* is sampled, and for each value of θ^* on the grid the corresponding estimates of θ^* are computed using on one hand a numerical minimization of $L''_\infty(T, \theta^*)$, and the identity (12) on the other hand. The scale invariance property enables us to restrict the study to a grid in (λ^*, r_x^*) only and set $r_n^* = 1$ for instance.

The numerical optimization of $L''_\infty(T, \theta^*)$ w.r.t. T has been performed using an exhaustive search on a grid as a first step. This rough optimum is then refined by a fixed point method.

The graphs presented on Figure 3, correspond to a "signal-to-noise ratio" (SNR) of 10dB. Values of λ^* are regularly sampled between 0.01 and 0.4. For the sake of clarity, we have reported only the results corresponding to the estimates of λ versus the true value λ^*.

The graphs compare an asymptotically unbiased estimator like ML, the GML estimator and the GL estimator of Gassiat *et al.*. Because the latter does not exhibit any scale invariance property we represented two graphs of GL estimates corresponding to $r_n^* = 1$ and $r_n^* = 0.1$. GL and GML estimates show a systematic bias, λ is always under-estimated. And it should be stressed that GL estimates do not always exist as shown on the graph for $r_n^* = 0.1$. The bias is moderate for small λ^*, and cannot be neglected otherwise. However, the estimates remain significant, at least for the chosen SNR. Further studies reported in [15] show that increasing the SNR diminishes the bias.

5 Conclusion

The question of the relevance of generalized likelihood techniques in the context of BG myopic deconvolution led us to the detailed study of a simpler problem, namely the MGML identification of a two-component Gaussian mixture.

The results obtained on this MGML estimator relieves some of the criticism against MGL estimation that concluded a former paper. In particular, we prove the existence of finite sample MGML estimates and propose an algorithm that computes exactly these estimates. Moreover the associated numerical cost that can be neglected compared to ML estimation. In the case of small samples, a Monte Carlo experiment shows that MGML can enjoy a smaller bias and MSE than ML.

Considering asymptotic properties, the convergence of MGML estimates is assessed under a reasonable assumption. A further numerical experiment supports this assumption and quantifies the asymptotic bias of MGML estimates. This bias may indeed range from moderate to large but corresponding estimates remain significant.

In the broader context of filtered mixtures, GL-like criterions have been used mainly for *practical* purposes, and showed a seemingly practical success. A common setback associated to GL methods is the impossibility to assess the existence of estimates because GL criterions are not bounded above and a local maxima may not exist. We believe that the use of GML criterions could be a satisfactory practical answer to this problem when consistent schemes cannot be implemented, provided that their finite sample behavior has been more seriously investigated.

References

[1] J. M. Mendel. *Optimal seismic deconvolution.* Academic Press, New York, 1983.

[2] G. Demoment, R. Reynaud, and A. Herment. Range resolution improvement by a fast deconvolution method. *Ultrasonic Imaging*, 6:435–451, 1984.

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

[4] Y. Bar-Shalom. Optimal simultaneous state estimation and parameter identification in linear discrete-time systems. *IEEE Trans. Automatic Control*, AC-17:308–319, 1972.

[5] S. Lakhsmanan and H. Derin. Simultaneous parameter estimation and segmentation of gibbs random fields using simulated annealing. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-11:799–813, 1989.

[6] P. Bryant and J. Williamson. Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika*, 65:273–281, 1978.

[7] A. C. Harvey and S. Peters. A note on the estimation of variance in state-space models using the maximum *a posteriori* procedure. *IEEE Trans. Automatic Control*, AC-30:1048–1050, 1985.

[8] J. Goutsias and J. M. Mendel. Maximum-likelihood deconvolution: an optimization theory perspective. *Geophysics*, 51:1206–1220, 1986.

[9] Y. Goussard. *Déconvolution de processus aléatoires non-gaussiens par maximisation de vraisemblances.* PhD thesis, Université de Paris-Sud, Centre d'Orsay, 1989.

[10] E. Gassiat, F. Monfront, and Y. Goussard. On simultaneaous signal estimation and parameter identification using a generalized likelihood approach. *IEEE Trans. Information Theory*, 38:157–162, 1992.

[11] R. Redner and H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26:195–239, 1984.

[12] D. Titterington, A. F. M. Smith, and U. Makov. *Statistical analysis of finite mixture distributions.* Wiley, New-York, 1985.

[13] N. M. Kiefer. Discrete parameter variation: efficient estimation of a switching regression mode. *Econometrica*, 46:427–434, 1978.

[14] R. E. Quandt and J. B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *J. Am. Stat. Ass.*, 73:730–738, 1978.

[15] F. Champagnat. *Déconvolution impulsionnelle et extensions pour la caractérisation des milieux inhomogènes en échographie.* PhD thesis, Université de Paris-Sud, centre d'Orsay, 1993.

[16] R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.*, 13:795–800, 1983.

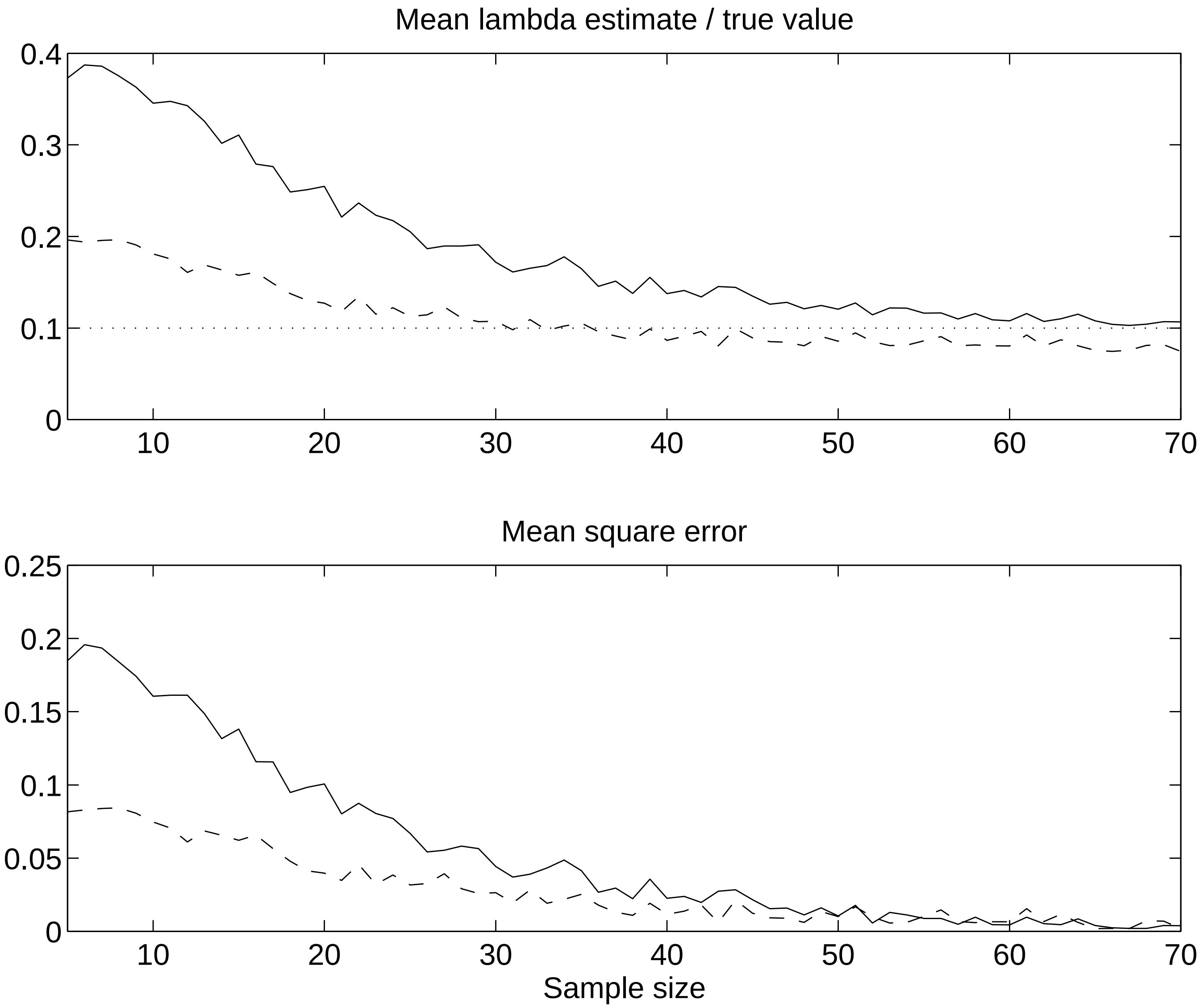

Figure 1: Comparison of finite sample estimates for parameter λ. Mean λ estimates (top figure) and mean square error (bottom figure) are represented as functions of sample size. $(\ldots)$: True value $\lambda^* = 0.1$. (- -) MGML estimation. (—) ML estimation. The comparative asymptotic behaviour of the estimators is best seen on the top figure where ML converges towards the true value whereas MGML converges to a slighly inferior value. Conversely, MGML performs better both in bias and MSE for sample sizes lower than 50.

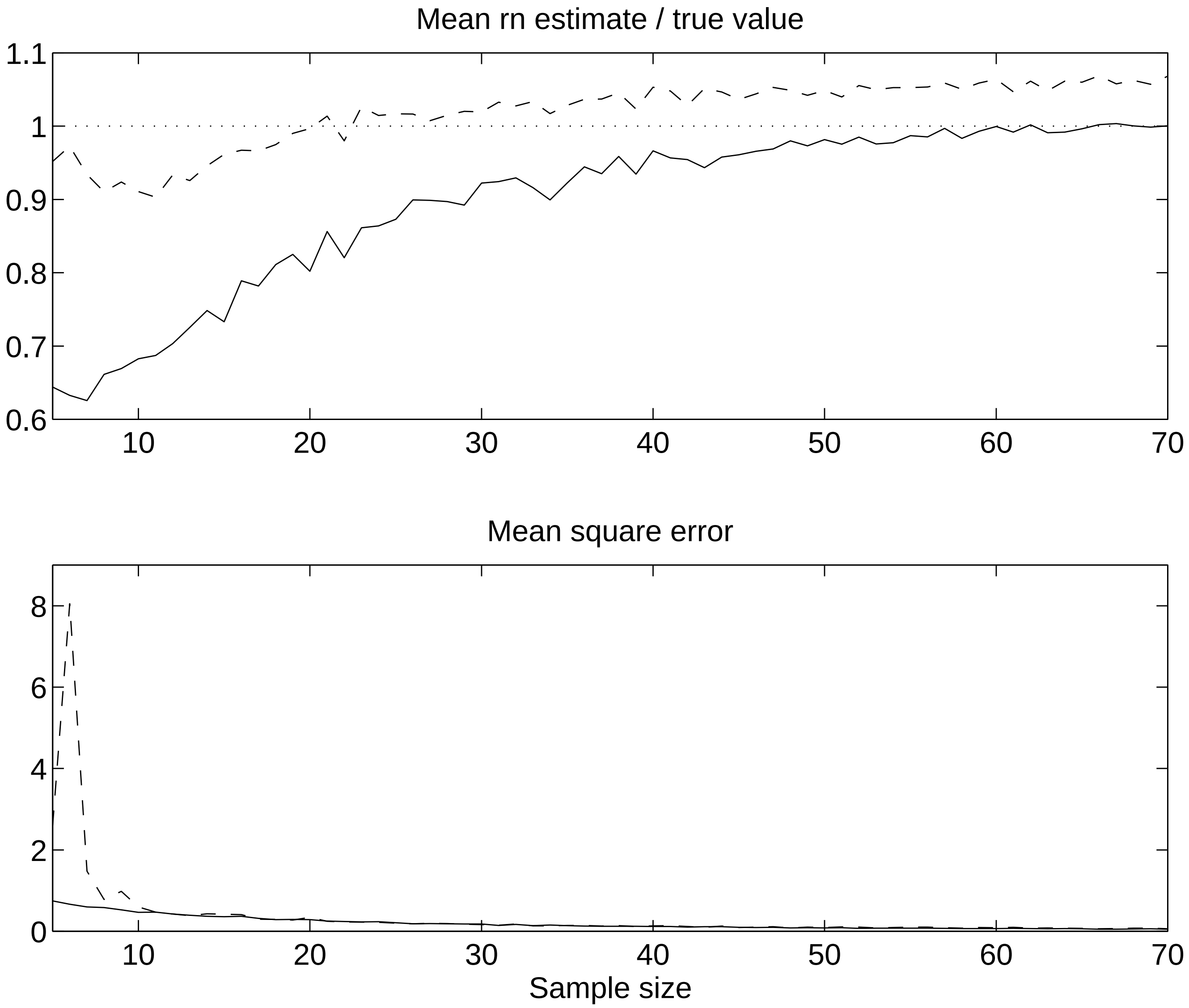

Figure 2: Comparison of finite sample estimates for parameter r_n. Mean r_n estimates (top figure) and mean square error (bottom figure) are represented as functions of sample size. $(\ldots)$: True value $\lambda^* = 0.1$. (- -) MGML estimation. (—) ML estimation. The comparative asymptotic behaviour of the estimators is best seen on the top figure where ML converges towards the true value whereas MGML converges to a slighly greater value. MGML performs better in bias for sample sizes lower than 40. MGML compares well in terms of MSE for sample sizes over 10, however for samples of smaller size the MSE of MGML ranges from 5 times to 10 times larger than that of ML, due to variance.

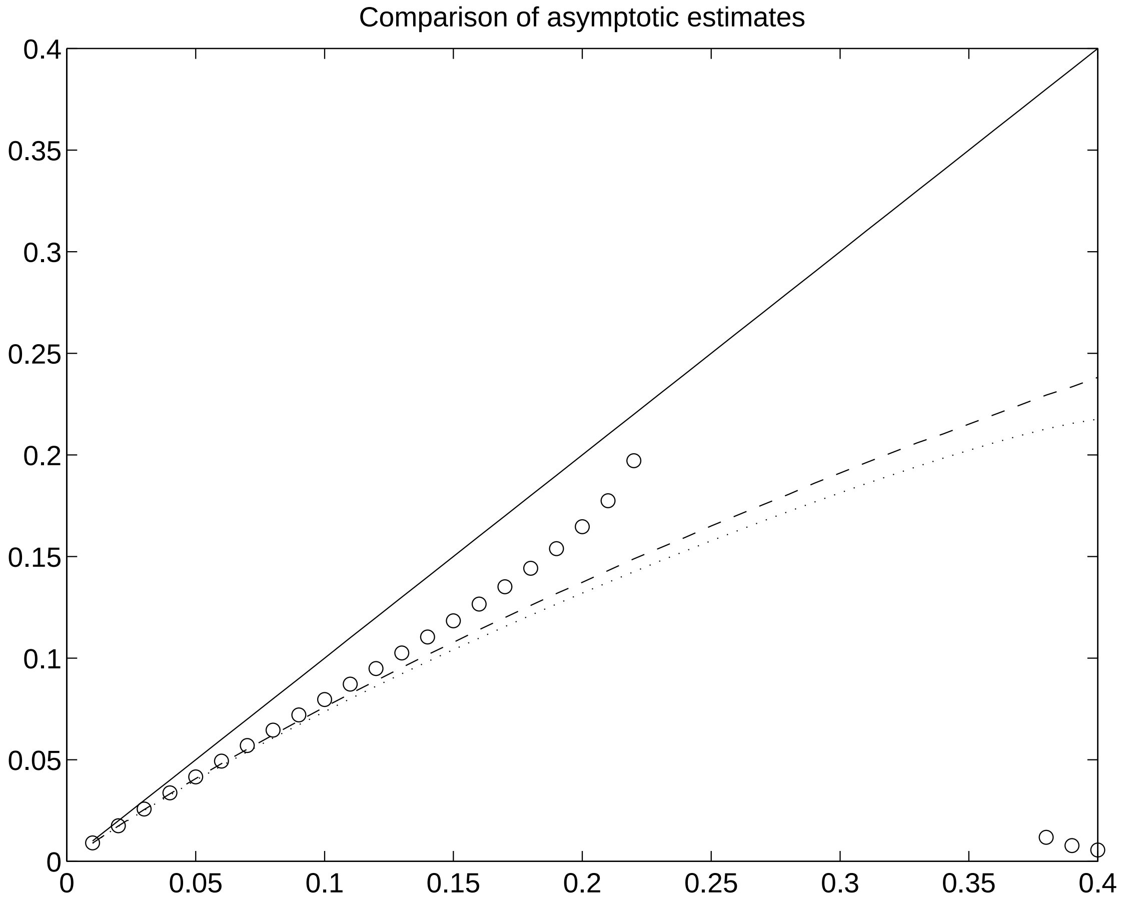

Figure 3: Different asymptotic estimates of λ versus λ^*, keeping $SNR = 10\log(\lambda^* r_x^*/r_n^*) = 10$. The estimators are systematically biased, but the estimates remain significant. (—) True λ. (- -) GML estimates. (...) GL estimates for $r_n^* = 1$. (∘∘∘) GL estimates $r_n = 0.1$. Note that the last curve is interrupted due to non existence of corresponding estimates.