Convex vs nonconvex approaches for sparse estimation: Lasso, Multiple Kernel Learning and Hyperparameter Lasso

Aleksander Aravkin, James V. Burke, Alessandro Chiuso and Gianluigi Pillonetto

Abstract— **We consider the problem of sparse estimation in a Bayesian framework. We outline the derivation of the Lasso in terms of marginalization of a particular Bayesian model. A different marginalization of the same probabilistic model leads also to a different nonconvex estimator where hyperparameters are optimized. The arguments are extended to problems where groups of variables have to be estimated. An approach alternative to Group Lasso is derived, also providing its connection with Multiple Kernel Learning approaches. Our estimator is nonconvex but one of its versions requires optimization with respect to only one scalar variable. Theoretical arguments and numerical experiments show that the new technique obtains sparse solutions which are more accurate than the other two convex estimators.**

Index Terms— **Lasso, Group Lasso, marginal density**

I. Introduction

We consider estimation of the parameters $\theta \in \mathbb{R}^m$ in a linear regression model. We also assume that the vector θ is sparse, i.e. many of its components are equal to zero or have a negligible influence on the output y, and that the number of "unknowns" m is is very large and possibly larger than the number of data available (say n, i.e. the number of "samples" available for statistical inference). In this scenario a key point is that the estimation procedure should be sparsity-favoring, i.e. able to extract from the large number of variables entering the model just that subset which influences the system output significantly. Linear problems of this sort are very general and have attracted the interest of many researchers in statistics, machine learning and signal processing; indeed, such a sparsity principle permeates many well known techniques in machine learning and signal processing such as feature selection, selective shrinkage and compressed sensing [10], [16], [6], [1].

We specifically became interested in a version of this problem since it also pops up in a "dynamic Bayesian network" identification scenario as discussed in [4], [2], [3]. Having this last application domain in mind, in this paper we shall be mainly concerned with a "group" version where the explanatory factors used to predict the output y can be grouped, i.e. the parameter vector θ can be partitioned as $\theta = [\theta^{(1)} \quad \theta^{(2)} \quad \dots \quad \theta^{(p)}]^\top$ To be concrete, in a dynamic network scenario the "explanatory variables" may be the past histories of different input signals and the "groups" $\theta^{(i)}$ be the impulse responses from the i-th input to the output y.

Several approaches have been put forward in the literature for joint estimation and variable selection problems. We cite the well known Lasso [16], Least Angle Regression (LAR), [7] their "group" versions Group Lasso (GLasso) and Group Least Angle Regression (GLAR) [19], Multiple Kernel Learning (MKL) [8], [12] as well as methods based on hierarchical Bayesian models such as the Relevance Vector Machine (RVM) [17] and the exponential hyperprior in [2]. Motivated by the stunning performance of the exponential hyperprior approach in the dynamic network identification scenario, see [2], [4], we believe an in depth comparison with other available methods is due. In this paper we initiate this comparison, discussing the relation among Lasso (and GLasso), the Exponential Hyperprior (HGLasso algorithm hereafter) and Multiple Kernel Learning by putting all these methods in a common Bayesian framework (similar to that discussed in [11]). Both Lasso/GLasso and MKL boil down to convex optimization problems, while HGLasso does not. However, one of the versions of HGLASSO here proposed requires optimization with respect to only one scalar variable. We discuss advantages and drawbacks of the nonconvex formulation and propose also a "forward selection" type of procedure for initializing the non-convex search, which may be seen as an instance of the "screening" type of approach for variable selection discussed in [18]. An optimization procedure for the HGLasso algorithm is also proposed.

II. Lasso and HLasso

Let $\theta = [\theta_1 \quad \theta_2 \dots \theta_m]^\top$ be an unknown parameter vector while $y \in \mathbb{R}^n$ denotes the vector containing some noisy data. In particular, our measurements model is

$$y = G\theta + v \tag{1}$$

where $G \in \mathbb{R}^{n \times m}$ and v is the vector whose components are white noise of known variance σ^2.

A. The Lasso approach

When θ is assumed to be sparse, i.e. many of its components are equal to zero or have a negligible influence on y, one popular approach to reconstruct the parameter vector is

A. Aravkin (saravkin@eos.ubc.ca) is with the Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, Canada.

J.V. Burke (burke@math.washington.edu) is with the Department of Mathematics, University of Washington, Seattle, USA.

G. Pillonetto (giapi@dei.unipd.it) is with the Department of Information Engineering, University of Padova, Padova, Italy.

A. Chiuso (chiuso@dei.unipd.it) is with the Department of Information Engineering, University of Padova, Padova, Italy.

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 and by the European Community's Seventh Framework Programme under agreement n. FP7-ICT-223866-FeedNetBack.

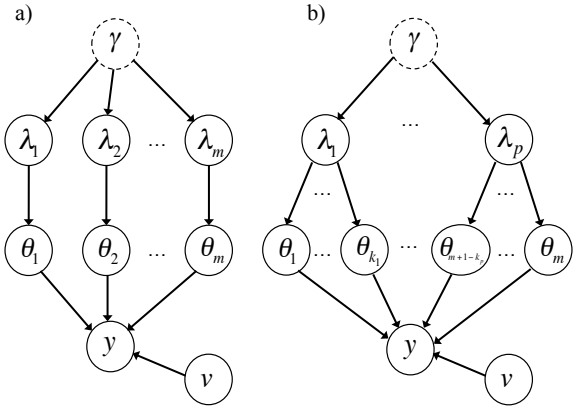

Fig. 1. Bayesian networks describing the stochastic model for sparse estimation (a) and group sparse estimation (b)

the so called Lasso [16]. It determines the estimate of θ as follows

$$\hat{\theta}_L = \arg\min_{\theta} \frac{(y-G\theta)^\top (y-G\theta)}{2\sigma^2} + \gamma_L \sum_{i=1}^{m} |\theta_i| \qquad (2)$$

where $\gamma_L \in \mathbb{R}_+$ is the regularization parameter. One can easily see that the above optimization problem is convex.

Now, we outline a derivation of the Lasso in terms of marginalization of a suitable probability density function, as also discussed in [11]. Our Bayesian model is depicted in Fig. 1(a). Nodes and arrows are either dotted or solid depending on being representative of, respectively, deterministic or stochastic quantities/relationships. Here, λ denotes a vector whose components $\{\lambda_i\}_{i=1}^m$ are independent exponential random variables, with the same probability density given by

$$p_\gamma(\lambda_i) = \gamma e^{-\gamma\lambda_i} \chi(\lambda_i) \qquad (3)$$

where γ is a positive scalar while $\chi(t) = 1$ if $t \geq 0$, 0 otherwise. In addition

$$\theta_i|\lambda_i \sim \mathcal{N}(0,\lambda_i), \qquad v \sim \mathcal{N}(0,\sigma^2 I_n) \qquad (4)$$

where $\mathcal{N}(\mu,\Sigma)$ is the Gaussian density of mean μ and autocovariance Σ while I_n is the $n \times n$ identity matrix. The following result then holds, see also Section 2 in [11] for details.

Proposition 1: Given the Bayesian network in Fig. 1(a), let

$$\hat{\theta} = \arg\max_{\theta\in\mathbb{R}^m} \int_{\mathbb{R}_+^m} p(\theta,\lambda|y)d\lambda \qquad (5)$$

Then $\hat{\theta} = \hat{\theta}_L$ provided that $\gamma_L = \sqrt{2\gamma}$.

B. The HLasso approach

The above result provides a hint for defining a different estimator. Instead of marginalizing with respect of λ, one could integrate out θ, finding the estimate of λ optimizing the marginal density $p(\lambda|y)$. Then, according to the empirical Bayes approach, the minimum variance estimate of θ is computed with λ set to its estimate. We call the resulting estimator Hyperparameter Lasso (HLasso). It is defined by the following proposition that exploits the fact that θ conditional on λ is Gaussian, so that the marginal density of λ becomes available in closed form.

Proposition 2: Given the Bayesian network in Fig. 1(a), let

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

Then

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

where

$$\Sigma_y = G\Lambda G^\top + \sigma^2 I_n, \qquad \Lambda = diag\{\lambda_i\}$$

Then, given $\lambda = \hat{\lambda}$, the HLasso estimate of θ is given by

$$\hat{\theta}_{HL} := \mathbb{E}[\theta|y,\hat{\lambda}] = \Lambda G^\top(\Sigma_y(\hat{\lambda}))^{-1}y \qquad (8)$$

■

Note that the objective in (7) used to determine λ depends on m variables as in the Lasso case but the optimization problem is not convex any more.

III. GLasso and HGLasso

We now consider a situation where explanatory factors able to predict y can be represented by groups of components contained in θ. To be more specific, we factorize θ as follows

$$\theta = [\theta^{(1)} \quad \theta^{(2)} \quad \dots \quad \theta^{(p)}]^\top \qquad (9)$$

and denote with k_i the dimension of the i-th block, so that $m = \sum_{i=1}^p k_i$. Partitioning also the matrix G as done for θ, we obtain the measurement model

$$y = \sum_{i=1}^{p} G^{(i)}\theta^{(i)} + v \qquad (10)$$

In what follows, we assume 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.

A. The GLasso approach

One of the leading approaches adopted to solve this problem is the so called Group Lasso (GLasso) [19]. It determines the estimate of θ as

$$\hat{\theta}_{GL} = \arg\min_{\theta\in\mathbb{R}^m} \frac{(y-G\theta)^\top(y-G\theta)}{2\sigma^2} + \gamma_{GL}\sum_{i=1}^{p}\|\theta^{(i)}\| \qquad (11)$$

where $\|\cdot\|$ denotes the classical Euclidean norm. It is easy to see that, as in the Lasso case, the objective is convex. However, as we will discuss in the next subsection, GLasso cannot be derived from the Bayesian models reported in Fig. 1.

The next proposition, taken from Section 2 in [19], characterizes $\hat{\theta}_{GL}$ by the Karush Kuhn Tucker (KKT) conditions.

Proposition 3: Assume that $G^{(i)\top}G^{(i)} = I_{k_i}$ for $i = 1,\dots,p$. Then, a necessary a sufficient condition for $\theta = [\theta^{(1)} \quad \theta^{(2)} \quad \dots \quad \theta^{(p)}]^\top$ to be a solution of (11) is

$$-G^{(i)\top}(y - G\theta) + \frac{\theta^{(i)}\gamma_{GL}\sigma^2}{\|\theta^{(i)}\|} = 0, \qquad \forall \theta^{(i)} \neq 0 \tag{12}$$

$$\|-G^{(i)\top}(y - G\theta)\| \leq \gamma_{GL}\sigma^2, \qquad \forall \theta^{(i)} = 0 \tag{13}$$

B. The HGLasso approach

The alternative approach we propose, discussed also in [2], relies upon the group version of that in Fig. 1(a) and is illustrated in Fig. 1(b). In the network, λ is now a p-dimensional vector with $i-th$ component given by $\lambda_i \in \mathbb{R}_+$. In addition, conditional on λ, each block $\theta^{(i)}$ of the vector θ is zero-mean Gaussian with covariance $\lambda_i I_{k_i}$, $i = 1,..,p$, i.e.

$$\theta^{(i)}|\lambda_i \sim N(0, \lambda_i I_{k_i}) \tag{14}$$

Then, the new estimator we propose first optimizes the marginal density of λ. Then, still according to the empirical Bayes approach, the minimum variance estimate of θ is computed with λ thought as known and set to its estimate. We call this scheme Hyperparameter Group Lasso (HGLasso). It is described in the following proposition.

Proposition 4: Consider the Bayesian network in Fig. 1 (b) and define

$$\hat{\lambda} = \arg\max_{\lambda \in \mathbb{R}_+^p} \int_{\mathbb{R}^m} p(\theta, \lambda | y) d\theta \tag{15}$$

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

$$\arg\min_{\lambda \in \mathbb{R}_+^p} \frac{1}{2}\log\det(\Sigma_y) + \frac{1}{2}y^\top \Sigma_y^{-1} y + \gamma \sum_{i=1}^{p} |\lambda_i| \tag{16}$$

where

$$\Sigma_y = G\Lambda G^\top + \sigma^2 I_n, \qquad \Lambda = \text{blockdiag}(\{\lambda_i I_{k_i}\}) \tag{17}$$

In addition, given $\lambda = \hat{\lambda}$, the HGLasso estimate of θ is given by

$$\hat{\theta}_{HGL} := \mathbb{E}[\theta | y, \hat{\lambda}] = \Lambda G^\top (\Sigma_y(\hat{\lambda}))^{-1} y \tag{18}$$

■

It can easily be seen that the objective in (16) used to determine λ is not convex. However, the optimization must be performed in $\mathbb{R}^p$, in place of $\mathbb{R}^m$ as in the GLasso case, with possibly $p << m$.

Now, let the vector μ denote the dual variables associated to the constraint $\lambda \geq 0$. The Lagrangian for the problem (16) is then given by

$$L(\lambda, \mu) := \tfrac{1}{2}\log\det(\Sigma_y(\lambda)) + \tfrac{1}{2}y^\top \Sigma_y(\lambda)^{-1} y + \gamma \mathbf{1}^\top \lambda - \mu^\top \lambda \tag{19}$$

Using the fact that

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

the following result based on the KKT conditions for the problem (16) is obtained.

Proposition 5: The necessary conditions for λ to be a solution of (16) are

$$\begin{aligned} &\Sigma = \sigma^2 I_n + \textstyle\sum_{i=1}^p \lambda_i G^{(i)} G^{(i)\top} \\ &W\Sigma = I_n \\ &\text{tr}\left(G^{(i)\top} W G^{(i)}\right) - \|G^{(i)\top} W y\|_2^2 + 2\gamma - 2\mu_i = 0, \quad i = 1,\dots,p \\ &\mu_i \lambda_i = 0, \quad i = 1,\dots,p \\ &0 \leq \mu, \ \lambda \text{ and } 0 \preceq W, \Sigma \end{aligned}$$

C. Asymptotic behavior and BIC

It is well known [14] that the so-called BIC criterion for order estimation can be derived as the asymptotic approximation of an exact Bayes procedure which takes, as a prior on parameter space, a mixture of the form

$$p(\theta) = \sum_j \alpha_j p(\theta | j)$$

where j indexes the different model classes ($\theta \in \Theta_j$) and $p(\theta|j)$ is a probability measure for $\theta \in \Theta_j$. Under mild assumptions on $p(\theta|j)$ [14] the asymptotics (in the number of data) do not depend on the specific choice of α_j.

With respect to the priors (14) let us now define the $j-th$ model class as follows:

$$\theta^{(i)}|j \sim \mathcal{N}\left(0, w_{ji}\bar{\lambda} I_{k_i}\right), i = 1,..,p$$

where $w_j := [w_{j1},..,w_{jp}] \in \{0,1\}^p$, $j = 1,..,2^p$ is vector of indicators defining which blocks $\theta^{(i)}$ are allowed to be nonzero and which are zero, and denote:

$$p(\theta|j) := \prod_{i:w_{ji}=1} \left(2\pi\bar{\lambda}\right)^{-k_i/2} e^{-\frac{\left(\theta^{(i)}\right)^\top \theta^{(i)}}{2\bar{\lambda}}} \prod_{\ell:w_{j\ell}=0} \delta(\theta^{(\ell)}) \tag{20}$$

With this notation we can define now a prior on θ as follows:

$$p(\theta) := \frac{1}{2^p}\sum_{j=1}^{2^p} p(\theta|j) \tag{21}$$

Note that the prior model obtained from (20) and (21) is related, even though not equivalent, to the prior used for the Stochastic Search Variable Selection (SSVS) method in [9].

It now follows from the derivation in [14] that the exact Bayes procedure which selects $\hat{j}$ as

$$\hat{j} := \arg\max_{j=1,..,2^p} p(j|y) = \arg\max_{j=1,..,2^p} \int_{\mathbb{R}^m} p(y|\theta)p(\theta|j)\,d\theta$$

is asymptotically equivalent to minimizing

$$BIC(m_j) := \log(\hat{\sigma}_j^2) + \sum_{i=1}^{p}(w_{ji}p_i)\frac{\log n}{n} \tag{22}$$

where $\hat{\sigma}_j^2$ is the maximum likelihood estimator of the noise variance under model class $m_j := \{\theta : \theta^{(\ell)} = 0 \ \forall \ell : w_{j\ell} = 0\}$

Of course such an exhaustive search is infeasible for large p and greedy procedures such as that outlined in Section VI-B will have to be utilized. The arguments in this Section show that, indeed, asymptotically, the criterion (50) is equivalent to (22). This fact may be advantageous since

(22) will depend only on partial correlations which can be computed recursively as new candidate groups are introduced in the regression, thus greatly reducing the computational load.

D. Comparing GLasso and HGLasso

The two estimators discussed above do not derive from the same Bayesian model as in the previous case. In fact, consider the problem of integrating out λ from the joint density of θ and λ described by the model in Fig. 1(b). Then, the result is the product of multivariate Laplace densities. In particular, define $B^{(i)}(\cdot)$ as the modified Bessel function of the second kind and order $k_i/2-1$. Then, following also [15], we obtain

$$\int_{\lambda\in\mathbb{R}_+^p} p(\theta,\lambda)d\lambda = \frac{(2\gamma)^p}{(2\pi)^{m/2}}\prod_{i=1}^{p}(2\gamma))^{2-k_i/4}\frac{B^{(i)}(2\gamma\sqrt{\theta^{(i)\top}\theta^{(i)}})}{(\theta^{(i)\top}\theta^{(i)})^{k_i/4-2}} \tag{23}$$

whereas the prior density underlying the GLasso should be such that

$$p(\theta)\propto\exp(-\gamma_{GL}\sum_{i=1}^{p}\sqrt{\theta^{(i)\top}\theta^{(i)}}) \tag{24}$$

One can assess that, when $k_i>1$, for $\theta^{(i)}$ tending to zero the prior density on $\theta^{(i)}$ related to GLasso remains bounded, while the one related to the HGLasso, reported in (23), tends to ∞. As it will be also apparent in the numerical experiments section, this feature allows HGLasso to produce sparser solutions.

IV. MKL and HGLasso

A. MKL and its Bayesian interpretation

In order to introduce the Multiple Kernel Learning (MKL) approach, it is useful to start considering the following measurements model

$$y=f+v=\sum_{i=1}^{p}f^{(i)}+v \tag{25}$$

In the MKL framework, f in (25) represents the sampled version of a scalar function assumed to belong to a (generally infinite-dimensional) reproducing kernel Hilbert space (RKHS). For our purposes, we can consider a simplified scenario, where the domain of the functions in the RKHS is the finite set $[1,\dots,n]$. In this way, f represents the entire function and y is the noisy version of f sampled on all its domain. In addition, f is assumed to belong to the RKHS H_K whose kernel is defined by the matrix

$$K(\lambda)=\sum_{i=1}^{p}\lambda_i K^{(i)} \tag{26}$$

Then, each function $f^{(i)}$ is an element of the RKHS $H^{(i)}$ induced by the kernel $\lambda_i K^{(i)}$, with norm denoted by $\|\cdot\|_{(i)}$. According to the MKL approach, the estimates of the unknown functions $f^{(i)}$ are obtained *jointly* with those of the scale factors λ_i solving the following inequality constrained problem

$$\begin{aligned}(\{\hat f^{(i)}\},\hat\lambda) = \underset{\{f^{(i)}\},\lambda\in\mathbb{R}_+^p}{\arg\min}\ & \frac{(y-f)^\top(y-f)}{\sigma^2}+\sum_{i=1}^{p}\|f^{(i)}\|_{(i)}^2\\ \text{s.t.}\quad & \sum_{i=1}^{p}\lambda_i\le M\end{aligned} \tag{27}$$

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 [5], [8], the solution is

$$\hat f^{(i)}=\hat\lambda_i K^{(i)}\hat c,\qquad i=1,\dots,p \tag{28}$$

where

$$\begin{aligned}\{\hat c,\hat\lambda\} = \underset{c\in\mathbb{R}^n,\lambda\in\mathbb{R}_p^+}{\arg\min}\ & \frac{(y-K(\lambda)c)^\top(y-K(\lambda)c)}{\sigma^2}+c^\top K(\lambda)c\\ \text{s.t.}\quad & \sum_{i=1}^{p}\lambda_i\le M\end{aligned} \tag{29}$$

It can be shown that every local minimum of the above objective is also a global minimum, see [5] for details.

For our purposes, it is now useful to define ϕ as the Gaussian vector with independent components of unit variance such that

$$\theta_i=\sqrt{\lambda_i}\phi_i \tag{30}$$

We also factorize ϕ as done for θ, i.e.

$$\phi=[\phi^{(1)}\quad\phi^{(2)}\quad\dots\quad\phi^{(p)}]^\top \tag{31}$$

Then, the following connection with the Bayesian model in Fig. 1(b) holds.

Proposition 6: Consider the joint density of ϕ and λ conditional on y induced by the Bayesian network in Fig. 1(b). Let also $K^{(i)}=G^{(i)}G^{(i)\top}$. Then, there exists a value of γ such that the maximum a posteriori estimate of λ is the $\hat\lambda$ in (29). In addition, one has

$$\hat\lambda=\arg\min_{\lambda\in\mathbb{R}_+^p}\frac{y^\top(K(\lambda)+\sigma^2I_n)^{-1}y}{2}+\gamma\sum_{i=1}^{p}\lambda_i \tag{32}$$

Finally, the maximum a posteriori estimates of the blocks of ϕ are

$$\hat\phi^{(i)}=\sqrt{\lambda_i}G^{(i)\top}\hat c \tag{33}$$

where $\hat c$ is the same as in (29) and given by

$$\hat c(\hat\lambda)=(K(\hat\lambda)+\sigma^2I_n)^{-1}y \tag{34}$$

Proof: First, the expression for $\hat c(\hat\lambda)$ derives from (29) after simple computations that are omitted.

Now, given the Bayesian network in Fig. 1(b), apart from

constant factors we are not concerned with, the minus log of the joint density of y, ϕ, λ is given by

$$\frac{(y-\sqrt{\lambda_i}G\phi)^\top(y-\sqrt{\lambda_i}G\phi)}{2\sigma^2}+\frac{\sum_{i=1}^p \phi^{(i)\top}\phi^{(i)}}{2} +\gamma\sum_{i=1}^p \lambda_i \tag{35}$$

For known y and λ, ϕ is Gaussian so that the maximizer of the joint density with respect only to ϕ is given by

$$\begin{aligned}\phi^{(i)}(\lambda) &= \sqrt{\lambda_i}G^{(i)\top}(\sum_{i=1}^p \lambda_i G^{(i)}G^{(i)\top}+\sigma^2 I_n)^{-1}y \\ &= \sqrt{\lambda_i}G^{(i)\top}(K(\lambda)+\sigma^2 I_n)^{-1}y \end{aligned} \tag{36}$$

where the last equality exploits the relation $K(\lambda) = \sum_{i=1}^p \lambda_i G^{(i)}G^{(i)\top}$. Hence, from the above arguments, (33,34) are immediately obtained.
Finally, (32) is derived replacing the expression of $\phi^{(i)}(\lambda)$ obtained above in (35) and performing simple algebraic manipulations that exploit the following equality

$$I_n-\sum_{i=1}^p \lambda_i G^{(i)}G^{(i)\top}(K(\lambda)+\sigma^2 I_n)^{-1}=\sigma^2(K(\lambda)+\sigma^2 I_n)^{-1}$$

■

It is also of interest to give the KKT conditions for the objective (32). This is obtained in the next proposition.

Proposition 7: The necessary and sufficient conditions for λ to be a solution of (32) are

$$\Sigma = K(\lambda)+\sigma^2 I_n \tag{37}$$
$$W\Sigma = I_n \tag{38}$$
$$-\|G^{(i)\top}Wy\|_2^2+2\gamma-2\mu_i=0, \quad i=1,\dots,p \tag{39}$$
$$\mu_i\lambda_i=0, \quad i=1,\dots,p \tag{40}$$
$$0\le\mu,\ \lambda \text{ and } 0\preceq W,\Sigma \tag{41}$$

■

Finally, we notice that, starting from (33), a natural estimator for $\theta^{(i)}$ is

$$\hat{\theta}^{(i)}=\sqrt{\lambda_i}\hat{\phi}^{(i)} \tag{42}$$

We stress that the above expression does not provide the maximum a posteriori estimate of $\theta^{(i)}$. In fact, it is not difficult to see that the joint density of θ and λ, conditional on y, is not bounded above around the origin. Hence, this kind of MAP estimator would always return an estimate of θ equal to zero.

B. Comparing MKL and HGLasso

Proposition 6 points out how MKL derives from the same Bayesian model underlying HGLasso but the estimate of λ is now obtained maximizing a joint, in place of a marginal, density. The expression of the estimator (32) is interesting when compared with that reported in (16). In fact, recall that, under the assumptions stated in Proposition 6, $\Sigma_y(\lambda) = K(\lambda)+\sigma^2 I_n$. Hence, the two objectives in (32) and (16) are identical except that the term $\frac{1}{2}\log\det(\Sigma_y)$ is missing in the MKL objective (32). Notice also that this is the component which makes problem (16) non convex. On the other hand, this term allows HGLasso to favor sparser solutions than MKL since it makes the marginal density of λ more concentrated around zero.

V. Sparsity vs. Shrinking: Comparison via Optimality conditions

In this section we compare the sparsity conditions for HGLasso, MKL and GLasso; we show that HGLasso guarantees a more favorable tradeoff between sparsity and shrinkage, in the sense that it induces greater sparsity with the same shrinkage (or, equivalently, for a given level of sparsity it guarantees less shrinkage). In order to illustrate this behavior, we consider a specific example with 2 groups of dimension 1, i.e.

$$y = G^{(1)}\theta^{(1)}+G^{(2)}\theta^{(2)}+v \quad y\in\mathbb{R}^2, \theta_1\in\mathbb{R}, \theta_2\in\mathbb{R} \tag{43}$$

where $G^{(1)} = [1 \ \ \delta]^\top$, $G^{(2)} = [0 \ \ 1]^\top$, $v\sim\mathcal{N}(0,\sigma^2)$. We assume $\theta^{(1)} = 0$, $\theta^{(2)} = 1$; our aim is now to understand how the hyperparameter γ influences sparsity and estimates of $\theta^{(2)}$. In particular, we would like to understand which values of γ guarantee that $\hat{\theta}^{(1)} = 0$ and how the estimator $\hat{\theta}^{(2)}$ varies with γ. In order to do so we consider the KKT conditions obtained in Propositions 5 and 7.

For future recall that we have defined $K^{(i)} := G^{(i)}\left(G^{(i)}\right)^\top$; we find that necessary conditions for $\hat{\lambda}_1 = 0$ and $\hat{\lambda}_2$ be the hyperparameters estimators using the HGLasso estimator (for fixed γ) are:

$$\begin{aligned}&\gamma_{HGL}\ge\tfrac{1}{2}\mathrm{tr}\left(y^\top\Sigma^{-1}K^{(1)}\Sigma^{-1}y\right)-\tfrac{1}{2}\mathrm{tr}\left(K^{(1)}\Sigma^{-1}\right)\\ &\xi := \frac{-4\gamma_{HGL}\sigma^2-1+\sqrt{(1+4\gamma_{HGL}\sigma^2)^2-8\gamma_{HGL}\left(\sigma^2+2\gamma_{HGL}\sigma^4-\mathrm{tr}(K^{(2)}yy^\top)\right)}}{4\gamma_{HGL}}\\ &\hat{\lambda}_2^{HGL}=\max\{\xi,0\}\\ &\Sigma = K^{(2)}\hat{\lambda}_2+\sigma^2 I\end{aligned} \tag{44}$$

Similarly, the same conditions for MKL read as

$$\begin{aligned}\gamma_{MKL} &\ge \tfrac{1}{2}\mathrm{tr}\left(y^\top\Sigma^{-1}K^{(1)}\Sigma^{-1}y\right)\\ \xi &= \sqrt{\tfrac{1}{2\gamma_{MKL}}\mathrm{tr}\left(K^{(2)}yy^\top\right)}-\sigma^2\\ \hat{\lambda}_2^{MKL} &= \max\{\xi,0\}\\ \Sigma &= K^{(2)}\hat{\lambda}_2^{MKL}+\sigma^2 I\end{aligned} \tag{45}$$

The corresponding estimators for $\theta^{(1)}$ and $\theta^{(2)}$ are:

$$\begin{aligned}&\hat{\theta}^{(1)}_{HGL}=\hat{\theta}^{(1)}_{MKL}=0\\ &\hat{\theta}^{(2)}_{HGL}=\hat{\lambda}_2^{HGL}G^{(2)\top}\left(K^{(2)}\hat{\lambda}_2^{HGL}+\sigma^2 I\right)^{-1}y\\ &\hat{\theta}^{(2)}_{MKL}=\hat{\lambda}_2^{MKL}G^{(2)\top}\left(K^{(2)}\hat{\lambda}_2^{MKL}+\sigma^2 I\right)^{-1}y\end{aligned} \tag{46}$$

If we take, $\delta = 0$ in the definition of $G^{(2)}$, i.e. $G^{(2)} = [1 \ \ \delta]^\top = [1 \ \ 0]^\top$ and denote $y := [y_1 \ y_2^\top$, the expressions

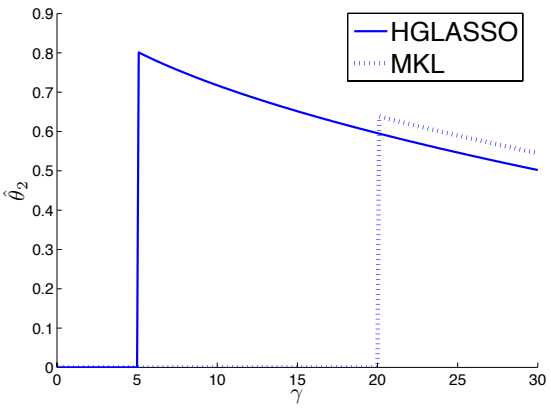

Fig. 2. Estimators $\hat{\theta}^{(2)}$ as a function of γ. The curves are plotted only for the values of γ which yield also $\hat{\theta}^{(1)} = 0$ (different for HGLasso ($\gamma_{HGL} > 5$) and MKL ($\gamma_{MKL} > 20$)).

simplify considerably, yielding for HGLasso:

$$\begin{array}{l} \gamma_{HGL} \geq \frac{1}{2\sigma^4}\left(y_1^2 - \sigma^2\right) \\ \hat{\lambda}_2^{HGL} = \max\{\xi, 0\} \\ \xi := \frac{-4\gamma_{HGL}\sigma^2 - 1 + \sqrt{(1+4\gamma_{HGL}\sigma^2)^2 - 8\gamma_{HGL}(\sigma^2 + 2\gamma_{HGL}\sigma^4 - y_2^2)}}{4\gamma_{HGL}} \end{array} \tag{47}$$

and

$$\begin{array}{rcl} \gamma_{MKL} & \geq & \frac{1}{2\sigma^4} y_1^2 \\ \xi & = & \sqrt{\frac{1}{2\gamma_{MKL}} y_2^2} - \sigma^2 \\ \hat{\lambda}_2^{MKL} & = & \max\{\xi, 0\} \end{array} \tag{48}$$

for MKL. It is clear that MKL requires a more stringent condition on γ (i.e. larger γ) in order to set $\hat{\lambda}_1^{MKL} = 0$ (and hence $\hat{\theta}_{MKL}^{(1)} = 0$). Of course having a larger γ tends to yield smaller $\hat{\lambda}_2$ and hence more shrinking on $\hat{\theta}^{(2)}$. This is illustrated in figure 2 where we report the estimators $\hat{\theta}_{HGL}^{(2)}$ (solid) and $\hat{\theta}_{MKL}^{(2)}$ (dotted) for $\sigma^2 = 0.005$, $\delta = 0.5$. The estimators are arbitrarily set to zero for the values of γ which do not yield $\hat{\theta}^{(1)} = 0$. In particular from (44) and (45) we obtain that GHLasso sets $\hat{\theta}_{HGL}^{(1)} = 0$ for $\gamma_{HGL} > 5$ while MKL sets $\hat{\theta}_{MKL}^{(1)} = 0$ for $\gamma_{MKL} > 20$. In addition it is clear that MKL tends to yield more shrinking on $\hat{\theta}_{MKL}^{(2)}$ (recall that $\theta^{(2)} = 1$).

Note that when the groups have dimension 1, as stated in Proposition 8, GLasso is equivalent to MKL with a proper rescaling of the regularization parameter, so that the comparison between HGLasso and MKL can be extended to GLasso.

Proposition 8: Assume that $k_1 = \ldots = k_p = 1$ for $i = 1, \ldots, p$ and $G = I_n$ so that GLasso reduces to Lasso. Then, the regularization paths of Lasso and MKL are the same.

Proof: The proof easily comes from the KKT conditions. In fact, assume that $\hat{\theta}_i$ is the i-th component of the solution obtained by Lasso and that it is different from zero. From Proposition 3 one obtains

$$-(y_i - \hat{\theta}_i) + \frac{\hat{\theta}_i}{\|\hat{\theta}_i\|}\gamma_L\sigma^2 = 0 \implies \hat{\theta}_i = y_i - \gamma_L\sigma^2$$

Instead, if $\hat{\theta}_i = 0$, one must have

$$\gamma_L \geq \frac{y_i}{\sigma^2}$$

Using also Proposition 7, one obtains that the MKL estimates of $\hat{\lambda}_i$ and $\hat{\theta}_i$ that are different from zero must satisfy

$$(\hat{\lambda}_i + \sigma^2)^2 = \frac{y_i^2}{2\gamma_{MKL}}, \qquad \hat{\theta}_i = \frac{\hat{\lambda}_i}{\hat{\lambda}_i + \sigma^2} y_i$$

that imply

$$\hat{\theta}_i = y_i - \sqrt{2\gamma_{MKL}}\sigma^2$$

On the other hand, the condition for $\hat{\theta}_i = 0$ becomes

$$\gamma_{MKL} \geq \frac{y_i^2}{2\sigma^4}$$

Hence, $\hat{\theta}$ returned by the two methods is the same provided that $\gamma_L = \sqrt{2\gamma_{MKL}}$. ■

VI. Implementing HGLasso

In this section we discuss the implementation of our HGLasso approach. This will also lead to the introduction of three different variants of this estimator.

A. Projected Quasi-Newton Method

The objective (16) is a differentiable function of λ with simple box constraints ($\lambda \geq 0$). Note that in order to compute the derivatives, the matrices $G^{(i)}G^{(i)^\top}$ need to be computed only once, and the inverse of the matrix $\Sigma_y(\lambda)$ needs to be computed once per iteration. Hence, an interesting feature of the problem is that the evaluation of the objective may be costly, as it depends on computing inverses of possibly large matrices and large matrix products. On the other hand, the dimension of the parameter vector λ can be small, and projection onto the feasible set is trivial.

We tried several methods, available from the Matlab package `minConf`, to optimize (16). The fastest method we implemented turned out to be a limited memory projected quasi-Newton algorithm detailed in [13]. It uses L-BFGS updates to build a diagonal plus low-rank quadratic approximation to the function, uses the Projected Quasi-Newton Method to minimize the quadratic approximation subject to the constraints present in the original problem, and uses a backtracking line search to generate new parameter vectors satisfying an Armijo-like sufficient decrease condition. The method is most effective since computing the projection onto the constraint set can be done much more efficiently than evaluating the function.

B. Bayesian Forward Selection

In this section we introduce a forward-selection type of procedure which will be useful to define a computationally efficient version of the HGLASSO estimator. In order to obtain an estimator of λ we consider the constraint $\kappa = \lambda_1 = \lambda_2 = \ldots = \lambda_p$ and treat κ as a deterministic hyperparameter whose knowledge makes Σ_y completely known. Therefore we set:

$$\hat{\kappa} := \arg\min_{\kappa \in \mathbb{R}_+} \frac{1}{2}\log\det(\Sigma_y) + \frac{1}{2} y^\top \Sigma_y^{-1} y \tag{49}$$

The forward-selection procedure is then designed as follows; let $I \subseteq \{1,2,..,p\}$ be the subset of currently selected groups and, considering now the Bayesian model in Fig. 1(b), define the marginal log posterior

$$L(I,\kappa,\gamma) := \log\left[p_\gamma(\tilde{\lambda}_I|y)\right] \tag{50}$$

where $\tilde{\lambda}_I := [\tilde{\lambda}_{I,1},...,\tilde{\lambda}_{I,p}]$ and $\tilde{\lambda}_{I,i} = \hat{\kappa}$ if $i \in I$ and $\tilde{\lambda}_{I,i} = 0$ otherwise.

Then do the following:

- set $\hat{\gamma} := \frac{1}{\hat{\kappa}}$
- initialize $I := \emptyset$
- repeat the following procedure:
  (a) for $j \in \{1,..,p\} \setminus I$, define $I_j^{'} := I \cup j$ and compute $L(I_j^{'};\hat{\kappa},\hat{\gamma})$.
  (b) select
  $$\bar{j} := \underset{j\in\{1,..,m\}\setminus I}{arg\,max}\ L(I_j^{'};\hat{\kappa},\hat{\gamma}) - L(I;\hat{\kappa},\hat{\gamma})$$
  (c) *if* $L(I_{\bar{j}}^{'};\hat{\kappa},\hat{\gamma}) - L(I;\hat{\kappa},\hat{\gamma}) > 0$
  set $I := I_{\bar{j}}^{'}$ and go back to (a)
  else
  finish.

Note that the set I contains the indexes of selected variables different from zero.

C. The three variants of HGLasso

The numerical procedures described above permit to introduce three different versions of HGLasso. They are listed below.

- **HGLa**: the optimization problem (49) is solved obtaining $\hat{\kappa}$. The regularization parameter γ is set to the inverse of $\hat{\kappa}$. Then, the forward-selection procedure described in the previous subsection is adopted to sparsify the solution, obtaining the estimate $\hat{\lambda}$ of the hyperparameter vector whose components are equal to 0 or $\hat{\kappa}$. Finally, the estimate $\hat{\theta}_{HGL}$ is obtained using (18).
- **HGLb**: the regularization parameter γ is set to the inverse of $\hat{\kappa}$ obtained by HGLa. Then, the optimization problem (16) is solved using the Projected Quasi-Newton method with starting point $\lambda_1 = \lambda_2 = \ldots = \lambda_p = \hat{\kappa}$ obtaining the estimate $\hat{\lambda}$ (notice that now all the components of $\hat{\lambda}$ different from zero may assume different values). Finally, $\hat{\theta}_{HGL}$ is obtained using (18).
- **HGLc**: this estimator performs the same operations of HGLb except that the components of λ set to zero by HGLa are kept at zero. Hence, problem (16) is in general optimized with respect to a restricted number of components of λ.

VII. Simulation results

We consider a Monte Carlo study of 500 runs where at any run a linear model of the form (10) is considered with $p = 10$ groups, each composed of $k_i = 5$ parameters, and $n = 100$. For each run, 5 of the groups $\theta^{(i)}$ are set to zero, one is always taken different from zero while each of the remaining 4 is set to zero with probability $p_i = 0.5$. The components of every block not set to zero are independent realizations from a uniform distribution on $[-a,a]$ where a is an independent realization (one for each block) from a uniform distribution on $[-100,100]$. The value of σ^2 is equal to the variance of the noiseless output divided by 25 and is assumed known. The columns of G are correlated, being defined at every run by

$$G_{i,j} = G_{i,j-1} + 0.2v_{i,j-1}, \quad i = 1,..,n, \quad j = 2,..,m$$
$$v_{i,j} \sim \mathcal{N}(0,1)$$

where $v_{i,j}$ are i.i.d. (as i and j vary) zero mean unit variance Gaussian and $G_{i,1}$ are i.i.d. zero mean unit variance Gaussian random variables. Note that correlated inputs renders the input selection problem more challenging.

We compare the following 5 estimators:

- **HGLa,HGLb,HGLc**: these are the three variants of our HGLasso procedure defined at the end of Section VI.
- **GLasso**: the regularization parameter is determined via cross validation, splitting the data set in two segments of the same size and testing a finite number of parameters from a pre-specified grid with 30 elements logarithmically distributed between $10^{-2}\hat{\gamma}$ and $10^{6}\hat{\gamma}$ where $\hat{\gamma}$ is the regularization parameter adopted by the three HLasso procedures. Finally, GLasso is reapplied to the full data set fixing the regularization parameter to its estimate.
- **MKL**: the regularization parameter is estimated using the same cross validation strategy adopted for GLasso.

The 5 estimators are compared computing the performance indexes listed below:

1) Percentage estimation error:
$$Err_1 = 100 \times \frac{\|\theta - \hat{\theta}\|}{\|\theta\|}\,\% \tag{51}$$
where $\hat{\theta}$ is the estimate of θ.
2) Absolute error on "zero" parameters:
$$Err_0 = \|\hat{\theta}^{(i)}\|, \quad i \quad \text{s.t.} \quad \|\theta^{(i)}\| = 0 \tag{52}$$
where $\hat{\theta}^{(i)}$ is the estimate of the i-th block of θ.
3) Percentage of the blocks equal to zero correctly set to zero by the estimator after the 500 runs.

Fig. 3 displays the boxplots of the 500 errors Err_1 and of Err_0. It is apparent that all of the three versions of the HGLasso outperform both GLasso and MKL. In addition, from the results reported in Table I one can see that the first and third versions of HGLasso obtain the remarkable performance of 99.5% of blocks correctly set to zero, while the second version obtains 72.5%. Instead, GLasso and MKL correctly set to zero 26.2% and 18.1% of the blocks, respectively. This result, which can appear surprising, is partially explained by the arguments in Section V; in a nutshell, MKL and GLasso need to trade sparsity for shrinking. The value of the regularization parameter γ needed to avoid oversmoothing is not large enough to induce "enough" sparsity. This drawback does not affect our new nonconvex estimators as described in Section V in a simplified scenario.

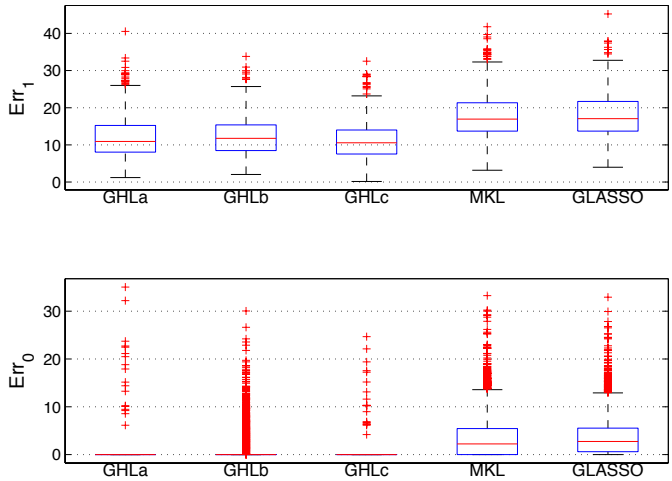

Fig. 3. Boxplot of the percentage errors in the reconstruction of θ (top) and of the absolute errors in the estimation of the null blocks of θ obtained by the 5 estimators after the 500 Monte Carlo runs.

HGLa	HGLb	HGLc	MKL	GLasso
99.5%	72.5%	99.5%	26.2%	18.1%

TABLE I

PERCENTAGE OF THE $\theta^{(i)}$ EQUAL TO ZERO CORRECTLY SET TO ZERO BY THE EMPLOYED ESTIMATORS.

VIII. CONCLUSIONS

We have presented a comparative study of three methods for sparse estimation, namely GLasso, MKL and the new HGLasso. It is shown that HGLasso and MKL derive from the same Bayesian model, yet in a different way; for GLasso, instead, this holds only for the case in which the groups have dimension 1. It is argued that the marginalization involved in HGLasso is advantageous, especially when the size of the groups is large. The tradeoffs between sparsity and shrinking are also studied in a simple example using the Karush Kuhn Tucker (KKT) conditions; our analysis suggests that HGLasso is able to achieve higher levels of sparsity without paying too much in terms of shrinking. This is indeed confirmed by the simulation experiments. Future work will include more efficient implementations of HGLasso, a thorough analysis of the optimality conditions and a more in depth study of the Bayesian forward selection used for initialization.

REFERENCES

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

[2] 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, 2010.

[3] 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, 2010.

[4] 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*.

[5] F. Dinuzzo. Kernel machines with two layers and multiple kernel learning. *arXiv:1001.2709*.

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

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

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

[9] E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[10] 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.

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

[12] Gianluigi Pillonetto, Francesco Dinuzzo, and Giuseppe De Nicolao. Bayesian online multitask learning of gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):193–205, 2010.

[13] Mark Schmidt, Ewout Van Den Berg, Michael P. Friedlander, and Kevin Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *Proc. of Conf. on Artificial Intelligence and Statistics*, pages 456–463, 2009.

[14] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[15] T.W. Lee T. Eltoft, T. Kim. On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13:300–303, 2006.

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

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

[18] H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

[19] 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.