JOINT ESTIMATION OF PARAMETERS AND HYPERPARAMETERS IN A BAYESIAN APPROACH OF SOLVING INVERSE PROBLEMS

Ali Mohammad–Djafari

Laboratoire des Signaux et Systèmes (CNRS–SUPELEC–UPS)
École Supérieure d'Électricité
Plateau de Moulon, 91192 Gif–sur–Yvette Cedex, France

ABSTRACT

In this paper we propose a joint estimation of the parameters and hyperparameters (the parameters of the prior law) when a Bayesian approach with Maximum Entropy (ME) priors is used to solve the inverse problems which arise in signal and image reconstruction and restoration problems. In particular we propose two methods: one based on the Expectation Maximization (EM) algorithm who aims to find the Marginalized MAP (MMAP) estimate and the second based on a joint MAP estimation (JMAP). We discuss and compare these methods and give some simulation results in image restoration to show the relative performances of the proposed methods.

1. INTRODUCTION

In many signal and image reconstruction and restoration problems, the observed data $g(\boldsymbol{s}_n)$ are related to the quantities of the direct physical interest $f(\boldsymbol{r})$ by a linear transformation:

$$g(\boldsymbol{s}_m) = \int_D f(\boldsymbol{r}) h_m(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} + b(\boldsymbol{s}_m),\ m = 1, \ldots, M, \quad (1)$$

where $h_m(\boldsymbol{r})$ is the instrument function which we assume to be known. When this integral equation is discretized, we have to solve a linear system of equations:

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{b}, \quad (2)$$

where $\boldsymbol{H}$ is a matrix whose components depend on h_m. $\boldsymbol{H}$ has, in general, very large dimensions and is very ill-conditionned. $\boldsymbol{f}$ is a vector whose components represents the unknown parameters to estimate (for example the pixel values of an image). $\boldsymbol{g}$ is a vector whose components are the observed data, and $\boldsymbol{b}$ is a vector whose components represents both the measurement noise and any other unmodeled error.

The objective of an inversion procedure is to obtain an unique and stable solution $\widehat{\boldsymbol{f}}$ to this *ill–posed* problem. This can be achieved by introducing prior knowledge on $\boldsymbol{f}$ and on $\boldsymbol{b}$ by adopting a Bayesian approach which consists in translating our prior knowledge on $\boldsymbol{f}$ and $\boldsymbol{b}$ by the probability laws $p(\boldsymbol{f}|\boldsymbol{\theta})$ and $p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\beta})$ and using the Baye's rule to obtain $p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\beta}, \boldsymbol{\theta})$. The final step then is to choose a decision rule (for example Maximum *a posteriori* MAP) to determine a solution $\widehat{\boldsymbol{f}}$ to the problem.

One of the difficulties in this bayesian approach is to assign the prior law $p(\boldsymbol{f}|\boldsymbol{\theta})$ in a way to reflect our prior knowledge about the solution $\boldsymbol{f}$. In cases where the prior knowledge is in the form of expectations then maximum entropy principle (MEP) can give us the solution. However, in general, no matter how $p(\boldsymbol{f}|\boldsymbol{\theta})$ is assigned, it may depend on some unknown parameters $\boldsymbol{\theta}$, called *hyperparameters*, so that, in the inversion procedure we want to estimate also them from the data. To be more specific, let us describe a method that we developed in preceeding works [1, 2, 3] which can be resumed as follows:

• Information about noise $\boldsymbol{b}$ is the covariance matrix $\boldsymbol{W} = \sigma^2\boldsymbol{I} = (2/\beta)\boldsymbol{I}$. So that, using this only information with MEP we obtain :

$$p(\boldsymbol{g}|\boldsymbol{f}, \beta) \propto \exp\left[-\beta Q(\boldsymbol{f})\right] \text{ with } Q(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 \quad (3)$$

• Information about $\boldsymbol{f}$ is assumed to be in the form :

$$\mathrm{E}\left\{\phi_i(\boldsymbol{f})\right\} = \mu_i \text{ with } \phi_i(\boldsymbol{f}) = \sum_{n=1}^{N} \Omega_i(f_n),\ i = 1, 2 \quad (4)$$

Using this information with MEP we obtain :

$$p(\boldsymbol{f}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(f_n|\boldsymbol{\theta}) \quad (5)$$

with

$$p(f_n|\boldsymbol{\theta}) = \frac{1}{Z(\theta_1, \theta_2)} \exp\left[-\theta_1\Omega_1(f_n) - \theta_2\Omega_2(f_n)\right] \quad (6)$$

and

$$Z(\theta_1, \theta_2) = \int \exp\left[-\theta_1\Omega_1(f) - \theta_2\Omega_2(f)\right] \mathrm{d}f \quad (7)$$

A scale invariance argument [3, 4] limits the set of possible functions $\Omega_1(f)$ and $\Omega_2(f)$:

$$\left\{(f^{r_1}, f^{r_2}), (f^{r_1}, \ln f), (f^{r_1}, f^{r_1} \ln f), (\ln f, \ln^2 f)\right\}. \quad (8)$$

• Applying the Bayes' rule to obtain the posterior law $p(\boldsymbol{f}|\boldsymbol{g})$ and MAP estimation rule we obtain :

$$\widehat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \{p(\boldsymbol{f}|\boldsymbol{g}\} = \arg\min_{\boldsymbol{f}} \{J(\boldsymbol{f})\}, \quad (9)$$

$$\text{with} \quad J(\boldsymbol{f}) = \beta Q(\boldsymbol{f}) + \theta_1 \phi_1(\boldsymbol{f}) + \theta_2 \phi_2(\boldsymbol{f}). \quad (10)$$

So, the well-posed Bayesian problem solving should be stated as follows:

Given $\boldsymbol{H}, \boldsymbol{g}, \beta = 2/\sigma^2$, and $\{\mu_1, \mu_2\}$ or equivalently $\{\theta_1, \theta_2\}$, estimate $\boldsymbol{f}$.

However, in practical applications (μ_1, μ_2) or equivalently (θ_1, θ_2) are not given to us and we want also to estimate them from the data $\boldsymbol{g}$.

2. JOINT ESTIMATION PROBLEM

In this paper we considered the following problem :

Given $\boldsymbol{H}, \boldsymbol{g}$ and β estimate $\boldsymbol{f}$ and the hyperparameters $\boldsymbol{\theta} = [\theta_1, \theta_2]$.

Two main approaches studied and compared here are :

• Joint Maximum *a posteriori* (JMAP) :
The main idea behind this approach is to consider the hyperparameters $\boldsymbol{\theta}$ on the same level that the other parameters $\boldsymbol{f}$ and try to estimate them by

$$\text{where} \quad (\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\theta}}) = \arg\max_{(\boldsymbol{f}, \boldsymbol{\theta})} \{p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})\} \quad (11)$$

$$p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{f}, \boldsymbol{\theta}, \boldsymbol{g}) = p(\boldsymbol{g}|\boldsymbol{f})\, p(\boldsymbol{f}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \quad (12)$$

Note that, we can choose either an improper or an uniform prior for $p(\boldsymbol{\theta})$.

This optimization problem can be implemented by successively maximizing with respect to $\boldsymbol{\theta}$ and to $\boldsymbol{f}$:

$$\begin{cases} \widehat{\boldsymbol{\theta}}^{(k)} = \arg\max_{\boldsymbol{\theta}} \left\{p(\widehat{\boldsymbol{f}}^{(k)}, \boldsymbol{\theta}|\boldsymbol{g})\right\} = \arg\max_{\boldsymbol{\theta}} \left\{p(\widehat{\boldsymbol{f}}^{(k)}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\right\} \\ \widehat{\boldsymbol{f}}^{(k+1)} = \arg\max_{\boldsymbol{f}} \left\{p(\boldsymbol{f}, \widehat{\boldsymbol{\theta}}^{(k)}|\boldsymbol{g})\right\} = \arg\max_{\boldsymbol{f}} \left\{p(\boldsymbol{f}|\boldsymbol{g}, \widehat{\boldsymbol{\theta}}^{(k)})\right\} \end{cases} \quad (13)$$

Note that the first equation can be interpreted as a maximum likelihood estimate of $\boldsymbol{\theta}$ if $p(\boldsymbol{\theta})$ is choosed to be uniforme and if $\boldsymbol{f}^{(k)}$ could be considered as a sample of the prior law $p(\boldsymbol{f}|\boldsymbol{\theta})$.

• Marginalized MAP (MMAP) :
The main idea behind this approach is to consider the hyperparameters $\boldsymbol{\theta}$ on a different level than $\boldsymbol{f}$. So, $\boldsymbol{\theta}$ is first estimated by marginalizing with respect to $\boldsymbol{f}$:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{L(\boldsymbol{\theta}) = \int p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})\, \mathrm{d}\boldsymbol{f}\right\} \quad (14)$$

Then using $\widehat{\boldsymbol{\theta}}$ we can estimate the solution by

$$\widehat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \left\{p(\boldsymbol{f}|\boldsymbol{g}, \widehat{\boldsymbol{\theta}})\right\} \quad (15)$$

Unfortunately the analytical calculus of the integral $L(\boldsymbol{\theta})$ is rarely (excepted for the Gaussian case) possible. However, noting that when $p(\boldsymbol{\theta})$ is uniform $L(\boldsymbol{\theta})$ becomes actually the likelihood and $\widehat{\boldsymbol{\theta}}$ the maximum likelihood (ML) estimate, in some cases one can obtain the solution using the Expectation-Maximisation (EM) algorithm, which in our case, is given by :

$$\begin{cases} \text{E:} \quad Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}) = \mathrm{E}_{\boldsymbol{f}|\boldsymbol{g}, \widehat{\boldsymbol{\theta}}^{(k)}} \left\{\ln p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})\right\} \\ \text{M:} \quad \widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \left\{Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)})\right\} \end{cases} \quad (16)$$

3. COMPARISON OF JMAP AND EM-MMAP

Now, let us go a little further inside these two methods by assuming $p(\boldsymbol{\theta})$ is chosen to be uniforme. Noting back $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and replacing $p(\boldsymbol{f}|\boldsymbol{\theta})$ from (4) and $p(\boldsymbol{g}|\boldsymbol{f})$ from (3) in the equations (13) and (16) we can make the following comparison :

• **JMAP :** At iteration $(k+1)$ of the iterative optimization algorithm (13) we have to estimate $(\theta_1, \theta_2)^{(k+1)}$ by maximizing $\ln p(\widehat{\boldsymbol{f}}^{(k)}|\theta_1, \theta_2)$ with respect to (θ_1, θ_2). To do this we have to deal with the following system of nonlinear equations :

$$\frac{\partial \ln Z(\theta_1, \theta_2)}{\partial \theta_i} = \frac{1}{N} \phi_i \left(\widehat{\boldsymbol{f}}^{(k)}_{MAP}\right), \quad i = 1, 2 \quad (17)$$

where the right hand side (RHS) of these equations are

$$\frac{1}{N} \phi_i \left(\widehat{\boldsymbol{f}}^{(k)}_{MAP}\right) = \sum_{n=1}^{N} \Omega_i(\widehat{x}^{(k)}_{MAP_n}), \quad i = 1, 2 \quad (18)$$

• **EM-MMAP :** When using the EM algorithm (16) at iteration $(k+1)$ we have to estimate $(\theta_1, \theta_2)^{(k+1)}$ by maximizing $Q\left((\theta_1, \theta_2); (\theta_1, \theta_2)^{(k)}\right)$ with respect to (θ_1, θ_2). To do this we have to deal with the following system of nonlinear equations :

$$\frac{\partial \ln Z(\theta_1, \theta_2)}{\partial \theta_i} = \frac{1}{N} \int \phi_i(\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{g}, \widehat{\boldsymbol{\theta}}^{(k)})\, \mathrm{d}\boldsymbol{f}, \quad i = 1, 2. \quad (19)$$

The RHS of these two equations can be written

$$RHS = \frac{1}{N} \sum_{n=1}^{N} \int \Omega_i(f_n) p(f_n|\boldsymbol{g}, \widehat{\boldsymbol{\theta}}^{(k)})\, \mathrm{d}f_n, \; i = 1, 2. \quad (20)$$

Thus, comparing (17) and (19), we can say that if $p(\boldsymbol{f}|\boldsymbol{g},\widehat{\boldsymbol{\theta}}^{(k)})$ is very concentrated around the $\widehat{\boldsymbol{f}}^{(k)}_{MAP}$, i.e., $p(\boldsymbol{f}|\boldsymbol{g},\widehat{\boldsymbol{\theta}}^{(k)}) \approx \delta(\boldsymbol{f}-\widehat{\boldsymbol{f}}^{(k)}_{MAP})$, then the integrals of the RHS of the equation (19) will be equivalent to RHS of (17) and the two methods will give the same numerical result. But this is not true in general.

However JMAP is easy to implement, because it does not need any integration. This is not true in the case of the EM-MMAP. Note however that, thanks to the entropic priors, the N-dimentional integration in (19) is replaced by N one-dimentional integrations in (20). However, in general, these integrals have rarely analytical solutions, so, implementing the EM algorithm is very difficult due to the need of calculating the marginal law $p(f_n|\boldsymbol{g})$ and these integrals. Two solutions have been proposed in litteratures [5, 6, 7, 8] to surround this difficulty:

1) making a Gaussian approximation for $p(\boldsymbol{f}|\boldsymbol{g},\widehat{\boldsymbol{\theta}}^{(k)})$ around the maximum $\widehat{\boldsymbol{f}}_{MAP}$ which allows to obtain an analytic solution to these integrals, and
2) calculating the integrals in a stochastic method which leads us to Stochastic EM (SEM) like methods.

What we propose is an an *ad hoc* approach to approximate these integrals by

$$\int \Omega_i(f_n) p(f_n|\boldsymbol{g})\,\mathrm{d}f_n \simeq \frac{1}{k}\sum_{l=1}^{k}\Omega_i(f_n^{(k)}),\ i=1,2. \quad (21)$$

This means that the expectations of $\Omega_i(f_n)$ are replaced by the average values of them calculated during all the past iterations of the algorithm. Doing this, the two methods will have the same structure, i.e; to obtain new values for the (θ_1,θ_2) at each iteration we have to resolve the following system of equations:

$$\frac{\partial \ln Z(\theta_1,\theta_2)}{\partial\theta_i} = d_i, \quad i=1,2, \quad (22)$$

where

$$d_i = \begin{cases} \displaystyle\sum_{n=1}^{N}\Omega_i(\widehat{x}^{(k)}_{MAP_n}) & \text{for JMAP} \\ \displaystyle\frac{1}{k}\sum_{l=1}^{k}\sum_{n=1}^{N}\Omega_i(\widehat{x}^{(k)}_{MAP_n}) & \text{for EM-MMAP} \end{cases} \quad (23)$$

A final and very important remarque is that we have to insure that the criteria to be maximized admit at least a local maximum. This question may arises more specifically when dealing with JMAP, but it may also arises in the case of the EM-MMAP. This difficulty is indepent of the algorithm used to find the optimum. To explain more this let us have a look on the expression to optimize in JMAP (eq. 11) which can be written as

$$(\widehat{\boldsymbol{f}},\widehat{\boldsymbol{\theta}}) = \arg\min_{(\boldsymbol{f},\boldsymbol{\theta})}\left\{-\ln p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})\right\}, \quad (24)$$

where

$$\begin{aligned} -\ln p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g}) &= \beta\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|^2 + \theta_1\phi_1(\boldsymbol{f}) + \theta_2\phi_2(\boldsymbol{f}) \\ &\quad -N\ln Z(\theta_1,\theta_2) - \ln p(\boldsymbol{\theta}). \end{aligned} \quad (25)$$

Depending on the choice of $p(\boldsymbol{\theta})$ and $p(\boldsymbol{f}|\boldsymbol{\theta})$ and consequently $\phi_1(\boldsymbol{f})$, $\phi_2(\boldsymbol{f})$ and $Z(\theta_1,\theta_2)$, this criterion may even not have a local minimum.

In practice this difficulty is encountered when we are dealing with the system of equations (17), (19) or (22) which are obtained by equating to zero the gradient of the above criterion with respect to $\boldsymbol{\theta}$. In fact this system of equations may not have any solution for given values d_i. On the same way $\ln Z(\theta_1,\theta_2)$ which depends on the prior probability law $p(f|\theta_1,\theta_2)$ may also be defined only on a set of admissible values of (θ_1,θ_2). For example, in the Gaussian case where $\Omega_1(f)=f^2$ and $\Omega_2(f)=f$ we have $Z(\theta_1,\theta_2)=\sqrt{\pi/\theta_1}$ which does not depend even on θ_2 and θ_1 must be positive.

However, when we have taken the necessary cautions, in practical situations the two algorithms will give satisfactory results. This has been shown at least in our simulations as we will see below.

4. SIMULATION RESULTS

In this section, we present some simulation results which show the relative performances of the the two proposed methods in image restoration problems. For this, we have first created two synthetic (64 by 64) images (O1 and O2) (Fig. 1). According to the histograms of the images, we assigned

- a Gaussian law ($\Omega_1(f)=f^2$, $\Omega_2(f)=f$) to the first image, and
- a Gamma law ($\Omega_1(f)=\ln f, \Omega_2(f)=f, f>0$) to the second one.

Table 1. summarizes the results of parameter estimation.

images	$\Omega_1(f)$	$\Omega_2(f)$	f dom.	θ_1	θ_2
O1	f^2	f	$\mathbb{R}$	$4.5e{-4}$	$-2.5e{-2}$
O2	f^2	f	$f>0$	$6.9e{-1}$	$1.1e{-2}$

Table 1: Prior laws and their parameters.

In a next step, we created degraded images by blurring them with a (9 by 9) Gaussian point spread function (PSF) and added a Gaussian noise with a given variance so that the signal to noise ratio was fixed to 20dB. Fig. 2 shows the degraded images.

In the final step, using the JMAP and the EM-MMAP methods we restored these images and simultaneously estimated the hyperparameters (θ_1,θ_2). Fig. 3 and Fig. 4 show the restored images by JMAP and EM-MMAP methods and the Table 2. summarizes the hyperparameter estimation results.

	JMAP		EM-MMAP	
images	$\widehat{\theta_1}$	$\widehat{\theta_2}$	$\widehat{\theta_1}$	$\widehat{\theta_2}$
O1	$4.3e{-}4$	$-2.4e{-}2$	$4.3e{-}4$	$-2.5e{-}2$
O2	$6.2e{-}1$	$1.3e{-}2$	$6.3e{-}2$	$1.3e{-}2$

Table 2: Results of hyperparameter estimation and distance between the original and estimated images for the two methos JMAP and EM-MMAP.

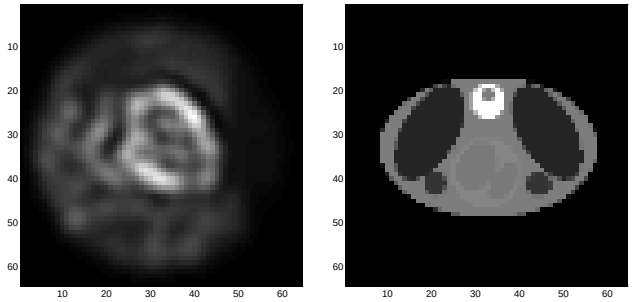

Fig. 1: Original images o1 and o2.

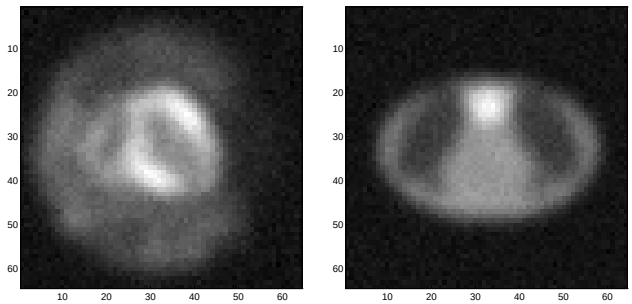

Fig. 2: Degraded images I1 and I2.

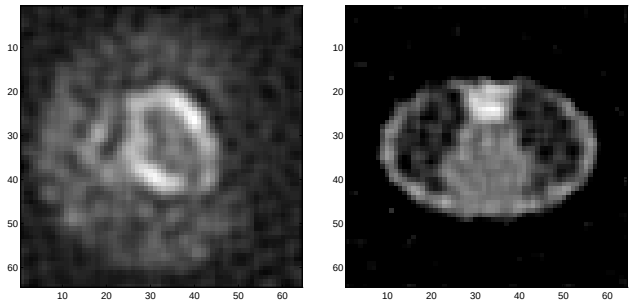

Fig. 3: Restored images by JMAP method.

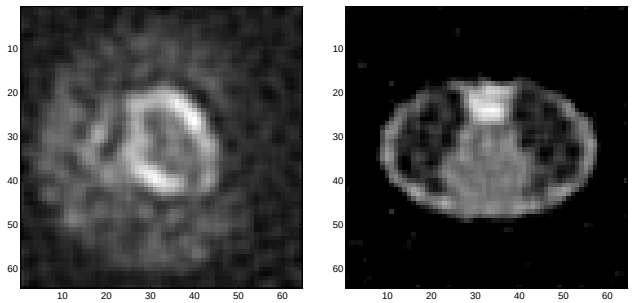

Fig. 4: Restored images by EM-MMAP method.

5. CONCLUSIONS

The main objective of this paper was to propose and compare two specific methods for joint estimation of the unknowns and the hyperparameters of the inverse problems in a Bayesian approach with ME priors. The first method, named EM-MMAP, is based on the marginalised maximum likelihood (MML) and the EM algorithm and the second, named JMAP, is based on generalized maximum likelihood (GML) methods. The relative performances of these two methods have been showed in image restoration problems.

6. REFERENCES

[1] A. Mohammad-Djafari and J. Idier, *Maximum Likelihood Estimation of the Lagrange Parameters of the Maximum Entropy Distributions*, pp. 131–140. Maximum entropy and Bayesian methods, Seattle, USA: Kluwer Academic Publ., smith, c.r. and erikson, g.j. and neudorfer, p.o. ed., 1991.

[2] A. Mohammad-Djafari, "On the estimation of hyperparameters in Bayesian approach of solving inverse problems," in *Proceedings of IEEE ICASSP*, (Minneapolis, U.S.A.), pp. 567–571, IEEE, Aprilxspace1993.

[3] A. Mohammad-Djafari and J. Idier, "Scale invariant Bayesian estimators for linear inverse problems," in *Proc. of the First ISBA meeting*, (San Fransisco, USA), Augustxspace1993.

[4] S. Brette, J. Idier, and A. Mohammad-Djafari, "Scale invariant Markov models for linear inverse problems," in *Proc. of the Section on Bayesian Statistical Sciences*, (Alicante, Spain), pp. 266–270, American Statistical Association, 1994.

[5] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[6] M. Miller and D. Snyder, "The role of likelihood and entropy in incomplete-data problems: Applications to estimating point-process intensities and toeplitz constrained covariances," *Proceedings of the IEEE*, vol. 75, pp. 892–906, Julyxspace1987.

[7] C. F. J. Wu, "On the convergence of the em algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

[8] F. Champagnat and J. Idier, "An alternative to standard maximum likelihood for Gaussian mixtures," in *Proceedings of IEEE ICASSP*, (Detroit), pp. 2020–2023, 1995.