2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

An EM-based estimation algorithm for a class of systems promoting sparsity

Boris I. Godoy, Rodrigo Carvajal, and Juan C. Agüero

Abstract— **In this paper we propose a Maximum a Posteriori (MAP) approach for estimating a random sparse parameter vector in the presence of nonlinearities of unknown parameters. In this Bayesian approach, the *a priori* probability distribution for the parameter vector is utilised as a mechanism to promote sparsity. We solve this identification problem by using a generalized Expectation Maximization algorithm in a MAP framework.**

I. Introduction

System identification considering sparsity has attracted increasing attention in the last decades, see e.g. [1], [2], [18], [20]–[22] and the references therein. In particular, real-world applications can be found in wireless communication systems (e.g. underwater acoustic channels [10], digital television channels [16], and residential ultrawideband channels [15]).

From a system identification perspective, sparsity can be promoted in different ways. For example, in [12], sparsity is promoted by generating a pool of possible models, and then performing model selection. On the other hand, one of the most used approaches is the Lasso algorithm [21], where an ℓ_1-norm regularization is used to obtain estimates with coefficients that are exactly zero in a linear regression. However, the Lasso algorithm is not applicable to the model presented in this work, since it is restricted to linear systems only, where the parameter vector defining the linear regression is the only unknown and there are no *hidden variables*. In contrast, the approach taken by the Lasso (applying the ℓ_1-norm regularization) can still be considered, but the solution cannot be obtained by applying the Lasso algorithm. In our case, we deal with a class of systems that does not satisfy the requirements of the traditional sparse estimation techniques (see e.g. [2], [4], [21]), and that includes applications such as the identification of communication channels (see e.g. [6] and [7]).

In this paper, we propose the utilization of a Bayesian approach for sparse parameter estimation, that can be related to the ℓ_1-norm regularization. Because we consider the presence of *hidden variables*, we use the Expectation-Maximization (EM) algorithm. This algorithm is extended from its classical formulation to a general MAP formulation, by defining an augmented auxiliary function. This augmented auxiliary function is based on the utilization of mixtures for representing the *a priori* distribution of the sparse parameters. In particular, we use variance-mean Gaussian mixtures (VMGM) [18], which can represent a wide range of distributions. This representation is inserted into the generalized EM algorithm. In addition, we consider the concentration of the cost function in the maximization step of the EM algorithm, to numerically optimize a reduced number of parameters.

Numerical examples presented here show that the estimation of a sparse parameter vector is improved when an *a priori* distribution is considered. In addition, we provide a method to estimate the parameter defining the *prior* distribution of the estimates.

The remainder of the paper is as follows. In § II we describe the problem of interest. In § III we describe the Maximum a posteriori approach (MAP) to this problem considering sparsity. In § IV we present the modified EM algorithm for MAP estimation. In § V we carry out examples, considering a particular case for the model treated in this work. Finally, in § VI we present our conclusions.

II. Problem formulation

Consider the general system model as follows:

$$\mathbf{y} = \mathbf{M}\mathbf{h} + \eta, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^N$, $\mathbf{M} \in \mathbb{R}^{N\times p}$, $\eta \sim \mathcal{N}(0, \sigma^2\mathbf{I}_N)$, $\mathbf{I}_N \in \mathbb{R}^{N\times N}$ is the identity matrix of order N, $\mathbf{h} \in \mathbb{R}^p$ is a sparse random vector, with *a priori* distribution given by $p(\mathbf{h})$.

A particular case of the system model in (1) is as follows:

$$\mathbf{y} = f(\varepsilon)\mathbf{A}(\mathbf{x},\mathbf{u})\mathbf{h} + \eta, \tag{2}$$

where $f : \mathbb{R}^{n_\varepsilon} \to \mathbb{R}^{N\times n_a}$. In addition, $\mathbf{A}(\mathbf{x},\mathbf{u}) \in \mathbb{R}^{n_a\times p}$ satisfies the following property

$$\mathbf{A}(\mathbf{x},\mathbf{u})\mathbf{h} = \mathbf{B}_1(\mathbf{h})\mathbf{x} + \mathbf{B}_2(\mathbf{h})\mathbf{u} + c = (\mathbf{A}_1(\mathbf{x}) + \mathbf{A}_2(\mathbf{u}))\mathbf{h} + c, \tag{3}$$

where $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{B}_1$, and $\mathbf{B}_2$ are known functions of $\mathbf{x}$ and $\mathbf{u}$, c is a known constant, $\mathbf{u} \in \mathbb{R}^{n_u}$ is a deterministic vector, and $\mathbf{x} \in \mathbb{R}^{n_x}$ is (in general) an unknown random vector $\mathbf{x} \sim p(\mathbf{x}|\beta)$.

The problem of interest consists of estimating the parameter $\gamma = [\mathbf{h}^T \operatorname{vec}(\beta)^T \varepsilon\, \sigma]^T$, subject to the structure given in (2) and (3), and where $\operatorname{vec}(\mathbf{P})$ stacks the columns of the matrix $\mathbf{P}$ on top of one another.

III. MAP estimation

Maximum a Posteriori estimation is a common optimization-based technique for parameter estimation of a posterior distribution, assuming that the parameters are random variables. In this approach, the distribution parameters are estimated so as to maximize the posterior

This work was supported in part by Chile's National Commission for Scientific and Technological Research (CONICYT) under Grant ACT-053. The work by B. I. Godoy was supported by CONICYT Chile through its Postdoctoral Fellowship Program 2011.

B. I. Godoy and J. C. Agüero are with the School of Electrical Engineering and Computer Science, The University of Newcastle, Callaghan, NSW, 2308, Australia.
R. Carvajal is with the Electronics Engineering Department, Universidad Técnica Federico Santa María, Chile.
Email addresses: boris.godoy@newcastle.edu.au, rodrigo.carvajalg@usm.cl, juan.aguero@newcastle.edu.au.

probability of the random parameters given the observed sample values. In particular, the *a priori* (or *prior*) probability density function (pdf) plays a key role in the parameter estimation procedure when the number of measurement points is limited.

The MAP estimator is obtained as

$$\hat{\gamma}_{\text{MAP}} = \arg\max_{\gamma} p(\gamma|\mathbf{y}), \tag{4}$$

where $p(\gamma|\mathbf{y})$ is the posterior distribution of the parameters γ given the observed data $\mathbf{y}$ (measurements). In general, Bayes' rule is applied in MAP estimation, in order to optimize an equivalent expression to $p(\gamma|\mathbf{y})$. In this sense, we have that

$$p(\gamma|\mathbf{y}) = \frac{p(\mathbf{y}|\gamma)p(\gamma)}{p(\mathbf{y})}, \tag{5}$$

where $p(\mathbf{y}|\gamma)$ is the *likelihood* function, and $p(\gamma)$ is the *prior* pdf for γ. Since the marginal pdf of the measurements, $p(\mathbf{y})$, does not depend on γ, its role in the posterior distribution is simply a normalization factor. Thus, it is possible to obtain the MAP estimator as

$$\hat{\gamma}_{\text{MAP}} = \arg\max_{\gamma}\ p(\mathbf{y}|\gamma)p(\gamma), \tag{6}$$

$$= \arg\max_{\gamma}\ \ell(\gamma) + \log p(\gamma), \tag{7}$$

where $\ell(\gamma)$ is the *log-likelihood* function, and $\log p(\gamma)$ is the *log-prior* function of γ.

A. Regularization and MAP estimation

When facing a sparse estimation problem, one of the key issues is the selection of a mechanism that promotes sparsity. In general, the approaches found in the literature deal with a regularized problem. The most common approach corresponds to including an ℓ_1-norm term of the parameter vector in the form of $\alpha||\gamma||_1$, $\alpha > 0$, and can be expressed as:

$$\hat{\gamma}_{\text{Reg}} = \arg\max_{\gamma}\ \ell(\gamma) - \alpha||\gamma||_1. \tag{8}$$

On the other hand, some other approaches have also explored different types of regularization for sparse parameter estimation (see e.g [24] and [25]).

From a Bayesian point of view, the regularization term in (8) can be related to an *a priori* pdf[1]. Hence, a MAP estimation problem can also be understood as a regularized Maximum Likelihood (ML) estimation problem. In this sense, it is possible to consider that

$$\log p(\gamma) = -\alpha||\gamma||_1 + \xi \ \Leftrightarrow\ p(\gamma) \propto e^{-\alpha||\gamma||_1}, \tag{9}$$

where ξ is a constant term. Given the relationship in (9), it is possible to define an *a priori* pdf for γ that promotes sparsity.

B. Laplace pdf

In our problem of interest, a Laplace pdf (see e.g. [11]) can be utilized for promoting sparsity (see e.g. [17] and the references therein). The Laplace pdf for an uncorrelated zero-mean vector is given by

$$p(\gamma) = \left(\frac{1}{2\tau}\right)^{\mathrm{p}} e^{-\|\gamma\|_1/\tau}, \tag{10}$$

[1]Notice the evident similarity between (7) and (8)

where p is the length of the sparse parameter vector γ, and $\tau \in \mathbb{R}^+$. At this stage, we will assume that τ is known. However, as we will explore later, this condition is not strictly necessary.

In this paper, we consider solving a sparse MAP estimation problem considering a Laplace *prior* for the parameter vector, considering also *hidden variables* and the utilization of the EM algorithm. In the following section, we depart from the classical formulation of the EM algorithm for MAP estimation, presenting an alternative formulation that encompasses hidden variables in the *a priori* pdf.

IV. The EM algorithm for MAP estimation

The EM algorithm is an iterative method that generates a succession of estimates $\hat{\gamma}^{(i)} = (\hat{\mathbf{h}}^{(i)}, \text{vec}(\hat{\beta}^{(i)}), \hat{\varepsilon}^{(i)}, \sigma^{(i)})$, $i = 1, 2, \ldots$, of the parameters γ, which converges to a local maximum of the log-likelihood function. The EM algorithm consists, basically, of an iterative two-step procedure: (i) an expectation step (E-step), and (ii) a maximization step (M-step). In our case, we develop an augmented EM algorithm to solve the MAP estimation problem presented in this paper.

The EM algorithm can be modified in a simply manner to include the extra term arising from the *prior* distribution of the parameters in MAP estimation.

A. The EM algorithm for ML estimation

In ML estimation, the parameter estimation procedure is based on maximizing the *likelihood function*, or equivalently the *log-likelihood* function, $\ell(\gamma)$. In the presence of *hidden variables*, the EM algorithm becomes a powerful tool for obtaining the ML estimates.

If we denote by $\mathbf{z}$ the *hidden variables*, then the EM algorithm for ML estimation is defined by (see e.g. [8], [14])

$$\ell(\gamma) = \int \ell(\gamma) p(\mathbf{z}|\mathbf{y}, \hat{\gamma}^{(i)}) d\mathbf{z}$$

$$= \int \log p(\mathbf{z}, \mathbf{y}|\gamma) p(\mathbf{z}|\mathbf{y}, \hat{\gamma}^{(i)}) d\mathbf{z} - \int \log p(\mathbf{z}|\mathbf{y}, \gamma) p(\mathbf{z}|\mathbf{y}, \hat{\gamma}^{(i)}) d\mathbf{z}$$

$$= \mathscr{Q}_{\text{ML}}(\gamma, \hat{\gamma}^{(i)}) - \mathscr{H}_{\text{ML}}(\gamma, \hat{\gamma}^{(i)}), \tag{11}$$

where $\mathscr{Q}_{\text{ML}}(\gamma, \hat{\gamma}^{(i)})$ and $\mathscr{H}_{\text{ML}}(\gamma, \hat{\gamma}^{(i)})$ are the auxiliary functions arising from ML. Then, the ML estimate can be obtained iteratively as

$$\gamma^{(i+1)} = \arg\max_{\gamma} \mathscr{Q}_{\text{ML}}(\gamma, \hat{\gamma}^{(i)}.) \tag{12}$$

B. The EM algorithm for MAP estimation: classical formulation

Given the expressions in (7) and (11), the *log-posterior* function can be expressed as

$$\log p(\gamma|\mathbf{y}) = \mathscr{Q}_{\text{ML}}(\gamma, \hat{\gamma}^{(i)}) - \mathscr{H}_{\text{ML}}(\gamma, \hat{\gamma}^{(i)}) + \log p(\gamma) - \log p(\mathbf{y}), \tag{13}$$

since the marginal pdf's $p(\gamma)$ and $p(\mathbf{y})$ do not depend on the *hidden variable* $\mathbf{z}$. Hence, the EM algorithm for MAP estimation is given by (see e.g. [14]):

E-step:

$$\mathscr{Q}_{\text{MAP}}(\gamma, \hat{\gamma}^{(i)}) = E[\log p(\mathbf{z}, \mathbf{y}|\gamma)|\mathbf{y}, \hat{\gamma}^{(i)}] + \log p(\gamma), \tag{14}$$

M-step:

$$\hat{\gamma}^{(i+1)} = \arg\max_{\gamma} \mathscr{Q}_{\text{MAP}}(\gamma, \hat{\gamma}^{(i)}). \tag{15}$$

In the next section, we will see how we can modify the EM algorithm for MAP estimation. One of the benefits of the proposed method is that, in general, a quadratic term of the parameters is obtained for the E-step. Also, the proposed method can be understood as a generalization of several methods of different nature found in the literature, see [5] and the references therein.

C. The EM algorithm combined with Infinite Mixtures

In general, it is possible to express a pdf as a marginal pdf, coming from a joint pdf of two or more variables (vectors). In this sense, we represent the marginal pdf as an infinite mixture, where there is an underlying process that generates the desired pdf for the parameters. In general, a mixture is given by (see e.g. [9]):

$$p(\gamma) = \int p(\gamma|\lambda)p(\lambda)d\lambda, \tag{16}$$

where $\gamma \in \mathbb{R}^{\mathrm{p}}$ and $\lambda \in \mathbb{R}^{\mathrm{r}}$. If the conditional pdf $p(\gamma|\lambda)$ is Gaussian, then the mixture is called variance-mean Gaussian mixture (VMGM) (see e.g. [18]), normal variance-mean mixture (NVMM, see e.g. [3]), or normal scale mixture (see e.g. [23]).

In (16), it is possible to consider the variable λ as a hidden variable. First, considering the variable of interest to be γ, the *log-prior* can be expressed as

$$\log p(\gamma) = \log p(\gamma,\lambda) - \log p(\lambda|\gamma). \tag{17}$$

Second, in the same manner as it was done with the *log-likelihood* function, both sides of (17) can be integrated with respect to $p(\lambda|\hat{\gamma}^{(i)})$ to obtain:

$$\begin{aligned}\log p(\gamma) = &\int \log p(\gamma,\lambda)p(\lambda|\hat{\gamma}^{(i)})d\lambda - \\ &\int \log p(\lambda|\gamma)p(\lambda|\hat{\gamma}^{(i)})d\lambda \\ &= \mathscr{Q}_{\mathrm{prior}}(\gamma,\hat{\gamma}^{(i)}) - \mathscr{H}_{\mathrm{prior}}(\gamma,\hat{\gamma}^{(i)}),\end{aligned} \tag{18}$$

where

$$\mathscr{Q}_{\mathrm{prior}}(\gamma,\hat{\gamma}^{(i)}) = \int \log p(\gamma,\lambda)p(\lambda|\hat{\gamma}^{(i)})d\lambda, \tag{19}$$

$$\mathscr{H}_{\mathrm{prior}}(\gamma,\hat{\gamma}^{(i)}) = \int \log p(\lambda|\gamma)p(\lambda|\hat{\gamma}^{(i)})d\lambda. \tag{20}$$

The function $\mathscr{H}_{\mathrm{prior}}(\gamma,\hat{\gamma}^{(i)})$ has the same properties that $\mathscr{H}_{\mathrm{ML}}(\gamma,\hat{\gamma}^{(i)})$ has in the *log-likelihood* function. That is, for any γ, using Jensen's inequality, we have:

$$\begin{aligned}&\mathscr{H}_{\mathrm{prior}}(\gamma,\hat{\gamma}^{(i)}) - \mathscr{H}_{\mathrm{prior}}(\hat{\gamma}^{(i)},\hat{\gamma}^{(i)}) = \\ &\int \log p(\lambda|\gamma)p(\lambda|\hat{\gamma}^{(i)})d\lambda - \int \log p(\lambda|\hat{\gamma}^{(i)})p(\lambda|\hat{\gamma}^{(i)})d\lambda \\ &= \int \log \frac{p(\lambda|\gamma)}{p(\lambda|\hat{\gamma}^{(i)})} p(\lambda|\hat{\gamma}^{(i)})d\lambda \\ &\leq \log \int \frac{p(\lambda|\gamma)}{p(\lambda|\hat{\gamma}^{(i)})} p(\lambda|\hat{\gamma}^{(i)})d\mathbf{z} \\ &= \log \int p(\mathbf{z}|\mathbf{y},\gamma)d\mathbf{z} \\ &= 0.\end{aligned} \tag{21}$$

Thus, from (14), we can re-write the E-step as:

$$\mathscr{Q}_{\mathrm{MAP}}(\gamma,\hat{\gamma}^{(i)}) = \mathscr{Q}_{\mathrm{ML}}(\gamma,\hat{\gamma}^{(i)}) + \mathscr{Q}_{\mathrm{prior}}(\gamma,\hat{\gamma}^{(i)}), \tag{22}$$

where $\mathscr{Q}_{\mathrm{prior}}(\gamma,\hat{\gamma}^{(i)})$ is the function corresponding to the *a priori* distribution.

Since the M-step in (15) involves the maximization of the function $\mathscr{Q}_{\mathrm{MAP}}(\gamma,\hat{\gamma}^{(i)})$ with respect to the parameter γ, in the following sections we calculate the expressions for $\mathscr{Q}_{\mathrm{prior}}(\gamma,\hat{\gamma}^{(i)})$, $\mathscr{Q}_{\mathrm{ML}}(\gamma,\hat{\gamma}^{(i)})$, and their derivatives.

D. Measurement noise variance estimation and reparametrization

Recalling the relationship between MAP estimation and regularized ML estimation, in an optimization problem such as (4), with a model like in (2) and (3), a good estimate $\hat{\sigma}$ is crucial. Thus, it is important to take into account the measurement noise variance or (equivalently) the standard deviation, σ, in the definition of the problem itself when the measurement noise variance is unknown (see e.g. [19]). If not, the optimization problem may be non-convex and exhibit numerical problems. To make it convex, we can use the procedure suggested in [19]. That is, we introduce the following change of variables:

$$\varphi = h/\sigma, \qquad \rho = \sigma^{-1}. \tag{23}$$

We therefore define a new parameter to be estimated: $\theta = [\varphi^T \operatorname{vec}(\beta)^T \varepsilon \rho]^T$ and the modified auxiliary function for MAP estimation is given by:

$$\mathscr{Q}_{\mathrm{MAP}}(\theta,\hat{\theta}^{(i)}) = \mathscr{Q}_{\mathrm{ML}}(\theta,\hat{\theta}^{(i)}) + \mathscr{Q}_{\mathrm{prior}}(\theta,\hat{\theta}^{(i)}), \tag{24}$$

and the M-step is defined by

$$\hat{\theta}^{(i+1)} = \arg\max_{\theta} \mathscr{Q}_{\mathrm{MAP}}(\theta,\hat{\theta}^{(i)}). \tag{25}$$

For the reminder of the paper, we will consider that the parameter vector is now given by θ.

E. Evaluation of $\mathscr{Q}_{prior}(\theta,\hat{\theta}^{(i)})$ and its derivative

Lemma 1: When $p(\theta|\lambda)$ is a Gaussian distribution $(\theta|\lambda \sim \mathscr{N}(\mu_\theta(\lambda),\Sigma_\theta(\lambda)))$, we have:

$$\frac{\partial \mathscr{Q}_{\mathrm{prior}}}{\partial \theta} = \mathbb{E}_{\lambda|\hat{\theta}^{(i)}}[-\Sigma_\theta^{-1}(\lambda)]\theta + \mathbb{E}_{\lambda|\hat{\theta}^{(i)}}[\Sigma_\theta^{-1}(\lambda)\mu_\theta(\lambda)]. \tag{26}$$

Proof: To calculate $\frac{\partial \mathscr{Q}_{\mathrm{prior}}}{\partial \theta}$, we notice that $\frac{\partial}{\partial \theta}\log p(\theta,\lambda) = \frac{\partial}{\partial \theta}\log p(\theta|\lambda)$. Then, we have:

$$\begin{aligned}\frac{\partial \mathscr{Q}_{\mathrm{prior}}}{\partial \theta} &= \int \frac{\partial}{\partial \theta}\log p(\theta|\lambda)p(\lambda|\hat{\theta}^{(i)})d\lambda \\ &= \int [-\Sigma_\theta^{-1}(\lambda)(\theta-\mu_\theta(\lambda))]p(\lambda|\hat{\theta}^{(i)})d\lambda \\ &= \int -\Sigma_\theta^{-1}(\lambda)p(\lambda|\hat{\theta}^{(i)})d\lambda\,\theta \\ &\quad + \int \mu_\theta(\lambda))p(\lambda|\hat{\theta}^{(i)})d\lambda. \\ &= \mathbb{E}_{\lambda|\hat{\theta}^{(i)}}[-\Sigma_\theta^{-1}(\lambda)]\theta + \mathbb{E}_{\lambda|\hat{\theta}^{(i)}}[\Sigma_\theta^{-1}(\lambda)\mu_\theta(\lambda)].\end{aligned} \tag{27}$$

(28)

■

The expected values in (26) can be computed in different ways. In particular, we consider the following aspects:

(i) Since $p(\theta|\lambda)$ is assumed Gaussian, $p(\theta)$ satisfies the following relations:

$$\frac{\partial p(\theta)}{\partial \theta} = \int [-\Sigma_\theta^{-1}(\lambda)(\theta - \mu_\theta(\lambda))] p(\theta|\lambda) p(\lambda) d\lambda \tag{29}$$

On the other hand, using Bayes theorem, we have that

$$p(\theta|\lambda)p(\lambda) = p(\lambda|\theta)p(\theta) \Rightarrow p(\lambda) = \frac{p(\lambda|\theta)p(\theta)}{p(\theta|\lambda)}$$

Then, we can replace $p(\lambda)$ in (29) to obtain

$$\begin{aligned}\frac{\partial p(\theta)}{\partial \theta} &= \int [-\Sigma_\theta^{-1}(\lambda)(\theta - \mu_\theta(\lambda))] p(\theta|\lambda) \times \\ &\quad \frac{p(\lambda|\theta)p(\theta)}{p(\theta|\lambda)} d\lambda \\ &= \int [-\Sigma_\theta^{-1}(\lambda)(\theta - \mu_\theta(\lambda))] p(\lambda|\theta) p(\theta) d\lambda.\end{aligned} \tag{30}$$

Finally, we have that

$$\begin{aligned}\frac{1}{p(\theta)} \frac{\partial p(\theta)}{\partial \theta} &= \int [-\Sigma_\theta^{-1}(\lambda)(\theta - \mu_\theta(\lambda))] p(\lambda|\theta) d\lambda \\ &= \mathbb{E}_{\lambda|\theta}[-\Sigma_\theta^{-1}(\lambda)(\theta - \mu_\theta(\lambda))].\end{aligned} \tag{31}$$

Note that the expression in (31) is equal to $\frac{d \log p(\theta)}{d\theta}$.

(ii) The function $\log p(\theta)$ is known, which implies that $\frac{d \log p(\theta)}{d\theta}$ can be obtained directly. Therefore, evaluating both (31) and $\frac{d \log p(\theta)}{d\theta}$ on $\theta = \hat{\theta}^{(i)}$, we have:

$$\begin{aligned}\left.\frac{d \log p(\theta)}{d\theta}\right|_{\hat{\theta}^{(i)}} &= \mathbb{E}_{\lambda|\hat{\theta}^{(i)}}[-\Sigma_\theta^{-1}(\lambda)]\hat{\theta}^{(i)} + \\ &\quad \mathbb{E}_{\lambda|\hat{\theta}^{(i)}}[\Sigma_\theta^{-1}(\lambda)\mu_\theta(\lambda))].\end{aligned} \tag{32}$$

(iii) We notice that the expected values in (32) are the same required to compute the M-step. We also notice that we obtain a system of linear equations in (32), where the unknowns are given by the expected values. This linear system may or may not have closed form solution, depending on the number of unknowns and the number of equations. Therefore, the solution of this linear system of equations can be obtained by computing some expectations, analytically or numerically, enough to yield a system with the same number of unknowns and equations. This can be done, because $p(\lambda|\hat{\theta}^{(i)})$ is a known pdf. Then, the expected values can be computed either in closed form (when possible) or using Monte Carlo techniques (e.g. Metropolis-Hastings algorithm, Gibbs sampler, etc., see e,g, [13]).

A particular case is when $\log p(\theta)$ can be expressed as a function of θ (or equivalently of the "individual" terms of θ, θ_j, $j = 1,2,...$) as

$$\log p(\theta) = \sum_{j=1}^{\mathrm{p}} g\left(\frac{\theta_j}{\tau}\right), \tag{33}$$

If we consider the relationship given by MAP estimation and regularization, common regularization terms can be obtained assuming Gaussian mixtures for the *prior* distribution. For instance, in Ridge regression, the function g corresponds to $g(\theta_j/\tau) = (\theta_j/\tau)^2$, and for Lasso, $g(\theta_j/\tau) = |\theta_j/\tau|$. Notice that in the regularization framework, τ represents the factor that controls the strength of the regularization..

Remark 1: The resulting algorithm for the derivative of the auxiliary function $\mathscr{Q}_{\text{prior}}(\theta, \hat{\theta}^{(i)})$ can be summarized as:

$$\frac{\partial \mathscr{Q}_{\text{prior}}(\theta, \hat{\theta}^{(i)})}{\partial \theta_j} = \frac{1}{\theta_j^{(i)}} \dot{g}(\theta_j)\big|_{\theta_j = \hat{\theta}_j^{(i)}} \theta_j. \tag{34}$$

▽

Lemma 2: if the mixture is obtained with $\theta_j|\lambda_j \sim \mathscr{N}(0, \tau^2\lambda_j)$, then

$$\frac{\partial \mathscr{Q}_{\text{prior}}(\theta, \hat{\theta}^{(i)})}{\partial \theta} = -\frac{1}{\tau^2}\mathbf{E}\theta, \tag{35}$$

where $\mathbf{E} = \text{diag}\left(\mathbb{E}_{\lambda_1|\hat{\theta}_1^{(i)}}\{\lambda_1^{-1}\}, ..., \mathbb{E}_{\lambda_L|\hat{\theta}_L^{(i)}}\{\lambda_L^{-1}\}\right)$, and $\mathbb{E}_{\lambda_j|\hat{\theta}_j^{(i)}}\{\lambda_j^{-1}\} = -\tau \,\text{sign}(\hat{\theta}_j^{(i)})/\hat{\theta}_j^{(i)}$.

Proof: Using (34) and that $\dot{g}(\theta_j)\big|_{\theta_j = \hat{\theta}_j^{(i)}} = \text{sign}(\hat{\theta}_j^{(i)})/\tau$ for a Laplace pdf, then we can obtain the expression in (35). ■

F. Evaluation of $\mathscr{Q}_{ML}(\theta, \hat{\theta}^{(i)})$ and its derivative

The E-step of the EM algorithm given in (14) can be expressed as

$$\begin{aligned}\mathscr{Q}_{\text{ML}}(\theta, \hat{\theta}^{(i)}) &= \mathbb{E}\{\log[p(\mathbf{y}, \mathbf{x})]|\mathbf{y}, \hat{\theta}^{(i)}\} \\ &= \mathbb{E}\{\log[p(\mathbf{y}|\mathbf{x})]|\mathbf{y}, \hat{\theta}^{(i)}\} + \mathbb{E}\{\log[p(\mathbf{x})]|\mathbf{y}, \hat{\theta}^{(i)}\} \\ &= K_y - \frac{1}{2}\mathbb{E}\{(\mathbf{y} - \mathbf{M}(\varepsilon)h)^T \Sigma_y^{-1} \times \\ &\quad (\mathbf{y} - \mathbf{M}(\varepsilon)h)|\mathbf{y}, \hat{\theta}^{(i)}\} + \mathbb{E}\{\log[p(\mathbf{x})]|\mathbf{y}, \hat{\theta}^{(i)}\}\end{aligned} \tag{36}$$

where $\mathbf{M}(\varepsilon)$ is a (matrix) function of the parameter ε, and $K_y = -0.5N_C \log(2\pi) + N_C \log(\rho)$. Here, we have that in $\mathbb{E}\{\log[p(\mathbf{x}|\beta)]|\mathbf{y}, \hat{\theta}^{(i)}\}$, $p(\mathbf{x}|\beta)$ can in general be any distribution. Hence, our algorithm is not restricted to any particular $p(\mathbf{x}|\beta)$.

When we take derivatives of $\mathscr{Q}_{\text{ML}}(\theta, \hat{\theta}^{(i)})$ with respect to φ, ρ, and β we obtain:

$$\begin{aligned}\frac{\partial \mathscr{Q}_{\text{ML}}}{\partial \varphi} &= -[-\rho \mathbb{E}\{\mathbf{M}(\varepsilon)|\mathbf{y}, \hat{\theta}^{(i)}\}^T \mathbf{y} + \\ &\quad \mathbb{E}\{\mathbf{M}(\varepsilon)^T \mathbf{M}(\varepsilon)|\mathbf{y}, \hat{\theta}^{(i)}\}\varphi],\end{aligned} \tag{37}$$

$$\frac{\partial \mathscr{Q}_{\text{ML}}}{\partial \rho} = \frac{N}{\rho} - [\rho \mathbf{y}^T \mathbf{y} - \varphi^T, \mathbb{E}\{\mathbf{M}(\varepsilon)|\mathbf{y}, \hat{\theta}^{(i)}\}^T \mathbf{y} \tag{38}$$

$$\frac{\partial \mathscr{Q}_{\text{ML}}}{\partial \beta} = \mathbb{E}\{\frac{1}{p(\mathbf{x}|\beta)} \frac{\partial p(\mathbf{x}|\beta)}{\partial \beta}|\mathbf{y}, \hat{\theta}^{(i)}\}. \tag{39}$$

If we assume that $\mathbf{x}|\beta \sim \mathscr{N}(0, \beta\mathbf{I})$, and using the property in (3), then the expectations on the right hand side of (37), (38), and (39) can be readily calculated by applying Kalman filtering to the model in (2).

Remark 2: Since we are interested in promoting sparsity in $\mathbf{h}$ (or equivalently in φ), it is possible to combine (37), and (35), obtaining $\varphi = \varphi(\varepsilon)$, that is,

$$\varphi = [\mathbb{E}\{\mathbf{M}(\varepsilon)^T\mathbf{M}(\varepsilon)|\mathbf{y},\hat{\theta}^{(i)}\} + \frac{\mathbf{E}}{\tau^2}]^{-1} \times \rho\mathbb{E}\{\mathbf{M}(\varepsilon)|\mathbf{y},\hat{\theta}^{(i)}\}^T\mathbf{y}. \tag{40}$$

That is, we only optimize with respect to a reduced number of parameters (β, ε, and ρ). The solution to (37), (38), and (39) can be obtained by using generalized EM (GEM) algorithms, applying a similar procedure to the one used in [8].

G. Combination of $\mathscr{Q}_{ML}(\theta,\hat{\theta}^{(i)})$ and $\mathscr{Q}_{prior}(\theta,\hat{\theta}^{(i)})$

The solution to the MAP estimation problem (considering the reparametrization) presented in this paper can be summarized in the following steps:

EM algorithm promoting sparsity, with τ known:

(i) Start with $\hat{\gamma}^{(i)} = [\hat{\mathbf{h}}^{(i)}, \text{vec}(\hat{\beta}^{(i)}), \hat{\varepsilon}^{(i)}, \hat{\sigma}^{(i)}]$, and form the new parameter vector $\hat{\theta}^{(i)} = [\hat{\varphi}^{(i)}, \text{vec}(\hat{\beta}^{(i)}), \hat{\varepsilon}^{(i)}, \hat{\rho}^{(i)}]$ where $\hat{\varphi}^{(i)} = \hat{\mathbf{h}}^{(i)}/\hat{\sigma}^{(i)}$, and $\hat{\rho}^{(i)} = 1/\hat{\sigma}^{(i)}$,
(ii) with the values for $\hat{\theta}^{(i)}$, equal (39) to zero and optimize for β, obtaining $\hat{\beta}^{(i+1)}$,
(iii) with the values for $\hat{\varphi}^{(i)}$, $\hat{\beta}^{(i+1)}$, $\hat{\varepsilon}^{(i)}$ $\hat{\rho}^{(i)}$ from (i) and (ii), optimize for ε after replacing (40) in (22),
(iv) with the estimate $\hat{\varepsilon}^{(i+1)}$ obtained from (iii), replace in (40) and obtain a new estimate $\hat{\varphi}^{(i+1)}$,
(v) with the new estimates $\hat{\varphi}^{(i+1)}$, $\hat{\beta}^{(i+1)}$, $\hat{\varepsilon}^{(i+1)}$ previously obtained, find $\hat{\sigma}^{(i+1)}$ from making zero the right-hand side of (39), and solving a quadratic equation,
(vi) set $i \to i+1$, and go back to (ii) until convergence.

H. Estimation of τ

So far, the proposed algorithm solves the MAP estimation problem assuming τ known, or at least for a good estimate of it. Its knowledge is important for accurate estimates of the sparse parameter $\mathbf{h}$, and having an *a priori* knowledge of this parameter is not always possible. It is known that this value controls the strength of the regularization, hence its value is of importance to estimate other parameters of interest in the system.

The reparametrization introduced in §IV-D also affects the *a priori* distribution for $\mathbf{h}$ in (10), which can be now re-written as:

$$p(\mathbf{h}|\sigma) = \left(\frac{1}{2\tau\sigma}\right)^{\mathrm{m}} \exp\left\{-\frac{||\mathbf{h}||_1}{\tau\sigma}\right\}. \tag{41}$$

Lemma 3: Using (41), an estimate $\hat{\tau}$ is given by

$$\hat{\tau}^{(i+1)} = \frac{\mathbb{E}_{||\mathbf{h}||_1/\sigma|\mathbf{y},\hat{\tau}^{(i)}}\left\{\frac{||\mathbf{h}||_1}{\sigma}\right\}}{\mathrm{p}}. \tag{42}$$

Proof: We define an auxiliary function $\mathscr{Q}(\tau,\tau^{(i)}) = \mathrm{p}\log(\tau^{-1}/2\sigma) - \tau^{-1}\mathbb{E}_{||\mathbf{h}||_1/\sigma|\mathbf{y},\hat{\tau}^{(i)}}\left\{\frac{||\mathbf{h}||_1}{\sigma}\right\}$, take the derivative with respect to τ^{-1}, and then equal to zero. ∎

In general, obtaining $\mathbb{E}_{||\mathbf{h}||_1/\sigma|\mathbf{y},\hat{\tau}^{(i)}}\left\{\frac{||\mathbf{h}||_1}{\sigma}\right\}$ is computational expensive, requiring, in addition, many observations $\mathbf{y}$. To circumvent this problem, we propose the following approximation:

$$\hat{\tau} \approx \frac{||\hat{\mathbf{h}}_{\mathrm{ML}}||_1}{\mathrm{p}\,\hat{\sigma}_{\mathrm{ML}}}, \tag{43}$$

where we have considered perfect knowledge of the input signals $\mathbf{x}$ and $\mathbf{u}$, and $\hat{\mathbf{h}}_{\mathrm{ML}}$ and $\hat{\sigma}_{\mathrm{ML}}$ are the estimates using ML, i.e. without the *prior* pdf for $\mathbf{h}$. For the sake of comparison, we have carried out several simulations regarding the estimation of τ. We have noticed that, in our problem of interest, the estimator proposed in this paper outperforms the estimator obtained using the Empirical Bayes approach.

The solution to the MAP estimation problem presented in this paper (considering the reparametrization) can be summarized in the following additional steps:

EM algorithm promoting sparsity, with τ unknown:

(i) find $\hat{\tau}$ by using no regularization term, and calculate $\hat{\tau}$ as in (43). Alternatively, empirical Bayes can be used. However, a larger number of measurements are needed to obtain a good estimate,
(ii) given that we now have an estimation for τ, we can apply same procedure as described in §IV-G.

V. Numerical Example

In this section, we consider two numerical examples that are inspired in communication systems. We consider the model given in (2) and (3). The nature of communication systems leads to having a partial knowledge of the transmitted signal. That is, only a reduced number of samples of the input signal is known. In the class of systems presented in this paper, this corresponds to $\mathbf{u} = 0$ and having a few samples from $\mathbf{x}$. Following a set-up of communication systems, we have the following conditions for the simulation:

- We consider a reduced number of measurements points, i.e. $N = 64$,
- For simplicity, the distribution of the input signal $\mathbf{x}$ ($\mathscr{N}(0,\mathbf{I})$) is assumed known.
- We consider $\mathbf{h} = [h_0\, h_1\, \ldots\, h_{\mathrm{p}-1}]^T \in \mathbb{R}^p$ with $p \in \mathbb{N}$. In particular, we consider $\mathrm{p} = 40$,
- $f(\varepsilon) = \begin{bmatrix} \cos\left(\text{diag}\left(\frac{\varepsilon k}{N}\right)\right) & 0 \\ 0 & \sin\left(\text{diag}\left(\frac{\varepsilon k}{N}\right)\right) \end{bmatrix}$, with $k = 0,1,\ldots,(N-1)/2$,
- $\mathbf{A}(\mathbf{x},\mathbf{u}) = \tilde{\mathbf{H}}\mathbf{x}$, where $\tilde{\mathbf{H}}$ is a circulant matrix composed by $\mathbf{h}$,
- $\eta \sim \mathscr{N}(0,\sigma^2\mathbf{I}_N)$ is additive white Gaussian noise (AWGN). We consider 2 cases: (i) when σ^2 is known, and (ii) when σ^2 is unknown.

A. Example 1: τ known

In our first example, we consider that τ is known. Therefore, we focus on the estimation of $\gamma = [\mathbf{h}^T, \varepsilon]^T$.

In Table I, our simulations compare both, the MAP estimate (with a Laplace *prior* pdf) and the ML estimate. We observe that we obtain better estimates of $\mathbf{h}$ when we include the *prior* distribution. This is somehow expected, since we are estimating the parameter $\mathbf{h}$ considering more information.

TABLE I

MSE PERFORMANCE FOR MAP AND ML ESTIMATION. KNOWN NOISE VARIANCE σ_η^2.

Approach	number of known samples of $\mathbf{x}$	Number of EM iterations	MSE
ML	64	50	2.92×10^{-4}
	40	200	2.8×10^{-3}
MAP	64	50	3.16×10^{-4}
	40	200	2.9×10^{-3}

TABLE II

MSE PERFORMANCE FOR MAP AND ML ESTIMATION. UNKNOWN NOISE VARIANCE σ^2.

Approach	number of known samples of $\mathbf{x}$	Number of EM iterations	MSE
ML	100%	200	2.94×10^{-4}
	62.5%	400	8.21×10^{-3}
MAP	100%	200	3.27×10^{-4}
	62.5%	400	4.41×10^{-3}

B. Example 2: τ unknown

In this section, we consider the same estimation problem § V-A, but considering now that τ needs to be estimated using (43). For the estimation of τ, we obtain $\hat{\tau}\approx 0.455$. The vector of parameters to be estimated is then $\gamma=[\mathbf{h}^T,\varepsilon,\sigma]^T$. The application of the algorithm gives the results in Table II. As we can observe, there is little difference in the MSE compared to Table I.

VI. CONCLUSIONS

We have shown a general sparse parameter estimation method, based on the utilization of mixtures for representing the *a priori* distribution of the sparse parameters. In particular, we have used Gaussian variance-mean mixtures, which can represent a wide range of distributions. This representation was inserted in the generalized EM algorithm. As a result, an auxiliary function $\mathcal{Q}_{\text{MAP}}$ was developed for the *a priori* distribution. In terms of the *a priori*, this generates an E-step that is quadratic with respect to the parameters, and, hence, simple to optimize. A great emphasis was put on sparse parameter estimation, which was promoted via a Laplace distribution. We have given expressions to estimate all the parameter involved for a particular class of sparse systems. This class is important in, for example, the estimation of communication channels. In particular, we have concentrated the cost function and then numerically optimize the M-step.

The numerical example illustrates the benefits of using the proposed Bayesian approach over the ML approach. In fact, we have seen that including the *prior* pdf yields a lower MSE for the estimated sparse parameter, compared with the ML estimate. In addition, we estimate the parameter defining the *prior* distribution of the estimates (τ), showing no different conclusions in terms of accuracy for the estimation of the sparse parameter $\mathbf{h}$.

REFERENCES

[1] A. Aravkin, J. Burke, A. Chiuso, and G. Pillonetto. On the MSE properties of empirical Bayes methods for sparse estimation. In *16th IFAC Symposium on System Identification*, Brussels, Belgium, 2012.

[2] L. Baldassarre, J.M. Morales, and M. Pontil. Incorporating additional constraints in sparse estimation. In *16th IFAC Symposium on System Identification*, Brussels, Belgium, 2012.

[3] O. Barndorff-Nielsen, J. Kent, and M. Sorensen. Normal variance-mean mixtures and z distributions. *Int. Stat. Review*, 50(2):145–159, 1982.

[4] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):23132351, 2007.

[5] R. Carvajal. *EM-based channel estimation for multicarrier communication systems*. PhD thesis, The University of Newcastle, Australia, 2013.

[6] R. Carvajal, J.C. Agüero, B.I. Godoy, and G.C. Goodwin. EM-based channel estimation in muticarrier systems with phase distorsion. *IEEE Trans. Veh. Technol.*, 62(1):152–160, 2013.

[7] R. Carvajal, B.I. Godoy, J.C. Agüero, and G.C. Goodwin. EM-based sparse channel estimation in OFDM systems. In *13th IEEE Workshop on Signal Process. Adv. in Wireless Commun.*, 2012.

[8] B.I. Godoy, G.C. Goodwin, J.C. Agüero, D. Marelli, and T. Wigren. On identification of fir systems having quantized output data. *Automatica*, 46(9):1905–1915, 2011.

[9] J. Keilson and F.W. Steutel. Mixtures of distributions, moment inequalities and measures of exponentiality and normality. *The Annals of Probability*, 2(1):112–130, 1974.

[10] D.B. Kilfoyle and A.B. Baggeroer. The state of the art in underwater acoustic telemetry. *IEEE J. Oceanic Eng.*, 25(1):4–27, 2000.

[11] S. Kotz, T. J. Kozubowski, and K. Podgórski. *The Laplace distribution and generalizations: A revisit with applications to communications, economics, engineering and finance*. Birkhäuser, 2001.

[12] E.G. Larsson and Y. Selén. Linear regression with a sparse parameter vector. *IEEE Trans. Signal Process.*, 55(2):451–460, 2007.

[13] J.S. Liu. *Monte Carlo strategies in scientific computing*. Springer, New York, 2004.

[14] G.J. Mclachlan and T. Krishnan. *The EM algorithm and Extensions*. John Wiley & Sons, Inc., 2nd edition, 2008.

[15] A. F. Molisch. Ultrawideband propagation channels – Theory, measurement, and modeling. *IEEE Trans. Veh. Technol.*, 54(5):1528–1545, 2005.

[16] S. Özen, W. Hillery, M. D. Zoltowski, S. M. Nereyanuru, and M. Fimoff. Structured channel estimation based decision feedback equalizers for sparse multipath channels with applications to digital tv receivers. In *36th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, November 2002.

[17] T. Park and G. Casella. The bayesian lasso. *J. Amer. Statist. Assoc.*, 103(482):681–686, June 2008.

[18] N.G. Polson and J.G. Scott. Data augmentation for non-gaussian regression models using variance-mean mixtures. Available online. doi: 10.1093/biomet/ass081, 2013.

[19] N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models. *Test*, 19:209–256, 2010.

[20] G. Tauböck and F. Hlawatsch. A compressed sensing technique for ofdm channel estimation in mobile environments: Exploting channel sparsity for reducing pilots. In *ICASSP*, 2008.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B*, 58(1):267–288, 1996.

[22] J. Welsh, C. Rojas, H. Hjalmarsson, and B. Wahlberg. Sparse estimation techniques for basis function selection in wideband system identification. In *16th IFAC Symposium on System Identification*, Brussels, Belgium, 2012.

[23] M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

[24] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68(1):49–67, 2006.

[25] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(2):301–320, 2005.