16th IFAC Symposium on System Identification
The International Federation of Automatic Control
Brussels, Belgium. July 11-13, 2012

On the estimation of hyperparameters for Empirical Bayes estimators: Maximum Marginal Likelihood vs Minimum MSE

A. Aravkin * **J.V. Burke** ** **A. Chiuso** *** **G. Pillonetto** ***

* *Department of Earth and Ocean Sciences, University of British Columbia (e-mail: saravkin@eos.ubc.ca)*
** *Department of Mathematics, University of Washington (e-mail: burke@math.washington.edu)*
*** *Dept. of Information Engineering, University of Padova (e-mail: chiuso,giapi@dei.unipd.it)*

Abstract: It has been recently argued that linear system identification can be tackled in a Bayesian framework provided a suitable class of priors is considered. These priors essentially encode stability of the system but have to be flexible enough to adapt to a wide range of situations. Part of this flexibility is achieved introducing hyperparameters in the prior distribution which have to be estimated from data. In this paper we study the properties of a class of empirical Bayes estimators in terms of their Mean Squared Error. We do so in a simplified scenario which however captures some of the essential features arising in system identification.

1. INTRODUCTION

Bayesian methods for system identification, whose origins can be perhaps traced back to the seventies and eighties [Doan et al., 1984, Kitagawa and Gersh, 1984], have been subject of significant progress in the last few years [Pillonetto and De Nicolao, 2010, Pillonetto et al., 2011a,b, Chen et al., 2011].

A common feature of these methods is the selection of a prior distribution for the unknown parameters [1], which often takes a preassigned form (e.g. the covariance called stable spline Kernel in [Pillonetto and De Nicolao, 2010, Pillonetto et al., 2011a]) but depends on a few hyperparameters [2] that have to be estimated from data. Then, the estimated prior is used in the Bayesian estimator. This approach is rather common in Bayesian statistic and goes under the name of Empirical Bayes method [Maritz and Lwin, 1989].

In [Pillonetto and De Nicolao, 2010, Pillonetto et al., 2011a,b] the hyperparameters are estimated using the so called marginal likelihood, where the dependence on the unknown parameters has been integrated out. This marginalization naturally takes into account the effect of uncertainty in the estimated parameters, which is described by their posterior distribution given the data.

In order to get more insight into this estimation procedure we have decided to simplify the problem assuming that the "unknown" system is linear and described by a finite number of parameters (e.g. an FIR as in [Chen et al., 2011]) and, in addition, their prior distribution is taken to be zero mean Gaussian with a covariance which is proportional to the identity matrix. This simplified setup still captures the essential features which we want to study.

In this paper we focus on the single input case. We shall see that in the asymptotic regime (as the number of measurements tend to infinity) maximizing the marginal likelihood is equivalent to minimizing a weighted version of the Mean Squared Error (MSE) on the parameters.

Note that studies of Bayes estimators under a squared loss criterion can be found in early papers such as [Efron and Morris, 1973] and are of course related to the so called "James-Stein" estimators [James and Stein, 1961], [Stein, 1981].

The structure of the paper is as follows: Section 2 introduces the class of estimators we consider, in Section 3 we report the main result which provides a link between marginal likelihood maximization and the Mean Squared Error. Then Section 4 discusses the relation between estimation of hyperparameters based on either marginal likelihood maximization or minimization of (weighted versions of) the MSE. Simulations are reported in Section 5 and conclusions end the paper.

2. EMPIRICAL BAYES ESTIMATORS BASED ON MARGINAL LIKELIHOOD OPTIMIZATION

We consider a linear measurement model of the form

$$y = G\theta + v \quad y \in \mathbb{R}^n \quad \theta \in \mathbb{R}^m \tag{1}$$

where v is the vector whose components are white noise of known variance σ^2.

In the system identification scenario one can think of θ as the coefficients of an FIR system and of G as the Hankel

[1] For simplicity of exposition we refer to the system to be estimated with the term "unknown parameters" even though it might be infinite dimensional.
[2] The word *hyperparameters* is used to denote the the parameters which describe the prior distribution.

978-3-902823-06-9/12/$20.00 © 2012 IFAC

10.3182/20120711-3-BE-2027.00353

matrix containing the input samples; note that (1) may describe a MISO model. In fact, as explained in [Aravkin et al., 2011b] the explanatory factors G used to predict y can be grouped, where the groups correspond to different inputs. As such θ can be partitioned into p sub-vectors $\theta^{(i)}$, $i = 1, \dots, p$, so that

$$\theta = [\theta^{(1)\top} \quad \theta^{(2)\top} \quad \dots \quad \theta^{(p)\top}]^\top. \tag{2}$$

In the paper [Aravkin et al., 2011a], with an eye to the problem of input selection, we study the MSE properties of these Bayes estimators with particular attention to the tradeoffs between sparsity and shrinking.

Hereafter, for the sake of exposition, we consider the case $p = 1$ and consider a class of Bayes estimators defined as follows. Let λ be a random variable, independent of the measurement noise, which is given an improper prior on $\mathbb{R}^+$ that includes information only on its positivity. Then, we model θ (conditionally on λ) as a zero-mean Gaussian vector with covariance [3]

$$\theta|\lambda \sim N(0, \lambda I_m) \tag{3}$$

In order to find an estimator of θ one first optimizes the marginal density of λ, and then again using an empirical Bayes approach, the minimum variance estimate of θ is computed with λ taken as known and set to its estimate. This is described in the following theorem.

Theorem 1. Consider the Bayesian prior in (3) and the measurement model given by (1). Define

$$\hat{\lambda} = \arg\max_{\lambda \in \mathbb{R}_+} \int_{\mathbb{R}^m} p(\theta, \lambda|y) d\theta \tag{4}$$

Then, $\hat{\lambda}$ is given by

$$\arg\min_{\lambda \in \mathbb{R}_+} \frac{1}{2} \log\det(\Sigma_y(\lambda)) + \frac{1}{2} y^\top \Sigma_y^{-1}(\lambda) y \tag{5}$$

where

$$\Sigma_y(\lambda) := \lambda G G^\top + \sigma^2 I \tag{6}$$

In addition, the estimate of θ, denoted $\hat{\theta}$, is given by setting $\lambda = \hat{\lambda}$ in the function

$$\theta(\lambda) := \mathbb{E}[\theta|y, \lambda] = \lambda G^\top (\Sigma_y(\lambda))^{-1} y. \tag{7}$$

■

The derivation of this estimate can be found in [Aravkin et al., 2011b] and is omitted for reasons of space.

Let the vector μ denote the dual vector for the constraint $\lambda \geq 0$. Then the Lagrangian for the problem (5) is given by

$$L(\lambda, \mu) := \frac{1}{2} \log\det(\Sigma_y(\lambda)) + \frac{1}{2} y^\top \Sigma_y(\lambda)^{-1} y - \mu\lambda \tag{8}$$

Using the fact that

$$\partial_\lambda L(\lambda, \mu) = \frac{1}{2} \mathrm{tr}\left(G^\top \Sigma_y(\lambda)^{-1} G\right) - \frac{1}{2} y^\top \Sigma_y(\lambda)^{-1} G G^\top \Sigma_y(\lambda)^{-1} y - \mu,$$

we obtain the following KKT conditions for (5).

[3] The results in this paper can be easily extended to priors of the form $\theta|\lambda \sim N(0, \lambda Q)$; however for sake of exposition we prefer to work with $Q = I_m$.

Proposition 2. The necessary conditions for λ to be a solution of (5) are

$$\begin{aligned} &\Sigma_y = \sigma^2 I + \lambda G G^\top \\ &W \Sigma_y = I \\ &\mathrm{tr}\left(G^\top W G\right) - \|G^\top W y\|_2^2 - 2\mu = 0, \\ &\mu\lambda = 0, \\ &0 \leq \mu, \ \lambda \text{ and } 0 \preceq W, \Sigma_y \end{aligned} \tag{9}$$

It is interesting to observe that one has

$$\mathbb{E}\left[\theta(\lambda)\theta(\lambda)^\top \mid \lambda\right] = \lambda^2 G^\top W G. \tag{10}$$

In addition

$$\|\theta(\lambda)\|^2 = \lambda^2 \|G^\top W y\|_2^2,$$

Equation (9) indicates that when tuning λ there should be a link between the "norm" of the actual estimator $\|\hat{\theta}(\lambda)\|^2$ to its a priori second moments (10). In particular, when the nonnegativity constraint is not active, i.e. $\mu = 0$, one finds that the optimal value of λ makes the norm of the estimator equal to (the trace of) its a priori matrix of second moments.

Remark 3. If the parameter vector θ were assumed to have a covariance of the form λQ then one should use the weighted norm $\|x\|_Q^2 := x^\top Q^{-1} x$ instead. Analogously, if the parameter θ is an infinite dimensional object in the Reproducing Kernel Hilbert Space (RKHS hereafter, Wahba [1990]) $\mathcal{H}$, which is of interest in the system identification scenario in which θ is an impulse response Pillonetto and De Nicolao [2010], Pillonetto et al. [2011a], Chiuso and Pillonetto [2011], then the 2-norm has to be replaced with the norm in the RKHS.

3. MEAN SQUARED ERROR PROPERTIES OF EMPIRICAL BAYES ESTIMATORS BASED ON MARGINAL LIKELIHOOD OPTIMIZATION

Our aim is to evaluate the performance of an estimator $\hat{\theta}$ using its Mean Squared Error (MSE) i.e. its expected quadratic loss

$$\mathrm{tr}\left[\mathbb{E}\left[\left(\hat{\theta} - \bar{\theta}\right)\left(\hat{\theta} - \bar{\theta}\right)^\top \,\middle|\, \lambda, \theta = \bar{\theta}\right]\right],$$

where $\bar{\theta}$ is the "true" but unknown value of θ. When we speak about "Bayes estimators" we think of estimators of the form $\hat{\theta}(\lambda) := \mathbb{E}[\theta \,|\, y, \lambda]$ computed using the probabilistic model (3).

We begin by deriving an expression for the MSE of the Bayes estimators $\hat{\theta}(\lambda) := \mathbb{E}[\theta \,|\, y, \lambda]$. In this section, it is convenient to introduce the following notation

$$\mathbb{E}_v[\,\cdot\,] := \mathbb{E}[\,\cdot \,|\, \lambda, \theta = \bar{\theta}] \quad \text{and} \quad \mathrm{Var}_v[\,\cdot\,] := \mathbb{E}[\,\cdot \,|\, \lambda, \theta = \bar{\theta}].$$

Proposition 4. Consider the model (1) under the probabilistic model described by (3). The Mean Squared Error of the Bayes estimator $\hat{\theta}(\lambda) := \mathbb{E}[\theta|y, \lambda]$ given λ, when $\theta = \bar{\theta}$, is

$$\begin{aligned} MSE(\lambda) &= \mathrm{tr}\left[\mathbb{E}_v\left[(\hat{\theta}(\lambda) - \bar{\theta})(\hat{\theta}(\lambda) - \bar{\theta})^\top\right]\right] \\ &= \mathrm{tr}\left[\sigma^2 R^{-1}(\lambda) P(\lambda, \bar{\theta}) R^{-1}(\lambda)\right]. \end{aligned} \tag{11}$$

where

$$R(\lambda) := G^\top G + \sigma^2 \lambda^{-1}, \quad P(\lambda, \bar{\theta}) := G^\top G + \sigma^2 \lambda^{-2} \bar{\theta}\bar{\theta}^\top.$$

Proof. See [Aravkin et al., 2011b].

In the sequel with think of G in (1) as the non deterministic regressor matrix $G_n(\omega)$, independent of the noise v, defined on the complete probability space $(\Omega, \mathcal{B}, \mathbb{P})$ with ω a generic element of Ω and $\mathcal{B}$ the sigma field of Borel regular measures. In particular, the rows of G_n are independent [4] realizations from a zero-mean random vector with covariance Ψ having strictly positive and distinct eigenvalues denoted by $\{d_k^2\}_{k=1}^m$ ($d_k > 0,\ k = 1, \dots, m$). We will also assume that the (mild) assumptions for the almost sure convergence of $G_n^\top G_n/n$ to Ψ, as n goes to ∞, are satisfied, see e.g. [Loève, 1963].

Remark 5. Under the stated assumptions, the MSE in (11) is a random variable which depends on n, therefore we shall denote it as $MSE_n(\lambda)$. This implies that also its minimizer

$$\hat{\lambda}_n := \arg\min_{\lambda \in \mathbb{R}^+} MSE_n(\lambda)$$

is a random variable that depends on n.

To simplify the notation, hereafter the dependence on ω is omitted. The SVD of $G_n/\sqrt{n}$ is denoted by $U_n D_n V_n^\top$ where $D_n := \mathrm{diag}\{d_{k,n}\}$ and the columns of U_n are restricted to be of unit length with one entry in each of its columns constrained to be positive. This, together with the assumption of distinct eigenvalues, ensures that the decomposition is unique [5], e.g. see the discussion in Section 4 of [Bauer, 2005].

The measurement model can now be rewritten as follows

$$z_{k,n} = d_{k,n}\eta_{k,n} + e_{k,n}, \qquad k = 1, \dots, m \tag{12}$$

where $\{d_{k,n}\}_{k=1}^m$ are the singular values of $G_n/\sqrt{n}$, $z_{k,n}$ and $\eta_{k,n}$ are the k-th entry of the vectors $z_n = (U_n^\top y_n)/\sqrt{n}$ and $\eta_n = V_n^\top \theta$, respectively, while $e_n = [e_{1,n} \quad \dots \quad e_{m,n}]^\top$ is white Gaussian noise of variance σ^2/n.

Note that λ and θ are seen as parameters, and the "true" value of θ is $\bar{\theta}$ with $\|\bar{\theta}\| > 0$. Hence, all the randomness present in the next formulas comes only from G_n and the measurement noise.

Define $\bar{\eta}_n = V_n^\top \bar{\theta}$ for all n. After simple computations, from (11) one finds that $MSE(\lambda)$ relative to $\bar{\eta}_n$ is the following random variable whose statistics depend on n (with all the randomness due to G_n):

$$\begin{aligned} MSE_n(\lambda) &= \frac{\sigma^2}{n}\sum_{k=1}^m \frac{d_{k,n}^2\lambda^2 + \bar{\eta}_{k,n}^2\sigma^2/n}{(d_{k,n}^2\lambda + \sigma^2/n)^2} \\ &= \frac{\sigma^2}{n}\sum_{k=1}^m d_{k,n}^{-2}\frac{\lambda^2 + \bar{\eta}_{k,n}^2 w_{k,n}}{(\lambda + w_{k,n})^2}\ , \end{aligned} \tag{13}$$

where $w_{k,n} := \sigma^2/(n d_{k,n}^2)$ for all k and n. Let λ_n^{wopt} denote the minimizer of the following weighted version of the $MSE_n(\lambda)$:

$$\lambda_n^{wopt} = \arg\min_{\lambda \in \mathbb{R}_+} \sum_{k=1}^m d_{k,n}^2 \frac{\lambda^2 + \bar{\eta}_{k,n}^2 w_{k,n}}{(\lambda + w_{k,n})^2}\ .$$

[4] The independence assumption can be removed and replaced by mixing conditions.

[5] The hypothesis of distinct eigenvalues is made just to simplify the exposition. The case of repeated values would just require the introduction of more complicated constraints to ensure the uniqueness of the SVD

Defining $\Sigma_n := \lambda D_n^2 + (\sigma^2/n) I_m$, we find from (5) that our estimator for λ is given by

$$\begin{aligned} \hat{\lambda}_n &= \arg\min_{\lambda \in \mathbb{R}_+} \frac{1}{2} z_n^\top \Sigma_n^{-1} z_n + \frac{1}{2}\log\det(\Sigma_n) \\ &= \arg\min_{\lambda \in \mathbb{R}_+} \frac{1}{2}\sum_{k=1}^m \left[\frac{\zeta_{k,n}^2}{\lambda + w_{k,n}} + \log(\lambda + w_{k,n})\right] \end{aligned}$$

where $\zeta_{k,n} := z_{k,n}/d_{k,n}$ and, again, $w_{k,n} := \sigma^2/(n d_{k,n}^2)$ for all k and n.

Theorem 6. For almost all ω, it holds that

$$\lim_{n\mapsto\infty} \lambda_n^{wopt} = \frac{\|\bar{\theta}\|^2}{m} = \lim_{n\mapsto\infty} \hat{\lambda}_n\ .$$

Proof. As noted above, the SVD given above is uniquely defined under the assumptions on Ψ and G_n. Moreover, under these assumptions, the almost sure convergence of sample covariances (see [Loève, 1963]) combined with a perturbation result for the Singular Value Decomposition of symmetric matrices (such as Theorem 1 in [Bauer, 2005], see also [Chatelin, 1983]), implies that V_n and each $d_{k,n}$ converge almost surely to V and d_k, respectively. Consequently, $(\bar{\eta}_n, \zeta_n, w_n)$ converges almost surely to $(V^T\bar{\theta}, V^T\bar{\theta}, 0)$.

Assuming $\lambda_n^{wopt} > 0$, the first-order optimality conditions for λ_n^{wopt} can be written as $f(\lambda, w_n, \eta_n) = 0$, where $f : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is given by

$$f(\lambda, w, \eta) := \sum_{k=1}^m \frac{\lambda - \eta_k^2}{(\lambda + w_k)^3}.$$

Similarly, assuming $\hat{\lambda}_n > 0$, the first-order optimality conditions for $\hat{\lambda}_n$ can be written as $f_0(\lambda, w_n, \zeta_n) = 0$, where $f_0 : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is given by

$$f_0(\lambda, w, \zeta) := \sum_{k=1}^m \frac{\lambda + w_k - \zeta_k^2}{(\lambda + w_k)^2}.$$

Since $\|\bar{\theta}\| = \|V\bar{\theta}\|$, we have

$$f\left(\frac{\|\bar{\theta}\|^2}{m}, 0, V\bar{\theta}\right) = 0 \quad \text{and} \quad f_0\left(\frac{\|\bar{\theta}\|^2}{m}, 0, V\bar{\theta}\right) = 0\ ,$$

with

$$\begin{aligned} \frac{\partial}{\partial\lambda} f\left(\frac{\|\bar{\theta}\|^2}{m}, 0, V\bar{\theta}\right) &= \frac{m^4}{\|\bar{\theta}\|^6} \\ \frac{\partial}{\partial\lambda} f_0\left(\frac{\|\bar{\theta}\|^2}{m}, 0, V\bar{\theta}\right) &= \frac{m^3}{\|\bar{\theta}\|^4}. \end{aligned}$$

Applying the Implicit Function Theorem to both f and f_0 at $\left(\frac{\|\bar{\theta}\|^2}{m}, 0, V\bar{\theta}\right)$ yields the existence neighborhoods $\mathcal{U}$ of $(0, V\bar{\theta})$ and $\mathcal{W}$ of $\|\bar{\theta}\|^2/m$ as well as uniquely defined continuously differentiable functions $\phi : \mathcal{U} \to \mathcal{W}$ and $\phi_0 : \mathcal{U} \to \mathcal{W}$ such that

$$\begin{aligned} f(\phi(w,\eta), w, \eta) &= 0 \\ f_0(\phi_0(w,\eta), w, \eta) &= 0 \end{aligned} \qquad \forall\, (w,\eta) \in \mathcal{U}\ .$$

In particular, $\phi(0, V\bar{\theta}) = \|\bar{\theta}\|^2/m = \phi_0(0, V\bar{\theta})$. Since $(w_n, V_n^\top\bar{\theta})$ and (w_n, ζ_n) both converge to $(0, V^\top\bar{\theta})$ almost surely, we know that for almost all ω both $(w_n, V_n^\top\bar{\theta})$ and (w_n, ζ_n) are in $\mathcal{U}$ for all n sufficiently large. Therefore, for almost all ω,

$$\lambda_n^{wopt} = \phi(w_n, V_n^\top\bar{\theta}) > 0 \quad \text{and} \quad \hat{\lambda}_n = \phi_0(w_n, \zeta_n) > 0$$

since $\|\bar{\theta}\|^2/m > 0$ by assumption. Hence, both λ_n^{wopt} and $\hat{\lambda}_n$ almost surely converge to $\|\bar{\theta}\|^2/m$ proving the result.

4. COMPARISON BETWEEN MARGINAL LIKELIHOOD MAXIMIZATION AND MINIMUM MSE ESTIMATORS

In the previous Sections we have seen that the empirical Bayes estimator based on maximizing the marginal likelihood possesses some interesting properties in terms of achieved MSE on estimated parameters.

In particular Theorem 6 implies that the estimator of λ based on maximization of the marginal likelihood converges to the estimator of λ which would be obtained minimizing a weighted version of the MSE. Note that this latter estimator would require knowledge of "true" parameter $\bar{\theta}$.

The following questions arise quite naturally:

(1) which is the meaning of this "weighted" MSE?
(2) what if instead of maximizing the marginal likelihood, one finds an estimate of λ minimizing and estimate of the MSE (or of a weighted version of the MSE)?

As for point 1 above, it is useful to provide also the connection between the weighted MSE version introduced in the previous section and the MSE of the output prediction. This is done in the following proposition.

Proposition 7. Let us consider the linear measurement model (1), where we make explicit the dependence on the data length n

$$y_n = G_n\theta + v_n$$

and consider the Mean Squared Error on the output prediction:

$$\begin{aligned} MSE_n^y(\lambda) &= \mathbb{E}_v\left[(y_n - \hat{y}_n(\lambda))^\top (y_n - \hat{y}_n(\lambda))\right] \\ &= \mathbb{E}_v\left[(y_n - G_n\hat{\theta}(\lambda))^\top (y_n - G_n\hat{\theta}(\lambda))\right] \\ &= \mathrm{tr}\left\{\mathbb{E}_v\left[G_n(\hat{\theta}(\lambda)-\bar{\theta})(\hat{\theta}(\lambda)-\bar{\theta})G_n^\top\right]\right\} + \\ &\quad +\sigma^2 n \end{aligned} \tag{14}$$

as well as its weighted version:

$$\begin{aligned} MSE_n^W(\lambda) &= \mathbb{E}_v\left[(y_n - \hat{y}_n(\lambda))^\top G_n G_n^\top (y_n - \hat{y}_n(\lambda))\right] \\ &= \mathbb{E}_v\left[(y_n - G_n\hat{\theta}(\lambda))^\top G_n G_n^\top (y_n - G_n\hat{\theta}(\lambda))\right] \\ &= \mathrm{tr}\left\{\mathbb{E}_v\left[(G_n^\top G_n)^2(\hat{\theta}(\lambda)-\bar{\theta})(\hat{\theta}(\lambda)-\bar{\theta})\right]\right\} + \\ &\quad +\sigma^2 \mathrm{tr}(G_n G_n^\top) \end{aligned} \tag{15}$$

Then

$$MSE_n^y(\lambda) = \frac{\sigma^2}{n}\sum_{k=1}^{m}\frac{\lambda^2 + \bar{\eta}_{k,n}^2 w_{k,n}}{(\lambda + w_{k,n})^2} \tag{16}$$

and

$$MSE_n^W(\lambda) = \frac{\sigma^2}{n}\sum_{k=1}^{m} d_{k,n}^2\frac{\lambda^2 + \bar{\eta}_{k,n}^2 w_{k,n}}{(\lambda + w_{k,n})^2} \tag{17}$$

The proof follows the same argument as the derivation of (13) and is therefore omitted. ◇

Hence, the maximum marginal likelihood estimator asymptotically minimizes the weighted version of the output Mean Squared Error $MSE_n^W(\lambda)$.

As far as the second question above, the answer is not entirely trivial. First of all it is not clear which estimate of the MSE one should consider. For instance if $m > n$ one cannot even estimate, say in the least squares sense, θ. Somewhat arbitrarily, in the following we shall make the following assumptions:

(i) $n > m$ and $G_n^\top G_n$ is of full rank so that the Least Squares estimator is $\hat{\theta}_n^{LS} := (G_n^\top G_n)^{-1}G_n^\top y_n$;
(ii) one find estimators

$$\widehat{MSE}_n(\lambda),\ \widehat{MSE}_n^y(\lambda),\ \widehat{MSE}_n^W(\lambda) \tag{18}$$

replacing the "true" but unknown value $\bar{\theta}$ in (11) with the Least Square Estimator $\hat{\theta}_n^{LS}$.

Some discussion is now in order: since the marginal likelihood is written in terms of $y_n - G_n\hat{\theta}$ where θ has been integrated out, it would seem quite natural that its maximization would involve a measure of how well the estimator $\theta(\hat{\lambda})$ would perform (on average) on output data. Thus one may be tempted to conjecture that maximizing the marginal likelihood is related to minimizing the "output" Mean Squared Error $MSE_n^y(\lambda)$. As we have seen this is not so and one has to add some extra weighting ($MSE_n^W(\lambda)$) to establish the relation (Theorem 6) between marginal likelihood maximization and MSE. However one has to keep in mind that the Mean Squared error is not available and, at best, only its estimate could be minimized. Of course estimating the MSE induces uncertainty also in the estimate. Ideally, when estimating the MSE, one should also introduce some form of "weighting" accounting for how well θ has been estimated. When the coefficients $d_{k,n}$ do not satisfy $d_{k,n} = c_n$ (which happens if $G_n^\top G_n \propto I_n$) the components of the estimators for θ have different uncertainties and this have to be accounted for. Unfortunately a sharp argument giving solid grounds to these conjectures is out of reach at the moment. For this reason we shall resort to some simulation experiments to investigate how estimators based on marginal likelihood maximization and MSE (eventually weighted) minimization compare.

In particular we shall consider the following four estimators:

(1) $\hat{\theta}(\hat{\lambda}_n)$ where $\hat{\lambda}_n$ maximizes the marginal likelihood;
(2) $\hat{\theta}(\hat{\lambda}_n^{wopt})$ where $\hat{\lambda}_n^{wopt}$ minimizes $\widehat{MSE}_n^W(\lambda)$ in (7);
(3) $\hat{\theta}(\hat{\lambda}_n^y)$ where $\hat{\lambda}_n^y$ minimizes $\widehat{MSE}_n^y(\lambda)$ in (7);
(4) $\hat{\theta}(\hat{\lambda}_n^\theta)$ where $\hat{\lambda}_n^\theta$ minimizes $\widehat{MSE}_n(\lambda)$ in (7);

These estimators are then compared using both MSE_n and MSE_n^y as criterions (one may be interested in estimating either θ or the output y).

The simulation results are reported in the next Section.

5. SIMULATION RESULTS

As anticipated in the previous Section we now report results comparing Empirical Bayes estimators using different estimators for the prior hyperparameter λ. We consider two different experimental setups:

(1) The true value $\bar{\theta} \in \mathbb{R}^{10}$ has all components equal to 1. The matrix $G_n \in \mathbb{R}^{10\times 10}$ is taken as a diagonal matrix with elements logarithmically spaced between

10^{-1} and 10 and the noise standard deviation is fixed to $\sigma = 0.1$.

(2) The true value of θ is fixed to $\bar{\theta} = [1\ 2\ 10\ 0\ -3\ -1\ 0\ -5\ 0\ 1]$. The matrix $G_n \in \mathbb{R}^{10\times 10}$ is taken as a diagonal matrix with elements logarithmically spaced between 10^{-1} and 10 and the noise standard deviation is fixed to $\sigma = 1$.

For each of these we perform 1000 Montecarlo experiments randomizing the noise realization; for each of these Montecarlo runs we estimated the parameter θ as described above. The histograms of the achieved MSE are reported in Figures 1 and 2.

The experimental results suggest that utilizing the marginal likelihood provides a smaller sensitivity to noise: in fact the histograms of achieved MSE are much more concentrated. Of course there is price to pay; in fact in both examples 3 and 4 the maximization of the marginal likelihood never reaches the minimum of the MSE which could be attained with knowledge of the true $\bar{\theta}$. It is also worth stressing that for values of $\bar{\theta}$ different from those here employed the performance of ML could get worse and estimating λ optimizing $\widehat{MSE}_n^y(\lambda)$ or $\widehat{MSE}_n(\lambda)$ could lead to better results. This point will deserve future investigation in the future.

6. CONCLUSIONS

We have presented an analysis of the asymptotic properties of marginal likelihood maximization in terms of MSE of the resulting empirical Bayes estimators. It has been shown that maximizing the marginal likelihood corresponds, asymptotically, to minimizing a weighted version of the MSE. We have also compared through numerical simulations different strategies for hyperparameter estimation via minimization of the (weighted) MSE. Future work will concentrate on (i) more general prior description possibly including Stable Spline Kernels [Pillonetto and De Nicolao, 2010, Pillonetto et al., 2011a] and (ii) on providing solid theoretical grounds for the simulation results.

ACKNOWLEDGEMENTS

This research has been partially supported by the PRIN grant n. 20085FFJ2Z "New Algorithms and Applications of System Identification and Adaptive Control" by the Progetto di Ateneo CPDA090135/09 funded by the University of Padova, by the European Community's Seventh Framework Programme [FP7/2007-2013] under agreement n. FP7-ICT-223866-FeedNetBack and under grant agreement n257462 HYCON2 Network of excellence

REFERENCES

A. Aravkin, J. Burke, A. Chiuso, and G. Pillonetto. On the MSE properties of empirical bayes methods for sparse estimation. Technical report, University of Padova, 2011a. submitted to IFAC SYSID 2012.

A. Aravkin, J. Burke, A. Chiuso, and G. Pillonetto. Convex vs nonconvex approaches for sparse estimation: GLasso, Multiple Kernel Learning and Hyperparameter GLasso. Technical report, University of Padova, 2011b. submitted to Journal of Machine Learning Research.

D. Bauer. Asymptotic properties of subspace estimators. *Automatica*, 41:359–376, 2005.

F. Chatelin. *Spectral approximation of linear operators*. Academic Press, NewYork, 1983.

T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularization and gaussian processes - revisited. In *IFAC World Congress 2011*, Milano, 2011.

A. Chiuso and G. Pillonetto. A Bayesian approach to sparse dynamic network identification. Technical report, University of Padova, 2011. *submitted to Automatica, available at http://automatica.dei.unipd.it/people/chiuso.html*.

T. Doan, R. Litterman, and C.A. Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3:1–100, 1984.

B. Efron and C. Morris. Stein's estimation rule and its competitors–an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

W. James and C. Stein. Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 361–379. Univ. California Press, Berkeley, Calif., 1961.

G. Kitagawa and H. Gersh. A smothness priors-state space modeling of time series with trends and seasonalities. *Journal of the American Statistical Association*, 79(386):378–389, 1984.

M. Loève. *Probability Theory*. Van Nostrand Reinhold, 1963.

J. S. Maritz and T. Lwin. *Empirical Bayes Method*. Chapman and Hall, 1989.

G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 45(2):291–305, 2011a.

G. Pillonetto, M.H. Quang, and A. Chiuso. A new kernel-based approach for nonlinear system identification. *IEEE Transactions on Automatic Control*, 2011b.

C.M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annalso of Statistics*, 9(6): 1135–1151, 1981.

G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

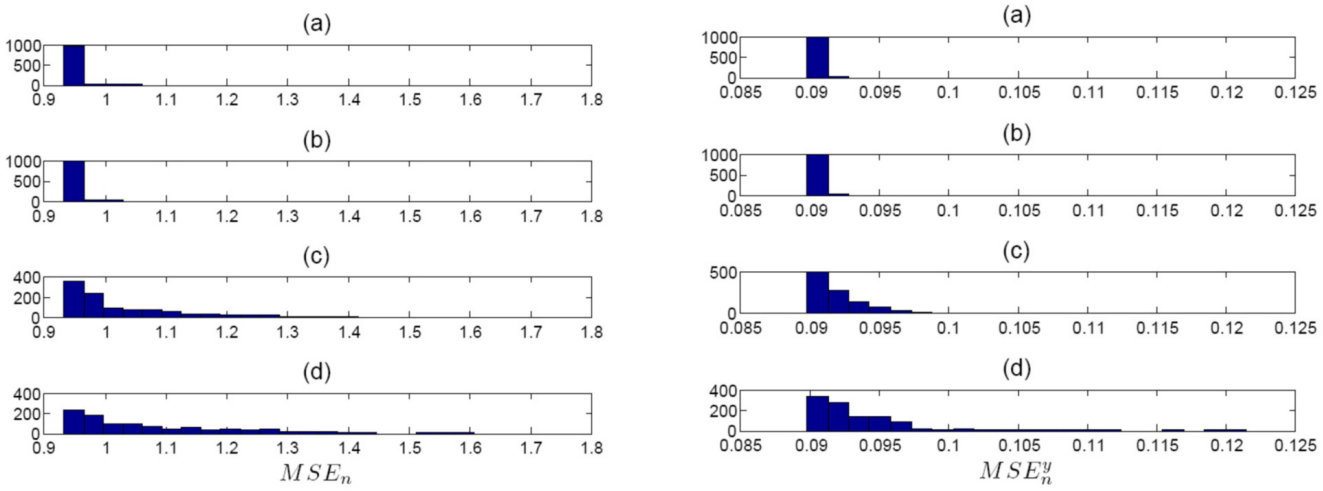

Fig. 1. Example #1: histogram of MSE_n (left) and MSE_n^y (right). Top to bottom: (a) using $\hat{\lambda}_n$ from maximizing marginal likelihood; (b) using $\hat{\lambda}_n^{wopt}$ which minimizes $\widehat{MSE}_n^W(\lambda)$; (c) using $\hat{\lambda}_n^y$ which minimizes $\widehat{MSE}_n^y(\lambda)$; (d) using $\hat{\lambda}_n^\theta$ which minimizes $\widehat{MSE}_n(\lambda)$.

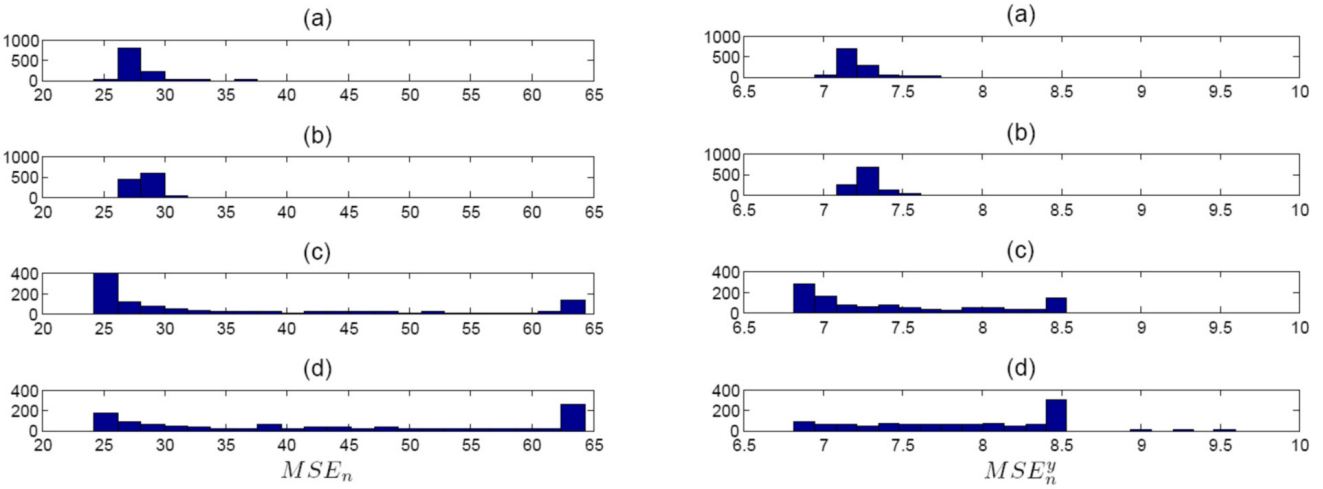

Fig. 2. Example #2: histogram of MSE_n (left) and MSE_n^y (right). Top to bottom: (a) using $\hat{\lambda}_n$ from maximizing marginal likelihood; (b) using $\hat{\lambda}_n^{wopt}$ which minimizes $\widehat{MSE}_n^W(\lambda)$; (c) using $\hat{\lambda}_n^y$ which minimizes $\widehat{MSE}_n^y(\lambda)$; (d) using $\hat{\lambda}_n^\theta$ which minimizes $\widehat{MSE}_n(\lambda)$.