AN ALTERNATIVE TO STANDARD MAXIMUM LIKELIHOOD FOR GAUSSIAN MIXTURES

Frédéric Champagnat

Institut de génie biomédical
École Polytechnique
P.O. Box 6079, Station "Centre-ville"
Montréal, Québec, H3C 3A7 Canada
email : champagnat@grbb.polymtl.ca

Jérôme Idier

Laboratoire des signaux et systèmes
École Supérieure d'Électricité
Plateau de Moulon
91192 GIF-SUR-YVETTE Cédex, France
email: idier@lss.supelec.fr

ABSTRACT

Because true *Maximum Likelihood* (ML) is too expensive, the dominant approach in Bernoulli-Gaussian (BG) myopic deconvolution consists in the joint maximization of a single *Generalized Likelihood* with respect to the input signal *and* the hyperparameters. This communication assesses the theoretical properties of a related *Maximum Generalized Marginal Likelihood* (MGML) estimator in a simplified framework: the filter is reduced to identity, so that the output data is a mixture of Gaussian populations. Our results are three-fold: first, *exact* MGML estimates can be efficiently computed; second, this estimator performs better than ML in the short sample case whereas it is drastically less expensive; third, 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 seismic exploration, non-destructive evaluation and biomedical engineering [1].

Such problems are classically dealt with a discrete-time convolution model for the observations: $z = h * r + n$. h is the filter, n is a stationary white Gaussian noise with variance r_n and r is the input to be restored. The filter h is assumed known hereafter. The *ill-posed* nature of the induced deconvolution problem may be coped within a Bayesian framework: prior information about the spiky structure of the input is introduced in the form of a prior probability model. Here we model the input r as the observable part of a Bernoulli-Gaussian process (BG) [1]. BG models may be seen as discrete-time compound processes (q, r). q and r model the time location for a spike, and its amplitude, respectively. A BG process is made up of independent samples, each one being defined as a pair of random variables (RVs) $X = (Q, R)$. Q is a Bernoulli variable such that $\lambda \triangleq \Pr(Q = 1)$ is the probability of occurrence of the spike. R is a zero-mean Gaussian RV with variance Qr_x. Thus the probability distributions associated with the problem are controlled by the vector of *hyperparameters* $\boldsymbol{\theta} = (\lambda, r_x, r_n)$. We address the practical problem of hyperparameter identification.

Up to now, Generalized Likelihood (GL) maximization has been the dominant method in BG deconvolution problems [1] mainly because of its practicability. GL estimation corresponds to the maximization of the joint likelihood $p(z, r, q, \boldsymbol{\theta})$ with respect to (w.r.t.) r, q and $\boldsymbol{\theta}$. Such methods have been successfully implemented in various areas [2, 3] but are generally disregarded because of their non-consistency [2]. However consistency is relevant only for large-sample signals and it is of no critical importance in short-data sets applications.

Gassiat *et al.* [4] presented a theoretical study of the maximum GL (MGL) estimator when the filter is reduced to a delta function in order to be able to carry out mathematical derivations. Then the output signal is a mixture of two zero-mean univariate Gaussian distributions. Estimation of the parameters governing a Gaussian mixture is yet a well documented area for consistent estimators such as ML [5]. But GL techniques (also referred to as *classification likelihood* methods [2]) are considered as *ad-hoc* techniques and are far less documented.

The results of Gassiat *et al.* [4] established the poor behavior of the GL criterion, in particular the inability to ensure existence of MGL estimates. Conversely, when estimates exist they may exhibit a small bias.

The aim of this correspondence is to provide, in the same context, an original statistical justification of one alternative methodology based on a Generalized Marginal Likelihood (GML) $p(z, q, \boldsymbol{\theta})$ w.r.t. q and $\boldsymbol{\theta}$, where amplitudes of the spikes have been "integrated out".

The conclusions of this study on MGML estimation qualify those drawn by Gassiat *et al.* on GL criterion: in the finite-sample case, existence of a global maximum for the GML is assessed and an efficient algorithm for *exact* maximization with a finite number of computations is derived. Unlike MGL estimates, MGML estimates possess an interesting *scale invariance property* (SIP). A presented Monte Carlo experiment shows that MGML estimation exhibit smaller bias and mean square error than ML estimation for small samples. Furthermore, the associated computational load is much lighter than that of ML estimation. Finite sample estimates converge toward the global maxi-

Part of this work was performed as the first author was in Ph. D. at Laboratoire des Signaux et Systèmes

0-7803-2431-5/95 $4.00 © 1995 IEEE

mum of the asymptotic GML under the reasonable assumption of uniqueness of this maximum. A further numerical experiment shows that the MGML estimator is not consistent and that asymptotic bias ranges from moderate to large, depending on the amount of noise and the density λ of the pulse process.

2. PROBLEM STATEMENT

2.1. Formulation as a mixture problem

In the absence of distortion, the input-output equation reduces to a spike process corrupted by an additive noise $\mathbf{Z} = \mathbf{R} + \mathbf{N}$. It turns out that $(Z_k \mid Q_k = q)$ is a zero-mean Gaussian RV of variance $qr_x + r_n$, in other words Z_k is a mixture of two univariate zero-mean Gaussian RVs. Let z denote a sample drawn after the distribution of $\mathbf{Z}^*$ controlled by the so-called "true" parameters $\theta^* = (\lambda^*, r_x^*, r_n^*)$. We assume these parameters belong to $\Theta \triangleq]0,1[\times]0,+\infty[^2$, and we address the problem of estimating θ^* on the basis of the sample $z = [z_1, \ldots, z_N]$.

Although estimation of the parameters of a Gaussian mixture has drawn a quantity of works in the statistical field [5, 6], we are not aware of any results specific to the problem addressed here. Beyond the methods available in the literature, great emphasis has been put on ML estimation [5]. Before proceeding on the GML estimation general and particular results pertaining to ML estimation are recalled.

2.2. Background on ML estimation

Let $f(z;r)$ denote the density of a univariate zero-mean Gaussian RV of variance r, then the ML estimate $\tilde{\theta}$ is the argument of the maximum of $p_{\mathbf{Z}}(z;\theta)$ when θ spans Θ, which takes the form:

$$p_{\mathbf{Z}}(z;\theta) = \prod_{i=1}^{N} \left(\lambda f(z_i, r_x + r_n) + (1-\lambda) f(z_i, r_n)\right). \quad (1)$$

$p_{\mathbf{Z}}(z;\theta)$ is bounded above and admits a global maximum w.r.t. θ [7], but local maxima might exist [6].

In the wider context of Gaussian mixtures the ML yields consistent and asymptotically efficient estimates, provided some regularity conditions on the likelihood be satisfied [5]. The EM algorithm [5] can be implemented for likelihood optimization in a very simple and heuristic manner, moreover it guarantees a monotonic increase of the likelihood and at each iteration it maintains the parameters inside the domain of definition. This feature is not shared by the general optimization tools such as gradient and Newton methods and can be important in practice. However, convergence of the EM algorithm is guaranteed only to a stationary point of the likelihood and it can be dramatically slow as was experienced during our simulations.

2.3. MGML estimation

The GML criterion is defined by

$$L_{\mathrm{GML}}(q,\theta) \triangleq p_{\mathbf{Z},\mathbf{Q}}(z,q;\theta) = p_{\mathbf{Z}|\mathbf{Q}}(z|q;r_x,r_n)\,\Pr(\mathbf{Q}=q;\lambda).$$

The MGML estimate $(\hat{q},\hat{\theta})$ is the argument of the maximum of L_{GML} when (q,θ) spans $\{0,1\}^N \times \Theta$. Then it can be easily shown that:

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

Comparison of the latter expression with (1) shows that 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. Conversely, the Monte Carlo experiments of Section 4.1 demonstrate that the MGML estimator yields smaller bias and MSE than ML in the finite sample case.

3. RESULTS ON GML ESTIMATION

In this section we state without demonstration[1] the main results pertaining to MGML estimation both in the finite sample case and the asymptotic limit.

3.1. Finite sample properties

We may consider that $z_1^2 \geq z_2^2 \geq \ldots \geq z_N^2$ at the expense of a swap of subscripts. Then we have the following theorem

Theorem 1 : *Let $J_N(n)$ be defined on $\{0,1,\ldots,N\}$ by*[2]

$$J_N(n) \triangleq n \ln \frac{\sum_{k=1}^{n} z_k^2}{n^3} + (N-n) \ln \frac{\sum_{k=n+1}^{N} z_k^2}{(N-n)^3}, \quad (2)$$

then the MGML estimate $\hat{\theta}$ exists almost surely (a.s.). It is given by:

$$\hat{\lambda} = \frac{N_e}{N},\ \hat{r}_n = \frac{\sum_{k=N_e+1}^{N} z_k^2}{N - N_e},\ \hat{r}_x = \frac{\sum_{k=1}^{N_e} z_k^2}{N_e} - \hat{r}_n$$

$$\textit{where: } N_e = \arg\min_{0\leq n\leq N} J_N(n).$$

A closed form expression for N_e could not be derived. Nevertheless, the computation of $\hat{\theta}$ associated to one signal sample z is extremely simple: it mainly requires the numerical evaluation of a simple function for $n \in \{0,1,\ldots,N\}$ whereas ML admits no exact computation involving a finite number of operations. The Monte Carlo study of Section 4.1 makes intensive use of Theorem 1 in order to efficiently compute MGML estimates.

Expression (2) enables us to assess easily a SIP for the MGML estimator, unlike the MGL estimator of [4]: $J_N, \hat{\lambda}, \hat{r}_x, \hat{r}_n$ and N_e depend implicitly on z, what happens if z is replaced by αz where $\alpha > 0$ is an arbitrary "scale factor" ? It can easily be seen from (2) that $J_N(n, \alpha z) = J_N(n, z) - 2N \ln \alpha$; then $J_N(n, \alpha z)$ and $J_N(n, z)$ have the same minimum. The SIP property follows immediately:

$$\hat{\lambda}(\alpha z) = \hat{\lambda}(z),\ \hat{r}_n(\alpha z) = \alpha^2 \hat{r}_n(z) \text{ and } \hat{r}_x(\alpha z) = \alpha^2 \hat{r}_x(z).$$

[1] Full developments can be found in [7]

[2] With the convention $0 \ln \frac{0}{0} = 0$

3.2. Asymptotic behavior

For all N exists a MGML estimate denoted $\hat{\boldsymbol{\theta}}_N$. This section examines the limiting behavior of the series $(\hat{\boldsymbol{\theta}}_N)$. As stated in the following Theorem 2 convergence of $(\hat{\boldsymbol{\theta}}_N)$ is linked to the existence and the uniqueness of a global minimum of function J_∞ of a unique threshold variable $T \in]0, +\infty[$:

$$J_\infty(T) \triangleq \lambda_\infty(T) \ln \frac{\sigma_\infty(T)}{\lambda_\infty^3(T)} + \bar{\lambda}_\infty(T) \ln \frac{\bar{\sigma}_\infty(T)}{\bar{\lambda}_\infty^3(T)}. \quad (3)$$

where $\lambda_\infty(T) \triangleq E\left[1_{\{Z^2 \geq T\}}\right]$, $\sigma_\infty(T) \triangleq E\left[Z^2 1_{\{Z^2 \geq T\}}\right]$, $\bar{\lambda}_\infty(T) \triangleq 1 - \lambda_\infty(T)$, $\bar{\sigma}_\infty(T) \triangleq E\left[Z^2\right] - \sigma_\infty(T)$ and Z denote a random variable distributed as Z_1^* for instance.

Theorem 2 *: Let $(\hat{\boldsymbol{\theta}}_N)$ be any series of MGML estimates. Assume J_∞ has a unique minimum $\hat{T}$, then* $\lim_{N\to\infty} \hat{\boldsymbol{\theta}}_N \overset{a.s.}{=} \hat{\boldsymbol{\theta}} \in \Theta$ *where:*

$$\hat{\lambda} = \lambda_\infty(\hat{T}), \quad \hat{r}_n = \frac{\bar{\sigma}_\infty(\hat{T})}{\bar{\lambda}_\infty(\hat{T})}, \quad \hat{r}_x = \frac{\sigma_\infty(\hat{T})}{\lambda_\infty(\hat{T})} - \frac{\bar{\sigma}_\infty(\hat{T})}{\bar{\lambda}_\infty(\hat{T})}. \quad (4)$$

J_∞ admits at least a global minimum $\hat{T} \in]0, +\infty[$ [7]. Up to date, no proof for the uniqueness has been found because of the tedious analytical expression for the derivative of J_∞. However, practical studies of J_∞ for values of $\boldsymbol{\theta}^*$ scattered over Θ support the assumption of uniqueness. Further study of the asymptotic bias can be performed numerically only, corresponding results are reported in section 4.2.

4. NUMERICAL EXPERIMENTS

4.1. Finite sample ML and MGML estimates

The mean estimate and mean square error (MSE) were computed for three data sets whose features are gathered in Table 1.

Due to the SIP we may keep the variance $r_n^* = 1$ and let the remaining parameters vary. The label "SNR" in Table 1 stands for "signal-to-noise ratio" which is defined as $10\log(\lambda^* r_x^*/r_n^*)$. The SNR indicates how difficult the problem is. The 10 dB SNR and $\lambda^* = 0.1$ parameters for set A are standard in the context of BG deconvolution. The samples were gathered by N within each set in order to study the statistical behavior (bias and MSE) of estimates based on samples of size N. The different graphs represent bias or MSE as a function of N.

Whereas MGML estimates are obtained quickly using the results of Section 3.1, computation of ML estimates is much more demanding. In order to deal with potential local maxima of the likelihood we proceed in two steps. First the likelihood is computed on a grid spanning the parameter space, then the maximum over the grid initiates an EM algorithm [5, 6].

Figure 1 summarizes the results relative to mean estimates (top figure) and MSE (bottom figure) for λ of set A. MGML performs better than ML until $N = 50$ in terms of both bias an MSE. Then asymptotic behavior of ML takes over MGML in terms of bias.

Figure 2 compares the performances of ML and MGML in terms of their "Total relative MSE". It is defined by $E\left[(\hat{\lambda}/\lambda^* - 1)^2 + (\hat{r}_x/r_x^* - 1)^2 + (\hat{r}_n/r_n^* - 1)^2\right]$ in order to account for the different parameter scales. Figure 2 indicates that the statistical behavior of MGML improves when λ^* decreases and when the SNR increases.

To a certain extend these results are consistent with previous reports of empirical success of GL-type approaches, and suggest they would perform better in the frequent context of small data set versus good contrast.

4.2. Asymptotic MGML estimates

The graphs on Figure 3 compare the performances of an asymptotically unbiased estimator like ML, the MGML estimator and the MGL estimator of Gassiat *et al.*, for different values of $\boldsymbol{\theta}^*$. For each value of $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}$ is computed using first a numerical minimization of $J_\infty(T, \boldsymbol{\theta}^*)$, and second the identity (4). The SNR is 10 dB, λ^* spans 0.01 and 0.4. For the sake of clarity, only the estimates of λ versus the true value λ^* are reported.

Because the MGL estimator does not exhibit any SIP two graphs of MGL estimates corresponding to $r_n^* = 1$ and $r_n^* = 0.1$ were presented. MGL and MGML estimates show a systematic negative bias. The bias is moderate for small λ^*, and cannot be neglected otherwise. However, the estimates remain significant, at least for the chosen SNR. MGL estimates do not always exist as shown on the graph for $r_n^* = 0.1$. Further results reported in [8] show that increasing the SNR diminishes the bias.

5. CONCLUSION

The question of relevance of GL techniques for BG myopic deconvolution led us to the study of a simpler problem, namely the MGML identification of a Gaussian mixture.

The results obtained on this MGML estimator alleviate some of the setbacks of a former MGL estimator [4]. In particular, MGML estimates always exist and enjoy a SIP. An algorithm for exact MGML estimation is derived and it is much faster than classical ML estimation methods. A presented Monte Carlo experiment shows that MGML should perform better than ML in the frequent context of small data set and good contrast.

The asymptotic convergence of MGML estimates is assessed under a reasonable assumption. A further numerical experiment quantifies the asymptotic bias of MGML estimates. This bias ranges from moderate to large but corresponding estimates remain significant.

In the broader context of BG deconvolution, GL-type criterions have been used mainly for *practical* purposes, and showed practical success. However MGL estimates do not always exist because GL criterions are not bounded above and a local maxima may not exist. MGML estimation provides a satisfactory answer to this problem.

6. REFERENCES

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

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

[3] 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.

[4] 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.

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

[6] D. M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions.* Wiley, Chichester ; Toronto, 1985.

[7] F. Champagnat and J. Idier. Generalized marginal likelihood for gaussian mixtures. LSS Internal Report GPI-94-01, 20 p., 1994.

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

Set	Total sample size	λ^*	r_x^*	r_n^*	SNR
A	10^4	0.1	100	1	10 dB
B	10^5	0.01	1000	1	10 dB
C	3.10^4	0.1	31.2	1	5 dB

Table 1: Features of tested data sets.

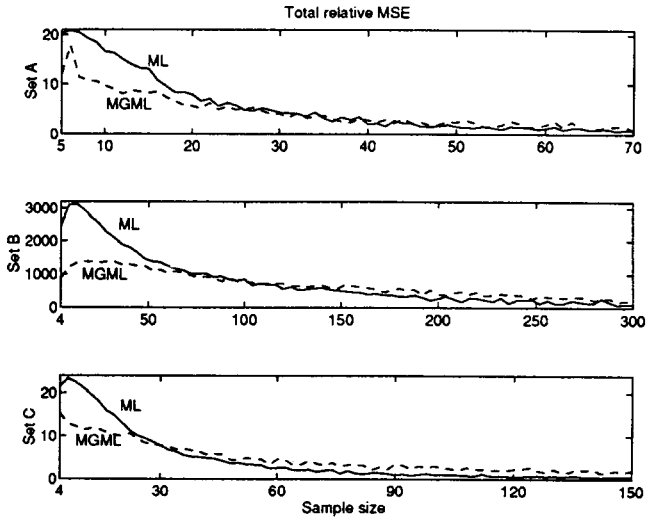

Figure 2: Comparison of total relative MSE for sets A (top), B (middle) and C (bottom) versus sample size. MGML always performs better than ML for small samples and ML takes over MGML in the asymptotic limit. The range where MGML remains competitive increases when the SNR increases and when λ^* decreases i.e. when the contrast is improved.

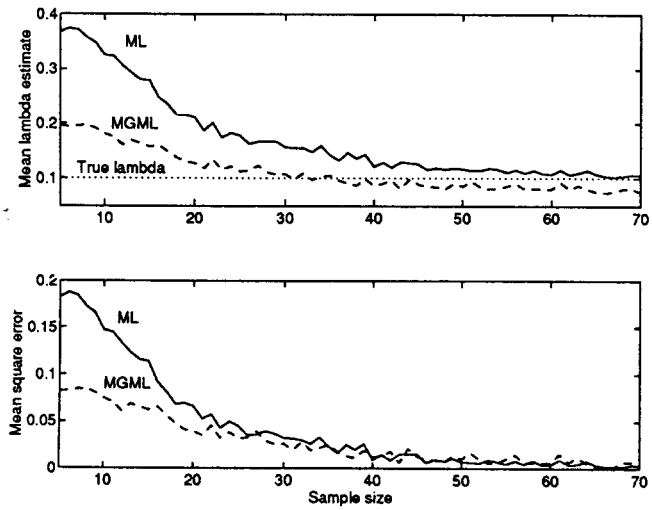

Figure 1: Bias (top) and MSE (bottom)for parameter λ of set A versus sample size.

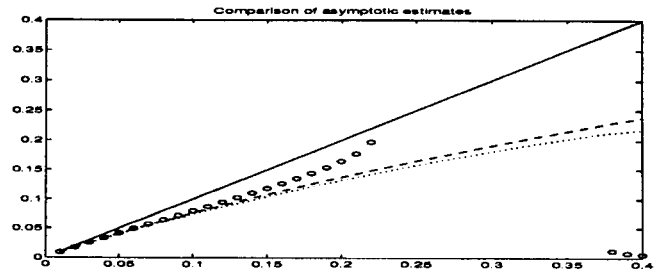

Figure 3: Different asymptotic estimates of λ versus λ^*, keeping $SNR = 10\log(\lambda^* r_x^*/r_n^*) = 10$. The estimators are systematically biased, but the estimates remain significant. (—) True λ. (- -) GML estimates. (...) GL estimates for $r_n^* = 1$. (ooo) GL estimates $r_n^* = 0.1$. Note that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