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

On the MSE Properties of Empirical Bayes Methods for Sparse Estimation

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: Popular convex approaches for sparse estimation such as Lasso and Multiple Kernel Learning (MKL) can be derived in a Bayesian setting, starting from a particular stochastic model. In problems where groups of variables have to be estimated, we show that the same probabilistic model, under a suitable marginalization, leads to a different non-convex estimator where hyperparameters are optimized. Theoretical arguments, independent of the correctness of the priors entering the sparse model, are included to clarify the advantages of our non-convex technique in comparison with MKL and the group version of Lasso under assumption of orthogonal regressors.

1. INTRODUCTION

In this paper we investigate sparse estimation in a linear regression model where the explanatory factors $\theta \in \mathbb{R}^m$ are naturally grouped so that θ is partitioned as $\theta = [\theta^{(1)^\top} \quad \theta^{(2)^\top} \quad \dots \quad \theta^{(p)^\top}]^\top$. In this context we assume that θ is group (or block) sparse, i.e. many of the constituent vectors $\theta^{(i)}$ are zero or have a negligible influence on the output $y \in \mathbb{R}^n$. In addition, we assume that the number of unknowns m is large, possibly larger than the size of the available data n. Interest in general sparsity estimation and optimization has attracted the interest of many researchers in statistics, machine learning, and signal processing with numerous applications in feature selection, compressed sensing, and selective shrinkage [Hastie and Tibshirani, 1990, Tibshirani, 1996, Donoho, 2006, Candes and Tao, 2007]. The motivation for our study of the group sparsity problem comes from the dynamic Bayesian network scenario identification problem as discussed in [Chiuso and Pillonetto, 2010b,a, 2012]. In a dynamic network scenario the explanatory variables are often the past histories of different input signals with the groups $\theta^{(i)}$ representing the impulse responses which describe the relationship between the i-th input and the output y. This application informs our view of the group sparsity problem as well as our measures of success for a particular estimation procedure.

Many approaches have been proposed in the literature for joint estimation and variable selection problems. We cite the well known Lasso [Tibshirani, 1996], Least Angle Regression (LAR) [Efron et al., 2004], their group versions Group Lasso (GLasso) and Group Least Angle Regression (GLAR) [Yuan and Lin, 2006], Multiple Kernel Learning (MKL) [Bach et al., 2004, Evgeniou et al., 2005, Pillonetto et al., 2010]. Methods based on hierarchical Bayesian models have also been considered such as Automatic Relevance Determination (ARD) [Mackay, 1994], the Relevance Vector Machine (RVM) [Tipping, 2001], and the exponential hyperprior in [Chiuso and Pillonetto, 2010b, 2012]. The Bayesian approach described in [Chiuso and Pillonetto, 2010b, 2012] and further developed in this paper is intimately related to [Mackay, 1994, Tipping, 2001]; in fact, the exponential hyperprior algorithm in [Chiuso and Pillonetto, 2010b, 2012] is a penalized version of ARD.

An interesting series of papers [Wipf and Rao, 2007, Wipf and Nagarajan, 2007, Wipf et al., 2011] provide a nice link between penalized regression problems like Lasso, also called type-I methods, and Bayesian methods (like RVM [Tipping, 2001] and ARD [Mackay, 1994]) with hierarchical hyperpriors where the hyperparameters are estimated via maximizing the marginal likelihood and then inserted in the Bayesian model following the Empirical Bayes paradigm [Maritz and Lwin, 1989]; these latter methods are also known as type-II methods [Berger, 1985]. Note that this Empirical Bayes paradigm has also been recently used in the context of System Identification [Pillonetto and De Nicolao, 2010, Pillonetto et al., 2011, Chen et al., 2011].

In [Wipf and Nagarajan, 2007, Wipf et al., 2011] it is argued that type-II methods have advantages over type-I methods; some of these advantages are related to the fact that, under suitable assumptions, the former can be written in the form of type-I with the addition of a non-separable penalty term (a function $g(x_1, .., x_n)$ is non-separable if it cannot be written as $g(x_1, \dots, x_n) = \sum_{i=1}^{n} = h(x_i)$). The analysis in [Wipf et al., 2011] also suggests that in the low noise regime the type-II approach results in a tighter approximation to the ℓ_0 norm.

This is supported by experimental evidence showing that these Bayesian approaches perform well in practice. Our experience is that the approach based on the marginal

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

10.3182/20120711-3-BE-2027.00301

likelihood is particularly robust w.r.t. noise regardless of the correctness of the Bayesian prior.

The scope of this work, which is also motivated by the stunning performance of the exponential hyperprior approach introduced in the dynamic network identification scenario [Chiuso and Pillonetto, 2010b, 2012], is to provide some new insights clarifying the above issues. In particular in the first part of the paper the relation among Lasso (and GLasso), the Exponential Hyperprior (HGLasso algorithm hereafter, for reasons which will become clear later on) and MKL is discussed by putting all these methods in a common Bayesian framework (similar to that discussed in [Park and Casella, 2008]). Both Lasso/GLasso and MKL boil down to convex optimization problems, while HGLasso does not. All these methods are then compared in terms of optimality (KKT) conditions and tradeoffs between sparsity and shrinkage are studied illustrating the advantages of HGLasso over GLasso and MKL assuming orthogonal regressors. In particular, the properties of Empirical Bayes estimators which form the basis of our computational scheme are studied in terms of their Mean Square Error properties. Such analysis avoids assumptions on the correctness of the priors entering the stochastic model and clarifies why HGLasso is likely to provide more sparse and accurate estimates in comparison with the other two convex estimators.

The paper is organized as follows. In Section 2 we introduce the HGLasso approach in a Bayesian framework. Section 3 introduces MKL. In Section 4 the Mean Squared Error properties of HGLasso and MKL are compared using orthogonal regressors. The analysis also includes the GLasso case, since, under orthogonal assumptions, the regularization paths of MKL and GLasso are the same, see [Aravkin et al., 2011]. Some conclusions then end the paper.

2. HGLASSO ESTIMATOR

We consider a linear measurement model of the form

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

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

For the reasons put forward in the introduction, we are interested in situations where the explanatory factors G used to predict y are grouped. As such we partition θ into p sub-vectors $\theta^{(i)}$, $i = 1, \dots, p$, so that

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

For $i = 1, \dots, p$, assume that the sub-vector $\theta^{(i)}$ has dimension k_i so that $m = \sum_{i=1}^p k_i$. Next, conformally partition the matrix $G = [G^{(1)}, \dots, G^{(p)}]$ to obtain the measurement model

$$y = G\theta + v = \sum_{i=1}^p G^{(i)}\theta^{(i)} + v. \tag{3}$$

In what follows, we assume that θ is *block sparse* in the sense that many of the blocks $\theta^{(i)}$ are null, i.e. with all of their components equal to zero, or have a negligible effect on y.

An possible approach to the block sparsity problem, discussed in [Chiuso and Pillonetto, 2010b], relies on the group version of the model in Fig. 1(a) illustrated in Fig. 1(b). In the network, λ is now a p-dimensional vector with independent and identically distributed components $\lambda_i \in \mathbb{R}_+$:

$$p_\gamma(\lambda_i) = \gamma e^{-\gamma\lambda_i}\chi(\lambda_i), \tag{4}$$

where γ is a positive scalar while $\chi(t) = 1$ if $t \geq 0$, 0 otherwise. In addition, conditional on λ, each block $\theta^{(i)}$ of the vector θ is zero-mean Gaussian with covariance [1] $\lambda_i I_{k_i}$, $i = 1, .., p$, i.e.

$$\theta^{(i)}|\lambda_i \sim N(0, \lambda_i I_{k_i}) \tag{5}$$

The proposed estimator 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. We call this scheme Hyperparameter Group Lasso (HGLasso). It is described in the following theorem.

Theorem 1. Consider the Bayesian network in Fig. 1(b) with measurement model given by (3), (5), and (4), and define

$$\hat{\lambda} = \arg\max_{\lambda \in \mathbb{R}_+^p} \int_{\mathbb{R}^m} p(\theta, \lambda|y)d\theta \tag{6}$$

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

$$\arg\min_{\lambda \in \mathbb{R}_+^p} \frac{1}{2}\log\det(\Sigma_y(\lambda)) + \frac{1}{2}y^\top\Sigma_y^{-1}(\lambda)y + \gamma\sum_{i=1}^p \lambda_i \tag{7}$$

where

$$\Sigma_y(\lambda) := G\Lambda G^\top + \sigma^2 I, \qquad \Lambda := \text{blockdiag}(\{\lambda_i I_{k_i}\}) \tag{8}$$

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

$$\theta_{HGL}(\lambda) := \mathbb{E}[\theta|y, \lambda] = \Lambda G^\top(\Sigma_y(\lambda))^{-1}y. \tag{9}$$

■

The derivation of this estimate can be found in [Aravkin et al., 2011] and is omitted for reasons of space. Note that the optimization (7) is performed in $\mathbb{R}^p$, rather than in $\mathbb{R}^m$ ($m > p$ for the group case) where the GLasso [Yuan and Lin, 2006] objective is optimized.

Let the vector μ denote the dual vector for the constraint $\lambda \geq 0$. Then the Lagrangian for the problem (7) is given by

$$\begin{aligned} L(\lambda, \mu) := \; & \frac{1}{2}\log\det(\Sigma_y(\lambda)) + \frac{1}{2}y^\top\Sigma_y(\lambda)^{-1}y + \\ & + \gamma\mathbf{1}^\top\lambda - \mu^\top\lambda \end{aligned} \tag{10}$$

Using the fact that

$$\begin{aligned} \partial_{\lambda_i} L(\lambda, \mu) = & \frac{1}{2}\text{tr}\left(G^{(i)\top}\Sigma_y(\lambda)^{-1}G^{(i)}\right) \\ & - \frac{1}{2}y^\top\Sigma_y(\lambda)^{-1}G^{(i)}G^{(i)\top}\Sigma_y(\lambda)^{-1}y + \gamma \; - \mu_i, \end{aligned}$$

we obtain the following KKT conditions for (7).

Proposition 2. The necessary conditions for λ to be a solution of (7) are

[1] The results in this paper can be easily extended to priors of the form $\lambda_i \sim N(0, \lambda_i Q_i)$; however for sake of exposition we prefer to work with $\Sigma_{k_i} = I_{k_i}$.

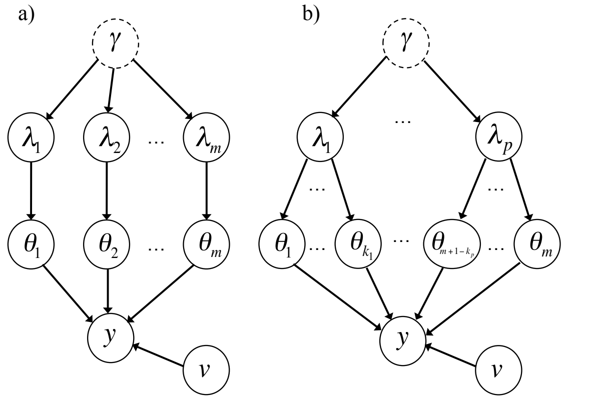

Fig. 1. Bayesian networks describing the stochastic model for sparse estimation (a) and group sparse estimation (b)

$$\begin{array}{l}\Sigma_y = \sigma^2 I + \sum_{i=1}^{p} \lambda_i G^{(i)} G^{(i)\top} \\ W\Sigma_y = I \\ \mathrm{tr}\left(G^{(i)\top} W G^{(i)}\right) - \|G^{(i)\top} W y\|_2^2 + 2\gamma - 2\mu_i = 0, \\ \mu_i \lambda_i = 0, \quad i = 1, \ldots, p \\ 0 \leq \mu, \ \lambda \text{ and } 0 \preceq W, \Sigma_y \end{array} \tag{11}$$

It is interesting to observe that one has

$$\mathbb{E}\left[\theta_{HGL}(\lambda)\theta_{HGL}(\lambda)^\top \mid \lambda\right] = \Lambda G^\top \Sigma_y(\lambda)^{-1} G \Lambda,$$

and so, for $i = 1, \ldots, p$,

$$\mathbb{E}\left[\theta^{(i)}_{HGL}(\lambda)\left(\theta^{(i)}_{HGL}(\lambda)\right)^\top \,\middle|\, \lambda\right] = \lambda_i^2 \left(G^{(i)\top} W G^{(i)}\right). \tag{12}$$

In addition

$$\|\theta^{(i)}_{HGL}(\lambda)\|^2 = \lambda_i^2 \|G^{(i)\top} W y\|_2^2, \quad i = 1, \ldots, p.$$

Equation (11) indicates that when tuning λ there should be a link between the "norm" of the actual estimator $\|\hat{\theta}^{(i)}(\lambda)\|^2$ to its a priori second moments (12). In particular, when no regularization is imposed on λ (i.e. $\gamma = 0$) and the nonnegativity constraint is not active, i.e. $\mu_i = 0$, one finds that the optimal value of λ_i makes the norm of the estimator equal to (the trace of) its a priori matrix of second moments.

Remark 3. If each block $\theta^{(i)}$ was assumed to have a covariance of the form $\lambda_i Q_i$ then one should use the weighted norm $\|x\|_{Q_i} := x^\top Q_i^{-1} x$ instead. Analogously, if the group $\theta^{(i)}$ are infinite dimensional objects in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, which is of interest in the system identification scenario in which each $\theta^{(i)}$ is an impulse response, then the 2-norm has to be replaced with the norm in the RKHS.

3. MULTIPLE KERNEL LEARNING

Multiple Kernel Learning (MKL) provides another approach to the block sparsity problem [Bach et al., 2004, Evgeniou et al., 2005, Dinuzzo, 2010, Bach, 2008]. To introduce this approach consider the measurements model

$$y = f + v = \sum_{i=1}^{p} f^{(i)} + v \ , \tag{13}$$

where v is as specified in (1). In the MKL framework, f represents the sampled version of a scalar function assumed to belong to a (generally infinite-dimensional) reproducing kernel Hilbert space (RKHS) [Wahba, 1990]. For our purposes, we consider a simplified scenario where the domain of the functions in the RKHS is the finite set $[1, \ldots, n]$. In this way, f represents the entire function and y is the noisy version of f sampled on all its whole domain. In addition, we assume that f belongs to the RKHS, denoted $\mathcal{H}_K$, having kernel defined by the matrix

$$K(\lambda) = \sum_{i=1}^{p} \lambda_i K^{(i)}, \tag{14}$$

where it is further assumed that each of the functions $f^{(i)}$ is an element of a RKHS, denoted $\mathcal{H}^{(i)}$, having kernel $\lambda_i K^{(i)}$ with associated norm denoted by $\|f^{(i)}\|_{(i)}$.

According to the MKL approach, the estimates of the unknown functions $f^{(i)}$ are obtained *jointly* with those of the scale factors λ_i by solving the following inequality constrained problem:

$$\begin{array}{c}(\{\hat{f}^{(i)}\}, \hat{\lambda}) = \underset{\{f^{(i)}\}, \lambda \in \mathbb{R}^p_+}{\arg\min} \dfrac{(y-f)^\top (y-f)}{\sigma^2} + \sum_{i=1}^{p} \|f^{(i)}\|^2_{(i)} \\ \text{s.t.} \quad \sum_{i=1}^{p} \lambda_i \leq M \ , \end{array} \tag{15}$$

where M plays the role of a regularization parameter. Hence, the "scale factors" contained in $\lambda \in \mathbb{R}^+_p$ are optimization variables, thought of as "tuning knobs" adjusting the kernel $K(\lambda)$ to better suit the measured data. Using the extended version of the representer theorem, e.g. see [Dinuzzo, 2010, Evgeniou et al., 2005], the solution is

$$\hat{f}^{(i)} = \hat{\lambda}_i K^{(i)} \hat{c}, \qquad i = 1, \ldots, p \tag{16}$$

where

$$\begin{array}{c}\{\hat{c}, \hat{\lambda}\} = \underset{c \in \mathbb{R}^n, \lambda \in \mathbb{R}^+_p}{\arg\min} \dfrac{(y - K(\lambda)c)^\top (y - K(\lambda)c)}{\sigma^2} + c^\top K(\lambda) c \\ \text{s.t.} \quad \sum_{i=1}^{p} \lambda_i \leq M \end{array} \tag{17}$$

It can be shown that every local solution the above optimization problem is also a global solution, see [Dinuzzo, 2010] for details.

In the following proposition we supply the KKT conditions for $\hat{\lambda}$ to be solutions of (17).

Proposition 4. The necessary and sufficient conditions for λ to be a solution of (17) are

$$\begin{array}{l}\Sigma_y = K(\lambda) + \sigma^2 I \\ W\Sigma_y = I \\ -\|G^{(i)\top} W y\|_2^2 + 2\gamma - 2\mu_i = 0, \quad i = 1, \ldots, p \\ \mu_i \lambda_i = 0, \quad i = 1, \ldots, p \\ 0 \leq \mu, \ \lambda \text{ and } 0 \preceq W, \Sigma_y \end{array} \tag{18}$$

■

4. MEAN SQUARED ERROR PROPERTIES OF EMPIRICAL BAYES ESTIMATORS

In this Section we evaluate the performance of an estimator $\hat{\theta}$ using its Mean Squared Error (MSE) i.e. its expected quadratic loss

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

where $\bar{\theta}$ is the "true" but unknown value of θ. When we speak about "Bayes estimators" we think of estimators of the form $\hat{\theta}(\lambda) := \mathbb{E}\left[\theta \,|\, y, \lambda\right]$ computed using the probabilistic model Fig. 1 with γ fixed.

We derive the MSE formulas under the simplifying assumption of "orthogonal" regressors ($G^{\top}G = nI$) and show that the Empirical Bayes estimator converges to an "optimal" estimator in terms of its MSE. This fact has close connections to the so called "Stein" estimators [James and Stein, 1961], [Stein, 1981], [Efron and Morris, 1973]. The same optimality properties are attained, asymptotically, when the columns of G are realizations of uncorrelated processes having the same variance. This is of interest in the system identification scenario considered in [Chiuso and Pillonetto, 2010a,b, 2012] since it arises when one performs identification with i.i.d. white noises as inputs.

We begin by deriving an expression for the MSE of the Bayes estimators $\hat{\theta}(\lambda) := \mathbb{E}\left[\theta \,|\, y, \lambda\right]$. 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 5. Consider the model (3) under the probabilistic model described in Fig. 1(b). The Mean Squared Error of the Bayes estimator $\hat{\theta}(\lambda) := \mathbb{E}\left[\theta|y,\lambda\right]$ given λ and $\theta = \bar{\theta}$ is

$$\begin{aligned} MSE(\lambda) &= \operatorname{tr}\left[\mathbb{E}_v\left[(\hat{\theta}(\lambda)-\theta)(\hat{\theta}(\lambda)-\theta)^{\top}\right]\right] \\ &= \operatorname{tr}\left[\sigma^2 R^{-1}(\lambda) P(\lambda,\bar{\theta}) R^{-1}(\lambda)\right]. \qquad (19) \end{aligned}$$

where

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

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

We can now minimize the expression for $MSE(\lambda)$ given in (19) with respect to λ to obtain the optimal minimum mean squared error estimator. In the case where $G^{\top}G = nI$ this computation is straightforward and is recorded in the following proposition.

Corollary 6. Assume that $G^{\top}G = nI$ in Proposition 5. Then MSE(λ) is globally minimized by choosing

$$\lambda_i = \lambda_i^{opt} := \frac{\|\bar{\theta}^{(i)}\|^2}{k_i}, \quad i = 1,\dots,p. \qquad (20)$$

Next consider the Maximum a Posteriori estimator of λ again under the simplifying assumption $G^{\top}G = nI$. Note that, under the noninformative prior ($\gamma = 0$), this Maximum a Posteriori reduces to the standard Maximum (marginal) Likelihood approach to estimating the prior distribution of θ. Consequently, we continue to call the resulting procedure Empirical Bayes (a.k.a. Type-II Maximum Likelihood, [Berger, 1985]).

Proposition 7. Consider model (3) under the probabilistic model described in Fig. 1(b), and assume that $G^{\top}G = nI$. Then the estimator of λ_i obtained maximizing the marginal posterior $\mathbf{p}(\lambda|y)$,

$$\begin{aligned} \{\hat{\lambda}_1(\gamma),...,\hat{\lambda}_p(\gamma)\} &:= \arg\max_{\lambda\in\mathbb{R}_+^p} \mathbf{p}(\lambda|y) \\ &= \arg\max_{\lambda\in\mathbb{R}_+^p} \int \mathbf{p}(y,\theta|\lambda)\mathbf{p}_{\gamma}(\lambda)\, d\theta, \end{aligned} \qquad (21)$$

is given by

$$\hat{\lambda}_i(\gamma) = \max\left(0, \frac{1}{4\gamma}\left[\sqrt{k_i^2 + 8\gamma\|\hat{\theta}_{LS}^{(i)}\|^2} - \left(k_i + \frac{4\sigma^2\gamma}{n}\right)\right]\right), \qquad (22)$$

where

$$\hat{\theta}_{LS}^{(i)} = \frac{1}{n}\left(G^{(i)}\right)^{\top} y$$

is the Least Squares estimator of the $i-$th block $\theta^{(i)}$. As $\gamma \to 0$ ($\gamma = 0$ corresponds to an improper flat prior) the expression (22) yields:

$$\lim_{\gamma\to 0}\hat{\lambda}_i(\gamma) = \max\left(0, \frac{\|\hat{\theta}_{LS}^{(i)}\|^2}{k_i} - \frac{\sigma^2}{n}\right). \qquad (23)$$

In addition, the probability $\mathbb{P}[\hat{\lambda}_i(\gamma) = 0 \,|\, \theta = \bar{\theta}]$ of setting $\hat{\lambda}_i = 0$ is given by

$$\mathbb{P}[\hat{\lambda}_i(\gamma) = 0 \,|\, \theta = \bar{\theta}] = \mathbb{P}\left[\chi^2_{k_i,\mu} \le \left(k_i + 2\gamma\frac{\sigma^2}{n}\right)\right], \qquad (24)$$

where $\chi^2(k_i,\mu)$ denotes a noncentral χ^2 random variable with d degrees of freedom and noncentrality parameter $\mu := \|\bar{\theta}^{(i)}\|^2\frac{n}{\sigma^2}$.

Proof. See [Aravkin et al., 2011]

Note that the expression of $\hat{\lambda}_i(\gamma)$ in Proposition 7 has the form of a "saturation". In particular, for $\gamma = 0$, we have

$$\hat{\lambda}_i(0) = \max(0, \hat{\lambda}_i^*), \quad \text{where} \quad \hat{\lambda}_i^* := \frac{\|\hat{\theta}_{LS}^{(i)}\|^2}{k_i} - \frac{\sigma^2}{n}. \qquad (25)$$

The following proposition shows that the "unsaturated" estimator $\hat{\lambda}_i^*$ is unbiased and consistent estimator of λ_i^{opt} which minimizes the Mean Squared Error while $\hat{\lambda}_i(0)$ is only asymptotically unbiased and consistent.

Corollary 8. Under the assumption $G^{\top}G = nI$, the estimator of $\hat{\lambda}^* := \{\lambda_1^*,..,\lambda_p^*\}$ in (25) is an unbiased and mean square consistent estimator of λ^{opt} which minimizes the Mean Squared Error, while $\hat{\lambda}(0) := \{\lambda_1(0),..,\lambda_p(0)\}$ is asymptotically unbiased and consistent, i.e.:

$$\mathbb{E}[\hat{\lambda}_i^* \,|\, \theta = \bar{\theta}] = \lambda_i^{opt} \qquad \lim_{n\to\infty}\mathbb{E}[\hat{\lambda}_i(0) \,|\, \theta = \bar{\theta}] = \lambda_i^{opt} \qquad (26)$$

and

$$\lim_{n\to\infty}\hat{\lambda}_i^* \overset{m.s.}{=} \lambda_i^{opt} \qquad \lim_{n\to\infty}\hat{\lambda}_i(0) \overset{m.s.}{=} \lambda_i^{opt} \qquad (27)$$

where $\overset{m.s.}{=}$ denotes convergence in mean square.

Proof. See [Aravkin et al., 2011]

Remark 9. Note that if $\bar{\theta}^{(i)} = 0$ the optimal value λ_i^{opt} is zero. Hence (27) shows that asymptotically $\hat{\lambda}_i(0)$ converges to zero. However, in this case, it is easy to see from (24) that

$$\lim_{n\to\infty}\mathbb{P}[\hat{\lambda}_i(0) = 0 \,|\, \theta = \bar{\theta}] < 1.$$

There is in fact no contradiction between these two statements because one can easily show that for all $\epsilon > 0$,

$$\mathbb{P}[\hat{\lambda}_i(0) \in [0, \epsilon) \,|\, \theta = \bar{\theta}] \overset{n\to\infty}{\longrightarrow} 1.$$

In order to guarantee that $\lim_{n\to\infty} \mathbb{P}[\hat{\lambda}_i(\gamma) = 0 \,|\, \theta = \bar{\theta}] = 1$ one must chose $\gamma = \gamma_n$ so that $2\frac{\sigma^2}{n}\gamma_n \to \infty$, so that γ_n grows faster than n. This is in line with the well known requirements for Lasso to be model selection consistent. In fact, Theorem 1 in [Aravkin et al., 2011] shows that the link between γ and the regularization parameter γ_L for Lasso is given by $\gamma_L = \sqrt{2\gamma}$. The condition $n^{-1}\gamma_n \to \infty$ translates into $n^{-1/2}\gamma_{Ln} \to \infty$, a well known condition for Lasso to be model selection consistent [Zhao and Yu, 2006, Bach, 2008].

The results obtained so far suggest that the Empirical Bayes resulting from HGLasso has desirable properties with respect to the MSE of the estimators. One wonder whether the same favorable properties are inherited by the Multiple Kernel Learning estimators. The next proposition shows that this is not the case. In fact, for $\bar{\theta}^{(i)} \neq 0$ MKL does yield consistent estimators for λ_i^{opt}; in addition, for $\theta^{(i)} = 0$ the probability of setting $\hat{\lambda}_i(\gamma)$ to zero (see equation (31)) is much smaller than that obtained using HGLasso (see equation (24)); this is also illustrated in Figure 2 (top). Also note that, as illustrated in Figure 2 (bottom), when the "true" θ is equal to zero MKL tends to give much larger values of $\hat{\lambda}$ than those given by HGLasso. This results in larger values of $\|\hat{\theta}\|$ (see Figure 2).

Proposition 10. Consider model (3) under the probabilistic model described in Fig. 1(b), and assume $G^\top G = nI$. Then the estimator of λ_i obtained by maximizing the joint posterior $\mathbf{p}(\lambda, \phi|y)$

$$\{\hat{\lambda}(\gamma), ..., \hat{\lambda}_p(\gamma)\} := \arg \max_{\lambda\in\mathbb{R}_+^p, \phi\in\mathbb{R}_+^m} \mathbf{p}(\lambda, \phi|y), \tag{28}$$

is given by

$$\hat{\lambda}_i(\gamma) = \max\left(0, \frac{\|\hat{\theta}_{LS}^{(i)}\|}{\sqrt{2\gamma}} - \frac{\sigma^2}{n}\right), \tag{29}$$

where

$$\hat{\theta}_{LS}^{(i)} = \frac{1}{n}\left(G^{(i)}\right)^\top y$$

is the Least Squares estimator of the $i-$th block $\theta^{(i)}$ for $i = 1, \ldots, p$. For $n \to \infty$ the estimator $\hat{\lambda}_i(\gamma)$ satisfies

$$\lim_{n\to\infty} \hat{\lambda}_i(\gamma) \overset{m.s.}{=} \frac{\|\bar{\theta}^{(i)}\|}{\sqrt{2\gamma}} \; . \tag{30}$$

In addition, the probability $\mathbb{P}[\hat{\lambda}_i(\gamma) = 0 \,|\, \theta = \bar{\theta}]$ of setting $\hat{\lambda}_i(\gamma) = 0$ is given by

$$\mathbb{P}_\theta[\hat{\lambda}_i(\gamma) = 0 \,|\, \theta = \bar{\theta}] = \mathbb{P}\left[\chi^2\left(k_i, \|\bar{\theta}^{(i)}\|^2\frac{n}{\sigma^2}\right) \leq 2\gamma\frac{\sigma^2}{n}\right] \; . \tag{31}$$

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

Note that the limit of the MKL estimators $\hat{\lambda}_i(\gamma)$ as $n \to \infty$ depends on γ. Therefore, using MKL, one cannot hope to get consistent estimators of λ_i^{opt}. Indeed, for $\|\bar{\theta}^{(i)}\|^2 \neq 0$, consistency of $\hat{\lambda}_i(\gamma)$ requires $\gamma \to \frac{k_i}{2\|\bar{\theta}^{(i)}\|^2}$, which is a circular requirement.

5. CONCLUSION

It has been shown that HGLasso and MKL derive from the same Bayesian model, yet in a different way. The HGLasso relies on a marginalized joint density with the resulting estimator involving optimization of a non-convex objective. However, the non-convex nature allows HGLasso to achieve higher levels of sparsity than MKL without introducing too much regularization in the estimation process. The MSE analysis reported in this paper, under assumptions of orthogonal regressors where MKL and GLasso share the same regularization paths, reveals the superior performance of HGLasso also in the reconstruction of the parameter groups different from zero. It shows the robustness of the empirical Bayes procedure, based on marginal likelihood optimization, independently of the correctness of the priors entering the stochastic model underlying HGLasso.

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. Convex vs nonconvex approaches for sparse estimation: glasso, multiple kernel learning and hyperparameter glasso. Technical report, University of Padova, 2011. submitted to Journal of Machine Learning Research.

F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, pages 41–48, 2004.

F.R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, second edition, 1985.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

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. Nonparametric sparse estimators for identification of large scale linear systems. In *Proceedings of IEEE Conf. on Dec. and Control*, Atlanta, 2010a.

A. Chiuso and G. Pillonetto. Learning sparse dynamic linear systems using stable spline kernels and exponential hyperpriors. In *Proceedings of Neural Information Processing Symposium*, Vancouver, 2010b.

A. Chiuso and G. Pillonetto. A Bayesian approach to sparse dynamic network identification. *Automatica, to appear*, 2012.

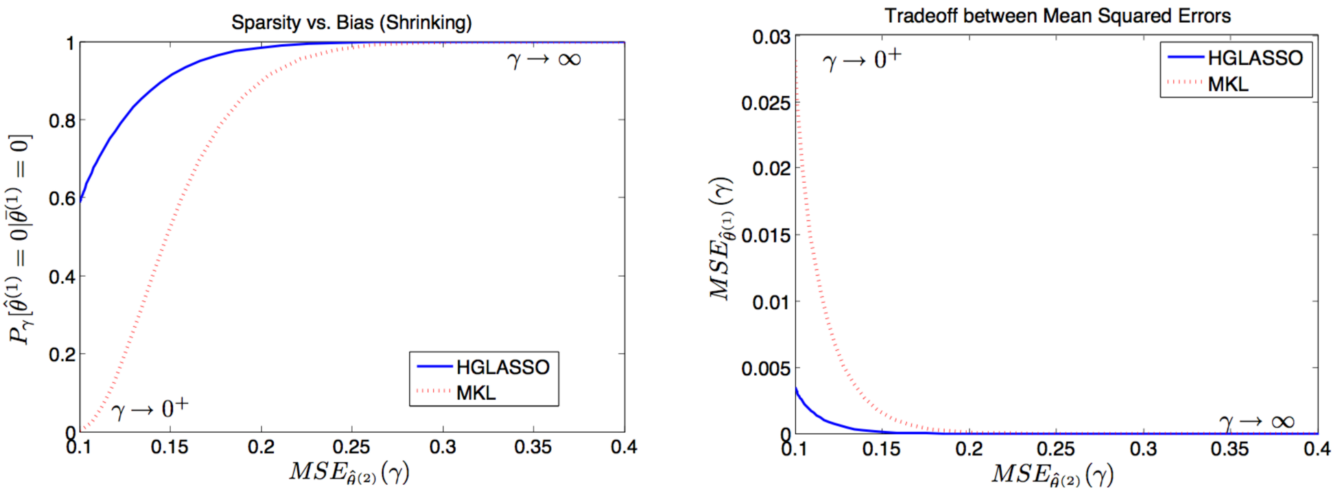

Fig. 2. This plot has been generated assuming that there are two blocks ($p = 2$) of dimension $k_1 = k_2 = 10$ with $\bar{\theta}^{(1)} = 0$ and all the components of the true $\bar{\theta}^{(2)} \in \mathbb{R}^{10}$ set to one.The matrix G equal to the identity, the noise variance equal to 0.1 and $n = 1$. Left: probability of setting $\hat{\theta}^{(1)}$ to zero vs Mean Squared Error in $\hat{\theta}^{(2)}$. Curves are parametrized in $\gamma \in [0, +\infty)$. Right: Mean Squared Error in $\hat{\theta}^{(1)}$ vs Mean Squared Error in $\hat{\theta}^{(2)}$. Curves are parametrized in $\gamma \in [0, +\infty)$.

F. Dinuzzo. Kernel machines with two layers and multiple kernel learning. *arXiv:1001.2709*, 2010.

D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

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.

B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

T. J. Hastie and R. J. Tibshirani. Generalized additive models. In *Monographs on Statistics and Applied Probability*, volume 43. Chapman and Hall, London, UK, 1990.

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.

D.J.C. Mackay. Bayesian non-linear modelling for the prediction competition. *ASHRAE Trans.*, 100(2):3704–3716, 1994.

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

T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, June 2008.

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

G. Pillonetto, F. Dinuzzo, and G. De Nicolao. Bayesian online multitask learning of gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):193–205, 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, 2011.

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

R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B.*, 58, 1996.

M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

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

D.P. Wipf and S. Nagarajan. A new view of automatic relevance determination. In *Proc. of NIPS*, 2007.

D.P. Wipf and B.D. Rao. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transactions on Signal Processing*, 55 (7):3704–3716, 2007.

D.P. Wipf, B.D. Rao, and S. Nagarajan. Latent variable Bayesian models for promoting sparsity. *IEEE Transactions on Information Theory (to appear)*, 2011.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, Nov. 2006.