Journal of Machine Learning Research 15 (2014) 217-252 Submitted 10/11; Revised 5/13; Published 1/13

Convex vs Non-Convex Estimators for Regression and Sparse Estimation: the Mean Squared Error Properties of ARD and GLasso

Aleksandr Aravkin SARAVKIN@US.IBM.COM
IBM T.J. Watson Research Center
1101 Kitchawan Rd, 10598
Yorktown Heights, NY, USA

James V. Burke BURKE@MATH.WASHINGTON.EDU
Department of Mathematics, Box 354350
University of Washington
Seattle, WA, 98195-4350 USA

Alessandro Chiuso CHIUSO@DEI.UNIPD.IT
Gianluigi Pillonetto GIAPI@DEI.UNIPD.IT
Department of Information Engineering
Via Gradenigo 6/A
University of Padova
Padova, Italy

Editor: Francis Bach

Abstract

We study a simple linear regression problem for grouped variables; we are interested in methods which jointly perform estimation and *variable selection*, that is, that automatically set to zero groups of variables in the regression vector. The Group Lasso (GLasso), a well known approach used to tackle this problem which is also a special case of Multiple Kernel Learning (MKL), boils down to solving convex optimization problems. On the other hand, a Bayesian approach commonly known as Sparse Bayesian Learning (SBL), a version of which is the well known Automatic Relevance Determination (ARD), lead to non-convex problems. In this paper we discuss the relation between ARD (and a penalized version which we call PARD) and Glasso, and study their asymptotic properties in terms of the Mean Squared Error in estimating the unknown parameter. The theoretical arguments developed here are independent of the correctness of the prior models and clarify the advantages of PARD over GLasso.

Keywords: Lasso, Group Lasso, Multiple Kernel Learning, Bayesian regularization, marginal likelihood

1. Introduction

We consider 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 setting we assume that θ is group (or block) sparse in the sense that 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 (Chiuso and Pillonetto, 2012, 2010b,a). In a dynamic network scenario, the *explanatory variables* are the past histories of different input signals, with the groups $\theta^{(i)}$ representing the impulse responses[1] describing 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.

Several approaches have been put forward 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.). Methods based on hierarchical Bayesian models have also been considered, including Automatic Relevance Determination (ARD) (Mackay, 1994), the Relevance Vector Machine (RVM) (Tipping, 2001), and the exponential hyperprior (Chiuso and Pillonetto, 2010b, 2012). The Bayesian approach considered by Chiuso and Pillonetto (2010b, 2012) is intimately related to that of Mackay (1994) and Tipping (2001); in fact the exponential hyperprior algorithm proposed by Chiuso and Pillonetto (2010b, 2012) is a penalized version of ARD (PARD) in which the prior on the groups $\theta^{(i)}$ is adapted to the structural properties of dynamical systems. A variational approach based on the golden standard spike and slab prior, also called two-groups prior (Efron, 2008), has been also recently proposed by Titsias and Lzaro-Gredilla (2011).

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).

Wipf and Nagarajan (2007) and Wipf et al. (2011) argue 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 $g(x_1, .., x_n) = \sum_{i=1}^{n} h(x_i)$). The analysis of 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

1. Impulse responses may, in principle, be infinite dimensional.

well in practice. Our experience is that the approach based on the marginal likelihood is particularly robust w.r.t. noise regardless of the "correctness" of the Bayesian prior.

Motivated by the strong performance of the exponential hyperprior approach introduced in the dynamic network identification scenario (Chiuso and Pillonetto, 2010b, 2012), we provide some new insights clarifying the above issues. The main contributions are as follows:

(i) We first provide some motivating examples which illustrate the superiority of PARD (and also of ARD) over GLasso both in terms of selection (i.e., detecting block of zeros in θ) as well as in estimation (i.e., reconstructing the non zero blocks).

(ii) Theoretical findings explaining the reasons underlying the superiority of PARD over GLasso are then provided. In particular, all the methods are compared in terms of optimality (KKT) conditions, and tradeoffs between sparsity and shrinkage are studied.

(iii) We then consider a non-Bayesian point of view, in which the estimation error is measured in terms of the Mean Squared Error, in the vein of Stein-estimators (James and Stein, 1961; Efron and Morris, 1973; Stein, 1981). The properties of Empirical Bayes estimators, which form the basis of the computational schemes, are studied in terms of their Mean Square Error properties; this is first established in the simplest case of orthogonal regressors and then extended to more general cases allowing for the regressors to be realizations from (possibly correlated) stochastic processes. This, of course, is of paramount importance for the system identification scenario studied by Chiuso and Pillonetto (2010b, 2012).

    Our analysis avoids assumptions on the correctness of the priors which define the stochastic model and clarifies why PARD is likely to provide sparser and more accurate estimates in comparison with GLasso (MKL). As a consequence of this analysis, our study clarifies the asymptotic properties of ARD.

Before we proceed with these results, we need to establish a common framework for these estimators (GLasso/MKL and PARD); this mostly uses results from the literature, which are recalled without proof in order to make the paper as self contained as possible.

The paper is organized as follows. In Section 2 we provide the problem statement while in Section 3 PARD and GLasso (MKL) are introduced in a Bayesian framework. Section 4 illustrates the advantages of PARD over GLasso using a simple example and two Monte Carlo studies. In Section 5 the Mean Squared Error properties of the Empirical Bayes estimators are studied, including their asymptotic behavior. Some conclusions end the paper while the Appendix gathers the proofs of the main results.

2. Problem Statement

We consider a linear model $y = G\theta + v$ where the explanatory factors G used to predict y are grouped (and non-overlapping). As such we partition θ into p sub-vectors $\theta^{(i)}$, $i = 1, \dots, p$, so that

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

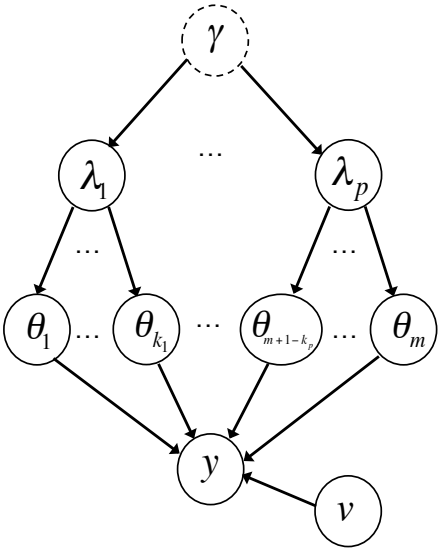

Figure 1: Bayesian networks describing the stochastic model for group sparse estimation

For $i = 1, \ldots, p$, 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 $G = [G^{(1)}, \ldots, G^{(p)}]$ to obtain the measurement model

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

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

Our problem is to estimate θ from y while also detecting the null blocks of $\theta^{(i)}$.

3. Estimators Considered

The purpose of this Section is to place the estimators we consider (GLasso/MKL and PARD) in a common framework that unifies the analysis. The content of the section is a collection of results taken from the literature which are stated without proof; the readers are referred to previous works for details which are not relevant to our paper's goal.

3.1 Bayesian Model for Sparse Estimation

Figure 1 provides a hierarchical representation of a probability density function useful for establishing a connection between the various estimators considered in this paper. In particular, in the Bayesian network of Figure 1, nodes and arrows are either dotted or solid depending on whether the quantities/relationships are deterministic or stochastic, respectively. Here, λ denotes a vector whose components $\{\lambda_i\}_{i=1}^{p}$ are independent and identically distributed exponential random variables with probability density

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

where γ is a positive scalar and

$$\chi(t) = \begin{cases} 1, & t \geq 0 \\ 0, & \text{elsewhere.} \end{cases}$$

In addition, let $\mathcal{N}(\mu, \Sigma)$ be the Gaussian density of mean μ and covariance Σ while, given a generic k, we use I_k to denote the $k \times k$ identity matrix. Then, conditional on λ, the blocks $\theta^{(i)}$ of the vector θ are all mutually independent and each block is zero-mean Gaussian with covariance $\lambda_i I_{k_i}$, $i = 1, .., p$, that is,

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

The measurement noise is also Gaussian, that is,

$$v \sim \mathcal{N}(0, \sigma^2 I_n).$$

3.2 Penalized ARD (PARD)

We introduce a sparse estimator, denoted by PARD in the sequel, cast in the framework of the Type II Bayesian estimators and consisting of a penalized version of ARD (Mackay, 1994; Tipping, 2001; Wipf and Nagarajan, 2007). It is derived from the Bayesian network depicted in Figure 1 as follows. First, the marginal density of λ is optimized, that is, we compute

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

Then, using an empirical Bayes approach, we obtain $\mathbb{E}[\theta|y, \lambda = \hat{\lambda}]$, that is, the minimum variance estimate of θ with λ taken as known and set to its estimate. The structure of the estimator is detailed in the following proposition (whose proof is straightforward and therefore omitted).

Proposition 1 (PARD) *Define*

$$\begin{aligned} \Sigma_y(\lambda) &:= G\Lambda G^\top + \sigma^2 I, & (2) \\ \Lambda &:= \mathit{blockdiag}(\{\lambda_i I_{k_i}\}). & (3) \end{aligned}$$

Then, the estimator $\hat{\theta}_{PA}$ of θ obtained from PARD is given by

$$\begin{aligned} \hat{\lambda} &= \arg\min_{\lambda \in \mathbb{R}_+^p} \frac{1}{2} \log\det(\Sigma_y(\lambda)) + \frac{1}{2} y^\top \Sigma_y^{-1}(\lambda) y + \gamma \sum_{i=1}^p \lambda_i, & (4) \\ \hat{\theta}_{PA} &= \hat{\Lambda} G^\top (\Sigma_y(\hat{\lambda}))^{-1} y. & (5) \end{aligned}$$

where $\hat{\Lambda}$ is defined as in (3) with each λ_i replaced by the i-th component of $\hat{\lambda}$ in (4).

■

One can see from (4) and (5) that the proposed estimator reduces to ARD if $\gamma = 0$.[2] In this case, the special notation $\hat{\theta}_A$ is used to denote the resulting estimator, that is,

$$\hat{\lambda} = \arg\min_{\lambda \in \mathbb{R}^p_+} \frac{1}{2}\log\det(\Sigma_y(\lambda)) + \frac{1}{2}y^\top \Sigma_y^{-1}(\lambda) y, \tag{6}$$

$$\hat{\theta}_A = \hat{\Lambda} G^\top (\Sigma_y(\hat{\lambda}))^{-1} y \tag{7}$$

where Σ_y is defined in (2), and $\hat{\Lambda}$ is defined as in (3) with each λ_i replaced by the i-th component of the $\hat{\lambda}$ in (6).

Observe that the objective in (4) is not convex in λ. Letting the vector μ denote the dual vector for the constraint $\lambda \geq 0$, the Lagrangian is given by

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

Using the fact that

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

we obtain the following KKT conditions for (4).

Proposition 2 (KKT for PARD) *The necessary conditions for λ to be a solution of (4) are*

$$\begin{aligned} &\Sigma_y = \sigma^2 I + \textstyle\sum_{i=1}^p \lambda_i G^{(i)} G^{(i)\top}, \\ &W \Sigma_y = I, \\ &tr\left(G^{(i)\top} W G^{(i)}\right) - \|G^{(i)\top} W y\|_2^2 + 2\gamma - 2\mu_i = 0, \quad i = 1, \ldots, p, \\ &\mu_i \lambda_i = 0, \quad i = 1, \ldots, p, \\ &0 \leq \mu, \ \lambda. \end{aligned} \tag{8}$$

3.3 Group Lasso (GLasso) and Multiple Kernel Learning (MKL)

A leading approach for the block sparsity problem is the Group Lasso (GLasso) (Yuan and Lin, 2006) which determines the estimate of θ as the solution of the following convex problem

$$\hat{\theta}_{GL} = \arg\min_{\theta \in \mathbb{R}^m} \frac{(y - G\theta)^\top (y - G\theta)}{2\sigma^2} + \gamma_{GL} \sum_{i=1}^p \|\theta^{(i)}\|, \tag{9}$$

where $\|\cdot\|$ denotes the classical Euclidean norm. Now, let ϕ be the Gaussian vector with independent components of unit variance such that

$$\theta_i = \sqrt{\lambda_i}\ \phi_i. \tag{10}$$

We partition ϕ conformally with θ, that is,

$$\phi = \left[\phi^{(1)\top} \quad \phi^{(2)\top} \quad \ldots \quad \phi^{(p)\top}\right]^\top. \tag{11}$$

2. Strictly speaking, what is called ARD in this paper corresponds to a group version of the original estimator discussed in Mackay (1994). A perfect correspondence is obtained when the dimension of each block is equal to one, that is, $k_i = 1\ \forall i$.

Then, interestingly, GLasso can be derived from the same Bayesian model in Figure 1 underlying PARD considering ϕ and λ as unknown variables and computing their maximum a posteriori (MAP) estimates. This is illustrated in the following proposition which is just a particular instance of the known relationship between regularization on kernel weights and block-norm based regularization. In particular, it establishes the known equivalence between GLasso and Multiple Kernel Learning (MKL) when linear models of the form (1) are considered, see the more general Theorem 1 of Tomioka and Suzuki (2011) for other details.

Proposition 3 (GLasso and its equivalence with MKL) *Consider the joint density of ϕ and λ conditional on y induced by the Bayesian network in Figure 1. Let $\hat{\lambda}$ and $\hat{\phi}$ denote, respectively, the maximum a posteriori estimates of λ and ϕ (obtained by optimizing their joint density). Then, for every γ_{GL} in (9) there exists γ such that the following equalities hold*

$$\hat{\lambda} = \arg\min_{\lambda \in \mathbb{R}^p_+} \frac{y^\top (\Sigma_y(\lambda))^{-1} y}{2} + \gamma \sum_{i=1}^{p} \lambda_i, \tag{12}$$

$$\begin{aligned} \hat{\phi}^{(i)} &= \sqrt{\hat{\lambda}_i} G^{(i)\top} (\Sigma_y(\hat{\lambda}))^{-1} y, \\ \hat{\theta}^{(i)}_{GL} &= \sqrt{\hat{\lambda}_i} \hat{\phi}^{(i)}. \end{aligned} \tag{13}$$

We warn the reader that MKL is more general than GLasso since it also embodies estimation in infinite dimensional models; yet in this paper we use interchangeably the nomenclature GLasso and MKL since they are equivalent for the considered model class.

Comparing Propositions 1 and 3, one can see that the sole difference between PARD and GLasso relies on the estimator for λ. In particular, notice that the objectives (12) and (4) differ only in the term $\frac{1}{2}\log\det(\Sigma_y)$ appearing in the PARD objective (4). This is the component that makes problem (4) non-convex but also the term that forces PARD to favor sparser solutions than GLasso (MKL), making the marginal density of λ more concentrated around zero. On the other hand, (12) is a convex optimization problem whose associated KKT conditions are reported in the following proposition.

Proposition 4 (KKT for GLasso and MKL) *The necessary and sufficient conditions for λ to be a solution of (12) are*

$$\begin{aligned} &K(\lambda) = \textstyle\sum_{i=1}^{p} \lambda_i G^{(i)} G^{(i)\top}, \\ &\Sigma_y = K(\lambda) + \sigma^2 I, \\ &W\Sigma_y = I, \\ &-\|G^{(i)\top} W y\|_2^2 + 2\gamma - 2\mu_i = 0, \quad i = 1, \ldots, p, \\ &\mu_i \lambda_i = 0, \quad i = 1, \ldots, p, \\ &0 \leq \mu,\ \lambda. \end{aligned} \tag{14}$$

■

Remark 5 (The LASSO case) *When all the blocks are one-dimensional, the estimator (9) reduces to Lasso and we denote the regularization parameter and the estimate by γ_L and $\hat{\theta}_L$, respectively. In this case, it is possible to obtain a derivation through marginalization. In fact, given the Bayesian network in Figure 1 with all the $k_i = 1$ and letting*

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

it follows from Section 2 in Park and Casella (2008) that $\hat{\theta} = \hat{\theta}_L$ provided that $\gamma_L = \sqrt{2\gamma}$.

4. Comparing PARD And GLasso (MKL): Motivating Examples

In this section, we present a sparsity vs. shrinkage example, and a Monte Carlo simulation to demonstrate advantages of the PARD estimator over GLasso.

4.1 Sparsity vs. Shrinkage: A Simple Experiment

It is well known that the ℓ_1 penalty in Lasso tends to induce an excessive shrinkage of "large" coefficients in order to obtain sparsity. Several variations have been proposed in the literature in order to overcome this problem, including the so called *Smoothly-Clipped-Absolute-Deviation* (SCAD) estimator (Fan and Li, 2001) and re-weighted versions of ℓ_1 like the *adaptive Lasso* (Zou, 2006). We now study the tradeoffs between sparsity and shrinking for PARD. By way of introduction to the more general analysis in the next section, we first compare the sparsity conditions for PARD and GLasso (or, equivalently, MKL) in a simple, yet instructive, two group example. In this example, it is straightforward to show that PARD 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).

Consider two groups of dimension 1, that is,

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

where $G^{(1)} = [1 \quad \delta]^\top$, $G^{(2)} = [0 \quad 1]^\top$, $v \sim \mathcal{N}(0, \sigma^2)$. Assume that the true parameter $\bar{\theta}$ satisfies: $\bar{\theta}^{(1)} = 0$, $\bar{\theta}^{(2)} = 1$. Our goal is to understand how the hyperparameter γ influences sparsity and the estimates of $\theta^{(1)}$ and $\theta^{(2)}$ using PARD and GLasso. In particular, we would like to determine which values of γ guarantee that $\hat{\theta}^{(1)} = 0$ and how the estimator $\hat{\theta}^{(2)}$ varies with γ. These questions can be answered by using the KKT conditions obtained in Propositions 2 and 4.

Let $y := [y_1\ y_2]^\top$. By (8), the necessary conditions for $\hat{\lambda}_1^{PA} = 0$ and $\hat{\lambda}_2^{PA} \geq 0$ to be the hyperparameter estimators for the PARD estimator (for fixed $\gamma = \gamma_{PA}$) are

$$\begin{aligned} &2\gamma_{PA} \geq \left[\tfrac{y_1}{\sigma^2} + \tfrac{\delta y_2}{\sigma^2 + \hat{\lambda}_2^{PA}}\right]^2 - \left[\tfrac{1}{\sigma^2} + \tfrac{\delta}{\sigma^2 + \hat{\lambda}_2^{PA}}\right] \quad \text{and} \\ &\hat{\lambda}_2^{PA} = \max\left\{\tfrac{-1+\sqrt{1+8\gamma_{PA}y_2^2}}{4\gamma_{PA}} - \sigma^2,\ 0\right\}. \end{aligned} \tag{15}$$

Similarly, by (14), the same conditions for $\hat{\lambda}_1^{GL} = 0$ and $\hat{\lambda}_2^{GL} \geq 0$ to be the estimators obtained using GLasso read as (for fixed $\gamma = \gamma_{GL}$):

$$\begin{aligned} 2\gamma_{GL} &\geq \left[\frac{y_1}{\sigma^2} + \frac{\delta y_2}{\sigma^2 + \hat{\lambda}_2^{GL}}\right]^2 \quad \text{and} \\ \hat{\lambda}_2^{GL} &= \max\left\{\frac{|y_2|}{\sqrt{2\gamma_{GL}}} - \sigma^2, 0\right\}. \end{aligned} \tag{16}$$

Note that it is always the case that the lower bound for γ_{GL} is strictly greater than the lower bound for γ_{PA} and that $\hat{\lambda}_2^{PA} \leq \hat{\lambda}_2^{GL}$ when $\gamma_{PA} = \gamma_{GL}$, where the inequality is strict whenever $\hat{\lambda}_2^{GL} > 0$. The corresponding estimators for $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ are

$$\hat{\theta}_{PA}^{(1)} = \hat{\theta}_{GL}^{(1)} = 0,$$

$$\hat{\theta}_{PA}^{(2)} = \frac{\hat{\lambda}_2^{PA} y_2}{\sigma^2 + \hat{\lambda}_2^{PA}} \qquad \text{and} \qquad \hat{\theta}_{GL}^{(2)} = \frac{\hat{\lambda}_2^{GL} y_2}{\sigma^2 + \hat{\lambda}_2^{GL}} \ .$$

Hence, $|\hat{\theta}_{PA}^{(2)}| < |\hat{\theta}_{GL}^{(2)}|$ whenever $y_2 \neq 0$ and $\hat{\lambda}_2^{GL} > 0$. However, it is clear that the lower bounds on γ in (15) and (16) indicate that γ_{GL} needs to be larger than γ_{PA} in order to set $\hat{\lambda}_1^{GL} = 0$ (and hence $\hat{\theta}_{GL}^{(1)} = 0$). 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 $\hat{\theta}_{PA}^{(2)}$ (solid) and $\hat{\theta}_{GL}^{(2)}$ (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 (15) and (16) we find that PARD sets $\hat{\theta}_{PA}^{(1)} = 0$ for $\gamma_{PA} > 5$ while GLasso sets $\hat{\theta}_{GL}^{(1)} = 0$ for $\gamma_{GL} > 20$. In addition, it is clear that MKL tends to yield greater shrinkage on $\hat{\theta}_{GL}^{(2)}$ (recall that $\bar{\theta}^{(2)} = 1$).

4.2 Monte Carlo Studies

We consider two Monte Carlo studies of 1000 runs. For each run a data set of size $n = 100$ is generated using the linear model (1) with $p = 10$ groups, each composed of $k_i = 4$ parameters. 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 groups $\theta^{(i)}$ are set to zero with probability 0.5. The components of every block $\theta^{(i)}$ not set to zero are independent realizations from a uniform distribution on $[-a_i, a_i]$ where a_i is an independent realization (one for each block) from a uniform distribution on $[0, 100]$. The value of σ^2 is the variance of the noiseless output divided by 25. The noise variance is estimated at each run as the sum of the residuals from the least squares estimate divided by $n - m$. The two experiments differ in the way the columns of G are generated at each run.

1. In the first experiment, the entries of G are independent realizations of zero mean unit variance Gaussian noise.

2. In the second experiment the columns of G are correlated, being defined at every run by

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

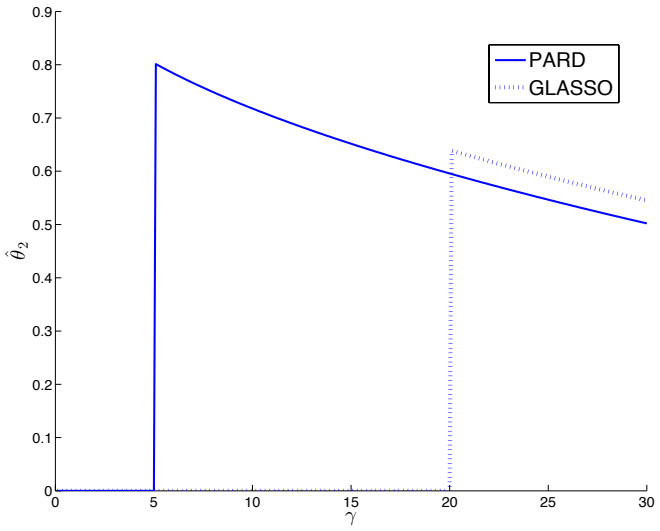

Figure 2: Estimators $\hat{\theta}^{(2)}$ as a function of γ. The curves are plotted only for the values of γ which yield also $\hat{\theta}^{(1)} = 0$ (different for PARD ($\gamma_{PA} > 5$) and MKL ($\gamma_{GL} > 20$)).

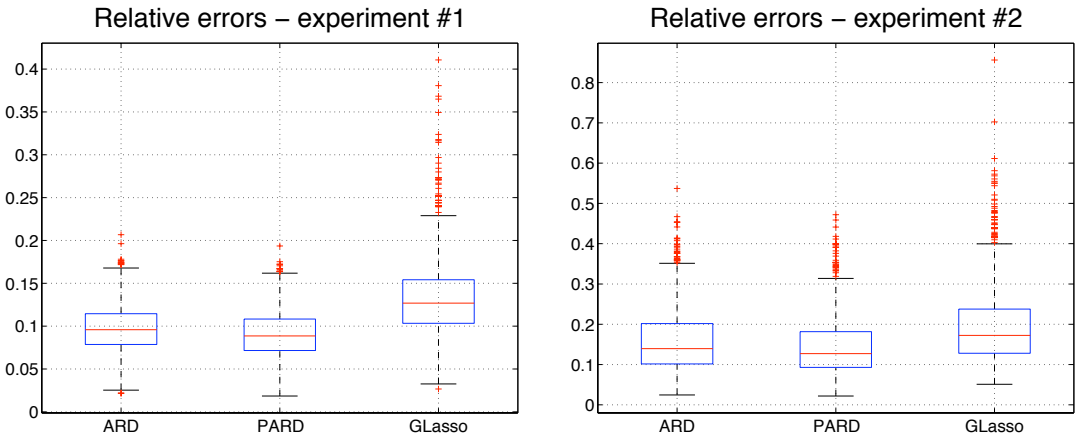

Figure 3: Boxplot of the relative errors in the reconstruction of θ obtained by the 2 non-convex estimators ARD and PARD and by the convex estimator GLasso (MKL) after the 1000 Monte Carlo runs in Experiment #1 (left panel) and #2 (right panel).

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 estimation problem more challenging.

Define $\hat{\kappa} \in \mathbb{R}$ as the optimizer of the ARD objective (6) under the constraint $\kappa = \lambda_1 = \ldots = \lambda_p$. Then, we define the following 3 estimators.

- **ARD**. The estimate θ_A is obtained by (6,7) using $\lambda_1 = \ldots = \lambda_p = \hat{\kappa}$ as starting point to solve (7) .
- **PARD**. The estimate θ_{PA} is obtained by (4,5) using cross validation to determine the regularization parameter γ. In particular, data are split into a training and validation set of equal size and the grid used by the cross validation procedure to select γ contains 30 elements logarithmically distributed between $10^{-2} \times \hat{\kappa}^{-1}$ and $10^{4} \times \hat{\kappa}^{-1}$. For each value of γ, (6) is solved using $\lambda_1 = \lambda_2 = \ldots = \lambda_p = \hat{\kappa}$ as starting point. Finally, θ_{PA} is obtained using the full data set fixing the regularization parameter to its estimate.
- **GLasso (MKL)**. The estimate θ_{GL} is obtained by (12-13) using the same cross validation strategy adopted by PARD to determine γ.

The three estimators described above are compared using the two performance indexes listed below:

1. Relative error: this is computed at each run as

$$\frac{\|\hat{\theta} - \theta\|}{\|\theta\|}$$

   where $\hat{\theta}$ is the estimator of θ.

2. Percentage of the blocks equal to zero correctly set to zero by the estimator after the 1000 runs.

The left and right panel of Figure 3 displays the boxplots of the 1000 relative errors obtained by the three estimators in the first and second experiment, respectively. The average relative error is also reported in Table 1. It is apparent that the performance of PARD and ARD is similar and that both of these non convex estimators outperform GLasso. Interestingy, in both of the experiments ARD and PARD return a reconstruction error smaller that that achieved by GLasso in more than 900 out of the 1000 runs.

In Table 2 we report the sparsity index. One can see that PARD exhibits the best performance, setting almost 75% of the blocks correctly to zero in the first and second experiment, respectively, while the performance of ARD is close to 67%. In contrast, GLasso (MKL) correctly set to zero no more than 40% of the blocks in each experiment.

Remark 6 (Projected Quasi-Newton Method) *We now comment on the optimization of (4). The same arguments reported below also apply to the objectives (6) and (12) which are just simplified versions of (4).*

	ARD	PARD	GLasso
Experiment #1	0.097	0.090	0.138
Experiment #2	0.151	0.144	0.197

Table 1: Comparison with MKL/GLasso (section 4.2). Average relative errors obtained by the three estimators.

	ARD	PARD	GLasso
Experiment #1	66.7%	74.5%	35.5%
Experiment #2	66.6%	74.6%	39.7%

Table 2: Comparison with MKL/GLasso (section 4.2). Percentage of the $\theta^{(i)}$ equal to zero correctly set to zero by the four estimators.

We notice that (4) is a differentiable function of λ. The computation of its derivative requires a one time evaluation of the matrices $G^{(i)}G^{(i)^\top}$, $i = 1, \ldots, p$. However, for each new value of λ, the inverse of the matrix $\Sigma_y(\lambda)$ also needs to be computed. Hence, the evaluation of the objective and its derivative may be costly since it requires computing the inverse of a possibly large matrix as well as 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 experimented with several methods available in the Matlab package `minConf` *to optimize (4). In these experiments, the fastest method was the limited memory projected quasi-Newton algorithm detailed in Schmidt et al. (2009). It uses L-BFGS updates to build a diagonal plus low-rank quadratic approximation to the function, and then uses the Projected Quasi-Newton Method to minimize a quadratic approximation subject to the original constraints to obtain a search direction. A backtracking line search is applied to this direction terminating at a step-size satisfying a Armijo-like sufficient decrease condition. The efficiency of the method derives in part from the simplicity of the projections onto the feasible region. We have also implemented the re-weighted method described by Wipf and Nagarajan (2007). In all the numerical experiments described above, we have assessed that it returns results virtually identical to those achieved by our method, with a similar computational effort. It is worth recalling that both the projected quasi-Newton method and the re-weighted approach guarantee convergence only to a stationary point of the objective.*

4.3 Concluding Remarks

The results in this section suggest that, when using GLasso, a suitable regularization parameter γ which does not induce oversmoothing (large bias) in $\hat{\theta}$ is not sufficiently large to induce "enough" sparsity. This drawback does not affect the nonconvex estimators. In addition, PARD and ARD seem to have the additional advantage of selecting the regularization parameters leading to more favorable Mean Squared Error (MSE) properties for the

reconstruction of the non zero blocks. The rest of the paper will be devoted to derivation of theoretical arguments supporting the intuition gained from these examples.

5. Mean Squared Error Properties of PARD and GLasso (MKL)

In this Section we evaluate the performance of an estimator $\hat{\theta}$ using its MSE, that is, the expected quadratic loss

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

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 Figure 1 with γ fixed.

5.1 Properties Using "Orthogonal" Regressors

We first 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 by Chiuso and Pillonetto (2010a,b, 2012) since it arises when one performs identification with i.i.d. white noises as inputs. We then consider the more general case of correlated regressors (see Section 5.2) and show that essentially the same result holds for a weighted version of the MSE.

In this section, it is convenient to introduce the following notation:

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

We now report an expression for the MSE of the Bayes estimators $\hat{\theta}(\lambda) := \mathbb{E}\left[\theta \,|\, y, \lambda\right]$ (the proof follows from standard calculations and is therefore omitted).

Proposition 7 *Consider the model* (1) *under the probabilistic model described in Figure 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*

$$\begin{aligned} MSE(\lambda) &= tr\left[\mathbb{E}_v\left[(\hat{\theta}(\lambda)-\theta)(\hat{\theta}(\lambda)-\theta)^{\top}\right]\right] \\ &= tr\left[\sigma^2\left(G^{\top}G+\sigma^2\Lambda^{-1}\right)^{-1}\left(G^{\top}G+\sigma^2\Lambda^{-1}\bar{\theta}\bar{\theta}^{\top}\Lambda^{-1}\right)\left(G^{\top}G+\sigma^2\Lambda^{-1}\right)^{-1}\right] \quad (17) \\ &= tr\left[\sigma^2\left(\Lambda G^{\top}G+\sigma^2\right)^{-1}\left(\Lambda G^{\top}G\Lambda+\sigma^2\bar{\theta}\bar{\theta}^{\top}\right)\left(G^{\top}G\Lambda+\sigma^2\right)^{-1}\right]. \end{aligned}$$

We can now minimize the expression for $MSE(\lambda)$ given in (17) 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 8 *Assume that $G^\top G = nI$ in Proposition 7. Then MSE(λ) is globally minimized by choosing*

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

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 estimator 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 9 *Consider model* (1) *under the probabilistic model described in Figure 1(b), and assume that $G^\top G = nI$. Then the estimator of λ_i obtained by maximizing the marginal posterior $\mathbf{p}(\lambda|y)$,*

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

is given by

$$\hat{\lambda}_i(\gamma) = \max\left(0, \frac{1}{4\gamma}\left[\sqrt{k_i^2 + 8\gamma\|\hat{\theta}_{LS}^{(i)}\|^2} - \left(k_i + \frac{4\sigma^2\gamma}{n}\right)\right]\right), \tag{18}$$

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 (18) *yields:*

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

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

$$\mathbb{P}[\hat{\lambda}_i(\gamma) = 0 \,|\, \theta = \bar{\theta}] = \mathbb{P}\left[\chi^2\left(k_i, \|\bar{\theta}^{(i)}\|^2 \frac{n}{\sigma^2}\right) \le \left(k_i + 2\gamma\frac{\sigma^2}{n}\right)\right], \tag{19}$$

where $\chi^2(d, \mu)$ denotes a noncentral χ^2 random variable with d degrees of freedom and noncentrality parameter μ.

Note that the expression of $\hat{\lambda}_i(\gamma)$ in Proposition 9 has the form of a "saturation". In particular, for $\gamma = 0$, we have

$$\hat{\lambda}_i(0) = \max(0, \hat{\lambda}_i^*), \quad \text{where} \quad \hat{\lambda}_i^* := \frac{\|\hat{\theta}_{LS}^{(i)}\|^2}{k_i} - \frac{\sigma^2}{n}. \tag{20}$$

The following proposition shows that the "unsaturated" estimator $\hat{\lambda}_i^*$ is an 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 10 *Under the assumption $G^\top G = nI$, the estimator of $\hat{\lambda}^* := \{\lambda_1^*, .., \lambda_p^*\}$ in* (20) *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, that is:*

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

and

$$\lim_{n\to\infty} \hat{\lambda}_i^* \overset{m.s.}{=} \lambda_i^{opt} \quad \lim_{n\to\infty} \hat{\lambda}_i(0) \overset{m.s.}{=} \lambda_i^{opt} \tag{21}$$

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

Remark 11 *Note that if $\bar{\theta}^{(i)} = 0$, the optimal value λ_i^{opt} is zero. Hence* (21) *shows that asymptotically $\hat{\lambda}_i(0)$ converges to zero. However, in this case, it is easy to see from* (19) *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 $2\frac{\sigma^2}{n}\gamma_n \to \infty$, with γ_n growing faster than n. This is in line with the well known requirements for Lasso to be model selection consistent. In fact, recalling remark 5, 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 PARD has desirable properties with respect to the MSE of the estimators. One wonders whether the same favorable properties are inherited by MKL or, equivalently, by GLasso. The next proposition shows that this is not the case. In fact, for $\bar{\theta}^{(i)} \neq 0$, MKL does not 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 (24)) is much smaller than that obtained using PARD (see Equation (19)); this is illustrated in Figure 4 (top). Also note that, as illustrated in Figure 4 (bottom), when the true $\bar{\theta}$ is equal to zero, MKL tends to give much larger values of $\hat{\lambda}$ than those given by PARD. This results in larger values of $\|\hat{\theta}\|$ (see Figure 4).

Proposition 12 *Consider model* (1) *under the probabilistic model described in Figure 1(b), and assume $G^\top G = nI$. Then the estimator of λ_i obtained by maximizing the joint posterior $\mathbf{p}(\lambda, \phi|y)$ (see Equations* (10) *and* (11)*),*

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

is given by

$$\hat{\lambda}_i(\gamma) = \max\left(0, \frac{\|\hat{\theta}_{LS}^{(i)}\|}{\sqrt{2\gamma}} - \frac{\sigma^2}{n}\right), \tag{22}$$

where

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

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

$$\lim_{n\to\infty} \hat{\lambda}_i(\gamma) \stackrel{m.s.}{=} \frac{\|\bar{\theta}^{(i)}\|}{\sqrt{2\gamma}} \ . \tag{23}$$

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

$$\mathbb{P}_\theta[\hat{\lambda}_i(\gamma) = 0 \,|\, \theta = \bar{\theta}] = \mathbb{P}\left[\chi^2\left(k_i, \|\bar{\theta}^{(i)}\|^2 \frac{n}{\sigma^2}\right) \leq 2\gamma \frac{\sigma^2}{n}\right] \ . \tag{24}$$

Note that the limit of the MKL estimators $\hat{\lambda}_i(\gamma)$ as $n \to \infty$ depends on γ. Therefore, using MKL (GLasso), 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 $\gamma \to \frac{k_i^2}{2\|\bar{\theta}^{(i)}\|^2}$, which is a circular requirement.

5.2 Asymptotic Properties Using General Regressors

In this subsection, we replace the deterministic matrix G with $G_n(\omega)$, where $G_n(\omega)$ represents an $n \times m$ matrix 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[3] realizations from a zero-mean random vector with positive definite covariance Ψ.

As in the previous part of this section, λ and θ are seen as parameters, and the true value of θ is $\bar{\theta}$. Hence, all the randomness present in the next formulas comes only from G_n and the measurement noise.

We make the following (mild) assumptions on G_n. Recalling model (1), assume that $G^{\top}G/n$ is bounded and bounded away from zero in probability, so that there exist constants $\infty > c_{max} \geq c_{min} > 0$ with

$$\lim_{n\to\infty} P[c_{min} I \leq G^{\top}G/n \leq c_{max} I] = 1 \ , \tag{25}$$

so, as n increases, the probability that a particular realization G satisfies

$$c_{min} I \leq G^{\top}G/n \leq c_{max} I \tag{26}$$

increases to 1.

In the following lemma, whose proof is in the Appendix, we introduce a change of variables that is key for our understanding of the asymptotic properties of PARD under these more general regressors.

Lemma 13 *Fix $i \in \{1, \ldots, p\}$ and consider the decomposition*

$$\begin{array}{rcl} y & = & G^{(i)}\theta^{(i)} + \sum_{j=1, j\neq i}^{p} G^{(j)}\theta^{(j)} + v \\ & = & G^{(i)}\theta^{(i)} + \bar{v} \end{array} \tag{27}$$

3. The independence assumption can be removed and replaced by mixing conditions.

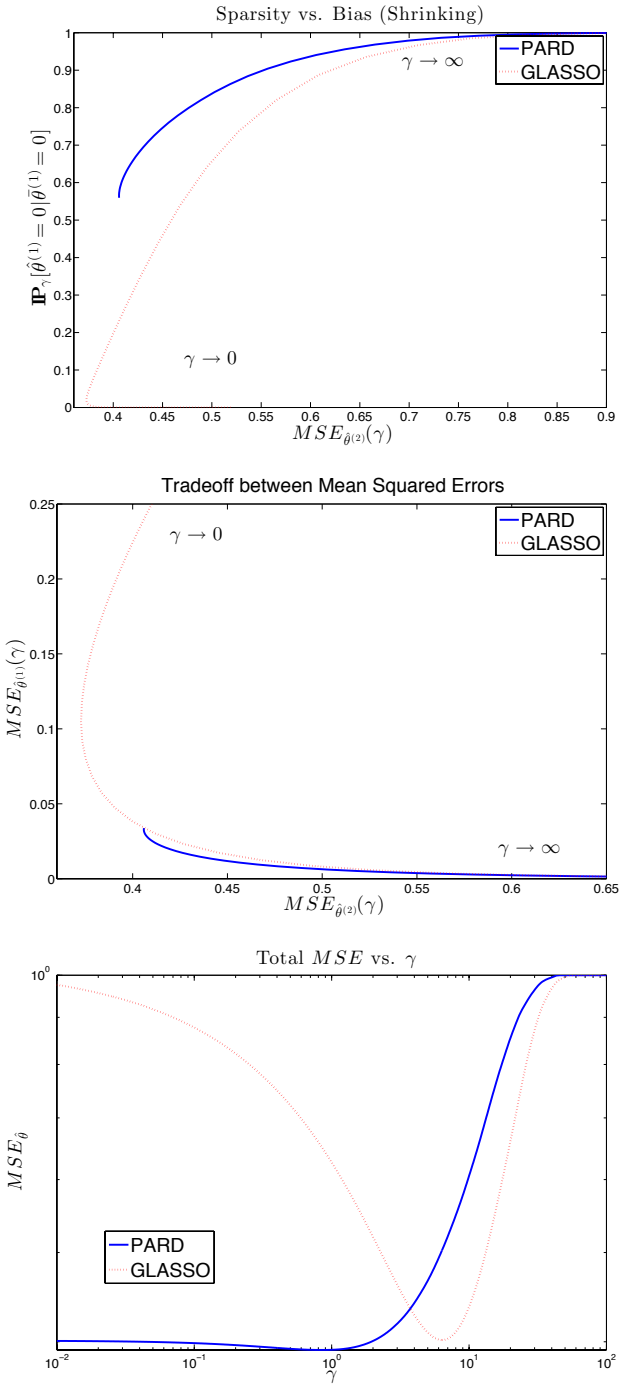

Figure 4: In this example we have two blocks ($p = 2$) of dimension $k_1 = k_2 = 10$ with $\bar{\theta}^{(1)} = 0$ and all the components of the true $\bar{\theta}^{(2)} \in \mathbb{R}^{10}$ set to one. The matrix G is the identity, so that the output dimension ($y \in \mathbb{R}^n$) is $n = 20$; the noise variance equals 0.5. Top: probability of setting $\hat{\theta}^{(1)}$ to zero vs Mean Squared Error in $\hat{\theta}^{(2)}$. Center: Mean Squared Error in $\hat{\theta}^{(1)}$ vs. Mean Squared Error in $\hat{\theta}^{(2)}$; both curves are parametrized in $\gamma \in [0, +\infty)$. Bottom: Total Mean Squared Error (on $\hat{\theta}$) as a function of γ.

of the linear measurement model (1) *and assume* (26) *holds. Define*

$$\Sigma_{\bar{v}}^{(i)} := \sum_{j=1, j\neq i}^{p} G^{(j)} \left(G^{(j)}\right)^{\top} \lambda_j + \sigma^2 I.$$

Consider now the singular value decomposition

$$\frac{\Sigma_{\bar{v}}^{(i)^{-1/2}} G^{(i)}}{\sqrt{n}} = U_n^{(i)} D_n^{(i)} \left(V_n^{(i)}\right)^{\top}, \tag{28}$$

where each $D_n^{(i)} = diag(d_{k,n}^{(i)})$ *is* $k_i \times k_i$ *diagonal matrix. Then* (27) *can be transformed into the equivalent linear model*

$$z_n^{(i)} \quad = \quad D_n^{(i)} \beta_n^{(i)} + \epsilon_n^{(i)}, \tag{29}$$

where

$$\begin{gathered} z_n^{(i)} := \left(U_n^{(i)}\right)^{\top} \frac{\Sigma_{\bar{v}}^{(i)^{-1/2}} y}{\sqrt{n}} = (z_{k,n}^{(i)}), \qquad \beta_n^{(i)} := \left(V_n^{(i)}\right)^{\top} \theta^{(i)} = (\beta_{k,n}^{(i)}), \\ \epsilon_n^{(i)} := \left(U_n^{(i)}\right)^{\top} \frac{\Sigma_{\bar{v}}^{(i)^{-1/2}} \bar{v}}{\sqrt{n}} = (\epsilon_{k,n}^{(i)}), \end{gathered} \tag{30}$$

and $D_n^{(i)}$ *is uniformly (in* n*) bounded and bounded away from zero.*

Below, the dependence of $\Sigma_y(\lambda)$ on G_n, and hence on n, is omitted to simplify the notation. Furthermore, $\longrightarrow_p$ denotes convergence in probability.

Theorem 14 *For known* γ *and conditional on* $\theta = \bar{\theta}$*, define*

$$\hat{\lambda}^n = \arg \min_{\lambda \in \mathcal{C} \bigcap \mathbb{R}_+^p} \frac{1}{2} \log \det(\Sigma_y(\lambda)) + \frac{1}{2} y^{\top} \Sigma_y^{-1}(\lambda) y + \gamma \sum_{i=1}^{p} \lambda_i, \tag{31}$$

where $\mathcal{C}$ *is any* p*-dimensional ball with radius strictly larger than* $\max_i \frac{\|\bar{\theta}^{(i)}\|^2}{k_i}$*. Suppose that the hypotheses of Lemma 13 hold. Consider the estimator (31) along its* $i-th$ *component* λ_i *that, in view of* (29)*, is given by:*

$$\hat{\lambda}_i^n = \arg \min_{\lambda \in \mathbb{R}_+} \frac{1}{2} \sum_{k=1}^{k_i} \left[\frac{\eta_{k,n}^2 + v_{k,n}}{\lambda + w_{k,n}} + \log(\lambda + w_{k,n}) \right] + \gamma \lambda \, , \tag{32}$$

where $\eta_{k,n} := \beta_{k,n}^{(i)}$*,* $w_{k,n} := 1/(n(d_{k,n}^{(i)})^2)$ *and* $v_{k,n} = 2 \frac{\epsilon_{k,n}^{(i)}}{d_{k,n}^{(i)}} \beta_{k,n}^{(i)} + \left(\frac{\epsilon_{k,n}^{(i)}}{d_{k,n}^{(i)}} \right)^2$*. Let*

$$\bar{\lambda}_i^{\gamma} := \frac{-k_i + \sqrt{k_i^2 + 8\gamma \|\bar{\theta}^{(i)}\|^2}}{4\gamma} \, , \quad \bar{\lambda}_i = \frac{\|\bar{\theta}^{(i)}\|^2}{k_i} \, .$$

We have the following results:

1. $\bar{\lambda}_i^{\gamma} \leq \bar{\lambda}_i$ *for all* $\gamma > 0$*, and* $\lim_{\gamma \to 0^+} \bar{\lambda}_i^{\gamma} = \bar{\lambda}_i$ *.*

2. *If* $\|\bar{\theta}^{(i)}\| > 0$ *and* $\gamma > 0$, *we have* $\hat{\lambda}_i^n \longrightarrow_p \bar{\lambda}_i^\gamma$.

3. *If* $\|\bar{\theta}^{(i)}\| > 0$ *and* $\gamma = 0$, *we have* $\hat{\lambda}_i^n \longrightarrow_p \bar{\lambda}_i$.

4. *if* $\bar{\theta}^{(i)} = 0$, *we have* $\hat{\lambda}_i^\gamma \longrightarrow_p 0$ *for any value* $\gamma \geq 0$.

We now show that, when $\gamma = 0$, the above result relates to the problem of minimizing the MSE of the i-th block with respect to λ_i, with all the other components of λ coming from $\hat{\lambda}^n$. For any index i, we define

$$I_1^{(i)} := \left\{ j \,:\, j \neq i \text{ and } \bar{\theta}^{(j)} \neq 0 \right\}, \quad I_0^{(i)} := \left\{ j \,:\, j \neq i \text{ and } \bar{\theta}^{(j)} = 0 \right\}. \tag{33}$$

If $\hat{\theta}_n^{(i)}(\lambda)$ denotes the i-th component of the PARD estimate of θ defined in (5), our aim is to optimize the objective

$$MSE_n(\lambda_i) := \operatorname{tr}\left[\mathbb{E}_v\left[(\hat{\theta}_n^{(i)}(\lambda) - \bar{\theta}^{(i)})(\hat{\theta}_n^{(i)}(\lambda) - \bar{\theta}^{(i)})^\top\right]\right] \quad \text{with} \quad \lambda_j = \bar{\lambda}_j^n \quad \text{for} \quad j \neq i$$

where is $\bar{\lambda}_j^n$ is any sequence satisfying condition

$$\lim_{n\to\infty} f_n = +\infty \quad \text{where} \quad f_n := \min_{j \in I_1^{(i)}} n\lambda_j^n, \tag{34}$$

(condition (34) appears again in the Appendix as (47)). Note that, in particular, $\bar{\lambda}_j^n = \hat{\lambda}_j^n$ in (31) satisfy (34) in probability.

Lemma 13 shows that we can consider the transformed linear model associated with the i-th block, that is,

$$z_{k,n}^{(i)} \;=\; d_{k,n}^{(i)}\beta_{k,n}^{(i)} + \epsilon_{k,n}^{(i)}, \quad k = 1, \ldots, k_i, \tag{35}$$

where all the three variables on the RHS depend on $\bar{\lambda}_j^n$ for $j \neq i$. In particular, the vector $\beta_n^{(i)}$ consists of an orthonormal transformation of $\theta^{(i)}$ while the $d_{k,n}^{(i)}$ are all bounded below in probability. In addition, by letting

$$\mathbb{E}_v\left[\epsilon_{k,n}^{(i)}\right] = m_{k,n}, \quad \mathbb{E}_v\left[(\epsilon_{k,n}^{(i)} - m_{k,n})^2\right] = \sigma_{k,n}^2,$$

we also know from Lemma 20 (see Equations (48) and (49)) that, provided $\bar{\lambda}_j^n$ ($j \neq i$) satisfy condition (34), both $m_{k,n}$ and $\sigma_{k,n}^2$ tend to zero (in probability) as n goes to ∞. Then, after simple computations, one finds that the MSE relative to $\beta_n^{(i)}$ is the following random variable whose statistics depend on n:

$$MSE_n(\lambda_i) = \sum_{k=1}^{k_i} \frac{\beta_{k,n}^2 + n\lambda_i^2 d_{k,n}^2(m_{k,n}^2 + \sigma_{k,n}^2) - 2\lambda_i d_{k,n} m_{k,n}\beta_{k,n}}{(1 + n\lambda_i d_{k,n}^2)^2} \quad \text{with } \lambda_j = \bar{\lambda}_j^n \text{ for } j \neq i.$$

Above, except for λ_i, the dependence on the block number i was omitted to improve readability.

Now, let $\breve{\lambda}_i^n$ denote the minimizer of the following weighted version of the $MSE_n(\lambda_i)$:

$$\breve{\lambda}_i^n = \arg\min_{\lambda \in \mathbb{R}_+} \sum_{k=1}^{k_i} d_{k,n}^4 \frac{\beta_{k,n}^2 + n\lambda_i^2 d_{k,n}^2(m_{k,n}^2 + \sigma_{k,n}^2) - 2\lambda_i d_{k,n} m_{k,n}\beta_{k,n}}{(1 + n\lambda_i d_{k,n}^2)^2}.$$

Then, the following result holds.

Proposition 15 *For $\gamma = 0$ and conditional on $\theta = \bar{\theta}$, the following convergences in probability hold*

$$\lim_{n\to\infty} \breve{\lambda}_i^n = \frac{\|\bar{\theta}^{(i)}\|^2}{k_i} = \lim_{n\to\infty} \hat{\lambda}_i^n \ , \quad i = 1, 2, \ldots, p. \tag{36}$$

The proof follows arguments similar to those used in last part of the proof of Theorem 14, see also proof of Theorem 6 in Aravkin et al. (2012), and is therefore omitted.

We can summarize the two main findings reported in this subsection as follows. As the number of measurements go to infinity:

1. regardless of the value of γ (provided γ does not depend on n; in such a case suitable conditions on the rate are necessary, see also Remark 11), the proposed estimator will correctly set to zero only those λ_i associated with null blocks;

2. when $\gamma = 0$, results 2 and 3 of Theorem 14 provide the asymptotic properties of ARD, showing that the estimate of λ_i will converge to the energy of the i-th block (divided by its dimension).

   This same value also represents the asymptotic minimizer of a weighted version of the MSE relative to the i-th block. In particular, the weights change with n, as they are defined by the singular values $d_{k,n}^{(i)}$ (raised at fourth power) that depend on the trajectories of the other components of λ (see (28)). This roughly corresponds to giving more emphasis to components of θ which excite directions in the output space where the signal to noise ratio is high; this indicates some connection with reduced rank regression where one only seeks to approximate the most important (relative to noise level) directions in output space.

Remark 16 (Consistency of $\hat{\theta}_{PA}$) *It is a simple check to show that, under the assumptions of Theorem 14, the empirical Bayes estimator $\hat{\theta}_{PA}(\hat{\lambda}_n)$ in (5) is a consistent estimator of $\bar{\theta}$. Indeed, Theorem 14 shows much more than this, implying that for $\gamma = 0$, $\hat{\theta}_{PA}(\hat{\lambda}_n)$ possesses some desirable asymptotic properties in terms on Mean Squared Error, see also Remark 17.*

5.3 Marginal Likelihood and Weighted MSE: Perturbation Analysis

We now provide some additional insights on point 2 above, investigating why the weights $d_{k,n}^4$ may lead to an effective strategy for hyperparameter estimation.
For our purposes, just to simplify the notation, let us consider the case of a single m-dimensional block. In this way, λ becomes a scalar and the noise $\epsilon_{k,n}$ in (35) is zero-mean of variance $1/n$.

Under the stated assumptions, the MSE weighted by $d_{k,n}^{\alpha}$, with α an integer, becomes

$$\sum_{k=1}^{m} d_{k,n}^{\alpha} \frac{n^{-1}\beta_{k,n}^2 + \lambda^2 d_{k,n}^2}{(n^{-1} + \lambda d_{k,n}^2)^2},$$

whose partial derivative with respect to λ, apart from the scale factor $2/n$, is

$$F_{\alpha}(\lambda) = \sum_{k=1}^{m} d_{k,n}^{\alpha+2} \frac{\lambda - \beta_{k,n}^2}{(n^{-1} + \lambda d_{k,n}^2)^3}.$$

Let $\beta_k = \lim_{n\to\infty} \beta_{k,n}$ and $d_k = \lim_{n\to\infty} d_{k,n}$.[4] When n tends to infinity, arguments similar to those introduced in the last part of the proof of Theorem 14 show that, in probability, the zero of F_α becomes

$$\breve{\lambda}(\alpha) = \frac{\sum_{k=1}^m d_k^{\alpha-4}\beta_k^2}{\sum_{k=1}^m d_k^{\alpha-4}}.$$

Notice that the formula above is a generalization of the first equality in (36) that was obtained by setting $\alpha = 4$. However, for practical purposes, the above expressions are not useful since the true values of $\beta_{k,n}$ and β_k depend on the unknown $\bar{\theta}$. One can then consider a noisy version of F_α obtained by replacing $\beta_{k,n}$ with its least squares estimate, that is,

$$\tilde{F}_\alpha(\lambda) = \sum_{k=1}^m d_{k,n}^{\alpha+2} \frac{\lambda - \left(\beta_{k,n} + \frac{v_{k,n}}{\sqrt{n}d_{k,n}}\right)^2}{(n^{-1} + \lambda d_{k,n}^2)^3},$$

where the random variable $v_{k,n}$ is of unit variance. For large n, considering small additive perturbations around the model $z_k = d_k\beta_k$, it is easy to show that the minimizer tends to the following perturbed version of $\breve{\lambda}$:

$$\breve{\lambda}(\alpha) + 2\frac{\sum_{k=1}^m d_k^{\alpha-5}\beta_k v_{k,n}}{\sqrt{n}\sum_{k=1}^m d_k^{\alpha-4}}. \tag{37}$$

We must now choose the value of α that should enter the above formula. This is far from trivial since the optimal value (minimizing MSE) depends on the unknown β_k. On one hand, it would seem advantageous to have α close to zero. In fact, $\alpha = 0$ relates $\breve{\lambda}$ to the minimization of the MSE on θ while $\alpha = 2$ minimizes the MSE on the output prediction, see the discussion in Section 4 of Aravkin et al. (2012). On the other hand, a larger value for α could help in controlling the additive perturbation term in (37) possibly reducing its sensitivity to small values of d_k. For instance, the choice $\alpha = 0$ introduces in the numerator of (37) the term β_k/d_k^5. This can destabilize the convergence towards $\breve{\lambda}$, leading to poor estimates of the regularization parameters, as, for example, described via simulation studies in Section 5 of Aravkin et al. (2012). In this regard, the choice $\alpha = 4$ appears interesting: it sets $\breve{\lambda}$ to the energy of the block divided by m, removing the dependence of the denominator in (37) on d_k. In particular, it reduces (37) to

$$\frac{\|\beta\|^2}{m} + \frac{2}{m}\sum_{k=1}^m \frac{\beta_k v_{k,n}}{\sqrt{n}d_k} = \sum_{k=1}^m \frac{\beta_k^2}{m}\left(1 + 2\frac{v_{k,n}}{\beta_k\sqrt{n}d_k}\right). \tag{38}$$

It is thus apparent that $\alpha = 4$ makes the perturbation on $\frac{\beta_k^2}{m}$ dependent on $\frac{v_{k,n}}{\beta_k\sqrt{n}d_k}$, that is, on the relative reconstruction error on β_k. This appears a reasonable choice to account for the ill-conditioning possibly affecting least-squares.

Interestingly, for large n, this same philosophy is followed by the marginal likelihood procedure for hyperparameter estimation up to first-order approximations. In fact, under

4. We are assuming that both of the limits exist. This holds under conditions ensuring that the SVD decomposition leading to (35) is unique, for example, see the discussion in Section 4 of Bauer (2005), and combining the convergence of sample covariances with a perturbation result for the Singular Value Decomposition of symmetric matrices (such as Theorem 1 in Bauer, 2005, see also Chatelin, 1983).

the stated assumptions, apart from constants, the expression for twice the negative log of the marginal likelihood is

$$\sum_{k=1}^{m} \log(n^{-1} + \lambda d_{k,n}^2) + \frac{z_{k,n}^2}{n^{-1} + \lambda d_{k,n}^2},$$

whose partial derivative w.r.t. λ is

$$\sum_{k=1}^{m} \frac{\lambda d_{k,n}^4 + n^{-1} d_{k,n}^2 - z_{k,n}^2 d_{k,n}^2}{(n^{-1} + \lambda d_{k,n}^2)^2}.$$

As before, we consider small perturbations around $z_k = d_k \beta_k$ to find that a critical point occurs at

$$\sum_{k=1}^{m} \frac{\beta_k^2}{m} \left(1 + 2\frac{v_{k,n}}{\beta_k \sqrt{n} d_k}\right),$$

which is exactly the same minimizer reported in (38).

5.4 Concluding Remarks and Connections to Subsection 4.2

We can now give an interpretation of the results depicted in Figure 3 in view of the theoretical analyses developed in this section.

When the regressors are orthogonal, which corresponds, asymptotically, to the case of white noise defining the entries of G, the results in subsection 5.1 (e.g., Corollary 10) show that ARD has a clear advantage over GLasso (MKL) in terms of MSE. This explains the outcomes from the first numerical experiment of Section 4.2 which are depicted in the left panel of Figure 3.

For the case of general regressors, subsection 5.2 provides insights regarding the properties of ARD, including its consistency. Ideally, a regularized estimator should adopt those hyperparameters contained in λ that minimize the MSE objective, but this objective depends on $\bar{\theta}$, which is what we aim to estimate. One could then consider a noisy version of the MSE function, for example, obtained replacing $\bar{\theta}$ with its least squares estimate. The problem is that this new objective can be very sensitive to the noise, leading to poor regularizers as, for example, described by simulation studies in Section 5 of Aravkin et al. (2012). On an asymptotic basis, ARD circumvents this problem using particular weights which introduce a bias in the MSE objective but make it more stable, that is, less sensitive to noise. This results in a form of regularization introduced through hyperparameter estimation. We believe that this peculiarity is key to understanding not only the results in Figure 3 but also the success of ARD in several applications documented in the literature.

Remark 17 [PARD: penalized version of ARD] *Note that, when one considers sparsity inducing performance, the use of a penalized version of ARD, for example, given by PARD, clearly may help in setting more blocks to zero, see Figure 4 (top). In comparison with GLasso, the important point here is that the non convex nature of PARD permits sparsity promotion without adopting too large a value of γ. This makes PARD a slightly perturbed version of ARD. Hence, PARD is able to induce more sparsity than ARD while*

maintaining similar performance in the reconstruction of the non null blocks. This is illustrated by the Monte Carlo results in Section 4.2. To better understand the role of γ, consider the orthogonal case discussed in Section 5.1, for sake of clarity. Recall the observation in Remark 11 that model selection consistency requires $\gamma = \gamma_n$. It is easy to show that the oracle property (Zou, 2006) holds provided $\frac{\gamma_n}{n} \to \infty$ and $\frac{\gamma_n}{n^2} \to 0$. However, large γ's tend to introduce excessive shrinkage, for example, see Figure 4 (center). It is well known (Leeb and Pötscher, 2005) that shrinkage estimators that possess the oracle property have unbounded normalized risk (normalized Mean Squared Error for quadratic loss), meaning that they are certainly not optimal in terms of Mean Squared Error. To summarize, the asymptotic properties suggest that to obtain the oracle properties γ_n should go to infinity at a suitable rate with n while γ should be set equal to zero to optimize the mean squared error. However, for finite data length n, the optimal mean squared error properties as a function of γ are found for a finite but nonzero γ. This fact, also illustrated in Figure 4 (bottom), is not in contrast w.r.t. Corollary 10: γ may induce a useful bias in the marginal likelihood estimator of the λ_i which can reduce the variance. This also explains the experimental results showing that PARD performs slightly better than ARD.

6. Conclusions

We have presented a comparative study of some methods for sparse estimation: GLasso (equivalently, MKL), ARD and its penalized version PARD, which is cast in the framework of the Type II Bayesian estimators. They derive from the same Bayesian model, yet in a different way. The peculiarities of PARD can be summarized as follows:

- in comparison with GLasso, PARD derives from a marginalized joint density with the resulting estimator involving optimization of a non-convex objective;
- the non-convex nature allows PARD to achieve higher levels of sparsity than GLasso without introducing too much regularization in the estimation process, thus providing a better tradeoff between sparsity and shrinking.
- the MSE analysis reported in this paper reveals the superior performance of PARD in the reconstruction of the parameter groups different from zero. Remarkably, our analysis elucidates this issue showing the robustness of the empirical Bayes procedure, based on marginal likelihood optimization, independently of the correctness of the priors which define the stochastic model underlying PARD. As a consequence of our analysis, the asymptotic properties of ARD have also been illuminated.

Many variations of PARD are possible, adopting different prior models for λ. In this paper, the exponential prior is used to compare different estimators that can be derived from the same Bayesian model underlying GLasso. In this way, it is shown that the same stochastic framework can give rise to an estimator derived from a posterior marginalization that has significant advantages over another estimator derived from posterior optimization.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement no 257462 HYCON2

Network of excellence, by the MIUR FIRB project RBFR12M3AC - Learning meets time: a new computational approach to learning in dynamic systems.

Appendix A. Proofs

In this Appendix, we present proofs of the main results in the paper.

A.1 Proof of Proposition 9

Under the simplifying assumption $G^\top G = nI$, one can use

$$\Sigma_y(\lambda)^{-1} = \sigma^{-2}\left[I - G(\sigma^2\Lambda^{-1} + G^TG)^{-1}G^\top\right]$$

which derives from the matrix inversion lemma to obtain

$$G^{(i)\top}\Sigma_y(\lambda)^{-1} = \frac{1}{n\lambda_i + \sigma^2}G^{(i)\top},$$

and so

$$\mathrm{tr}\left(G^{(i)\top}\Sigma_y^{-1}G^{(i)}\right) = \frac{nk_i}{n\lambda_i+\sigma^2} \quad \text{and} \quad \|G^{(i)\top}\Sigma_y^{-1}y\|_2^2 = \left(\frac{n}{n\lambda_i+\sigma^2}\right)^2\|\hat\theta_{LS}^{(i)}\|^2 .$$

Inserting these expressions into (8) with $\mu_i = 0$ yields a quadratic equation in λ_i which always has two real solutions. One is always negative while the other, given by

$$\frac{1}{4\gamma}\left[\sqrt{k_i^2 + 8\gamma\|\hat\theta_{LS}^{(i)}\|^2} - \left(k_i + \frac{4\sigma^2\gamma}{n}\right)\right]$$

is non-negative provided

$$\frac{\|\hat\theta_{LS}^{(i)}\|^2}{k_i} \geq \frac{\sigma^2}{n}\left[1 + \frac{2\gamma\sigma^2}{nk_i}\right] . \tag{39}$$

This concludes the proof of (18). The limiting behavior for $\gamma \to 0$ can be easily verified, yielding

$$\hat\lambda_i(0) = \max\left(0, \frac{\|\hat\theta_{LS}^{(i)}\|^2}{k_i} - \frac{\sigma^2}{n}\right) \quad i = 1, .., p.$$

Also note that $\hat\theta_{LS}^{(i)} = \frac{1}{n}\left(G^{(i)}\right)^\top y$ and $\left(G^{(i)}\right)^\top G^{(i)} = nI_{k_i}$ while $\left(G^{(i)}\right)^\top G^{(j)} = 0$, $\forall j \neq i$. This implies that $\hat\theta_{LS}^{(i)} \sim \mathcal{N}(\bar\theta^{(i)}, \frac{\sigma^2}{n}I_{k_i})$. Therefore

$$\|\hat\theta_{LS}^{(i)}\|^2\frac{n}{\sigma^2} \sim \chi^2(d,\mu) \quad d = k_i, \quad \mu = \|\bar\theta^{(i)}\|^2\frac{n}{\sigma^2}.$$

This, together with (39), proves also (19).

A.2 Proof of Proposition 10

In the proof of Proposition 9 it was shown that $\|\hat{\theta}_{LS}^{(i)}\|^2 \frac{n}{\sigma^2}$ follows a noncentral χ^2 distribution with k_i degrees of freedom and noncentrality parameter $\|\bar{\theta}^{(i)}\|^2 \frac{n}{\sigma^2}$. Hence, it is a simple calculation to show that

$$\mathbb{E}[\hat{\lambda}_i^* \,|\, \theta = \bar{\theta}] = \frac{\|\bar{\theta}^{(i)}\|^2}{k_i} \quad \mathrm{Var}[\hat{\lambda}_i^* \,|\, \theta = \bar{\theta}] = \frac{2\sigma^4}{k_i n^2} + \frac{4\|\bar{\theta}^{(i)}\|^2 \sigma^2}{k_i^2 n} \ .$$

By Corollary 8, the first of these equations shows that $\mathbb{E}[\hat{\lambda}_i^* \,|\, \theta = \bar{\theta}] = \lambda_i^{opt}$. In addition, since $\mathrm{Var}\{\hat{\lambda}_i^*\}$ goes to zero as $n \to \infty$, $\hat{\lambda}_i^*$ converges in mean square (and hence in probability) to λ_i^{opt}.

As for the analysis of $\hat{\lambda}_i(0)$, observe that

$$\mathbb{E}[\hat{\lambda}_i(0) \,|\, \theta = \bar{\theta}] = \mathbb{E}[\hat{\lambda}_i^* \,|\, \theta = \bar{\theta}] - \int_0^{k_i \frac{\sigma^2}{n}} \left(\frac{\|\hat{\theta}_{LS}^{(i)}\|^2}{k_i} - \frac{\sigma^2}{n} \right) dP(\|\hat{\theta}_{LS}^{(i)}\|^2 \,|\, \theta = \bar{\theta})$$

where $dP(\|\hat{\theta}_{LS}^{(i)}\|^2 \,|\, \theta = \bar{\theta})$ is the measure induced by $\|\hat{\theta}_{LS}^{(i)}\|^2$. The second term in this expression can be bounded by

$$-\int_0^{k_i \frac{\sigma^2}{n}} \left(\frac{\|\hat{\theta}_{LS}^{(i)}\|^2}{k_i} - \frac{\sigma^2}{n} \right) dP(\|\hat{\theta}_{LS}^{(i)}\|^2 \,|\, \theta = \bar{\theta}) \leq \frac{\sigma^2}{n} \int_0^{k_i \frac{\sigma^2}{n}} dP(\|\hat{\theta}_{LS}^{(i)}\|^2 \,|\, \theta = \bar{\theta}),$$

where the last term on the right hand side goes to zero as $n \to \infty$. This proves that $\hat{\lambda}_i(0)$ is asymptotically unbiased. As for consistency, it is sufficient to observe that $\mathrm{Var}[\hat{\lambda}_i(0) \,|\, \theta = \bar{\theta}] \leq \mathrm{Var}[\hat{\lambda}_i^* \,|\, \theta = \bar{\theta}]$ since "saturation" reduces variance. Consequently, $\hat{\lambda}_i(0)$ converges in mean square to its mean, which asymptotically is λ_i^{opt} as shown above. This concludes the proof.

A.3 Proof of Proposition 12

Following the same arguments as in the proof of Proposition 9, under the assumption $G^\top G = nI$ we have that

$$\|G^{(i)\top} \Sigma_y^{-1} y\|_2^2 = \left(\frac{n}{n\lambda_i + \sigma^2} \right)^2 \|\hat{\theta}_{LS}^{(i)}\|^2.$$

Inserting this expression into (14) with $\mu_i = 0$, one obtains a quadratic equation in λ_i which has always two real solutions. One is always negative while the other, given by

$$\frac{\|\hat{\theta}_{LS}^{(i)}\|}{\sqrt{2\gamma}} - \frac{\sigma^2}{n}.$$

is non-negative provided

$$\|\hat{\theta}_{LS}^{(i)}\|^2 \geq \frac{2\gamma\sigma^4}{n^2} \ . \tag{40}$$

This concludes the proof of (22).

The limiting behavior for $n \to \infty$ in Equation (23) is easily verified with arguments similar to those in the proof of Proposition 10. As in the proof of Proposition 9, $\|\hat{\theta}_{LS}^{(i)}\|^2 \frac{n}{\sigma^2}$ follows a noncentral $\chi^2(d, \mu)$ distribution with $d = k_i$ and $\mu = \|\bar{\theta}^{(i)}\|^2 \frac{n}{\sigma^2}$, so that from (40) the probability of setting $\hat{\lambda}_i(\gamma)$ to zero is as given in (24).

A.4 Proof of Lemma 13:

Let us consider the Singular Value Decomposition (SVD)

$$\frac{\sum_{j=1, j\neq i}^{p} G^{(j)} \left(G^{(j)}\right)^\top \lambda_j}{n} = PSP^\top, \tag{41}$$

where, by the assumption (26), using $\frac{\sum_{j=1,j\neq i}^{p} G^{(j)}\left(G^{(j)}\right)^\top \lambda_j}{n} \geq \frac{\sum_{j=1,j\neq i,\lambda_j\neq 0}^{p} G^{(j)}\left(G^{(j)}\right)^\top}{n} \min\{\lambda_j, j : \lambda_j \neq 0\}$ and Lemma 19 the minimum singular value $\sigma_{min}(S)$ of S in (41) satisfies

$$\sigma_{min}(S) \geq c_{min} \min\{\lambda_j, j : \lambda_j \neq 0\}. \tag{42}$$

Then the SVD of $\Sigma_{\bar{v}}^{(i)} = \sum_{j=1, j\neq i}^{p} G^{(j)} \left(G^{(j)}\right)^\top \lambda_j + \sigma^2 I$ satisfies

$$\Sigma_{\bar{v}}^{(i)\,-1} = \begin{bmatrix} P & P_\perp \end{bmatrix} \begin{bmatrix} (nS + \sigma^2)^{-1} & 0 \\ 0 & \sigma^{-2} I \end{bmatrix} \begin{bmatrix} P^\top \\ P_\perp^\top \end{bmatrix}$$

so that $\left\| \Sigma_{\bar{v}}^{(i)\,-1} \right\| = \sigma^{-2}$.

Note now that

$$D_n^{(i)} = \left(U_n^{(i)}\right)^\top \frac{\Sigma_{\bar{v}}^{(i)\,-1/2} G^{(i)}}{\sqrt{n}} V_n^{(i)}$$

and therefore, using Lemma 19,

$$\|D_n^{(i)}\| \leq \left\| \Sigma_{\bar{v}}^{(i)\,-1/2} \right\| \sqrt{c_{max}} = \sigma^{-1} \sqrt{c_{max}}$$

proving that $D_n^{(i)}$ is bounded. In addition, again using Lemma 19, condition (26) implies that $\forall a, b$ (of suitable dimensions) s.t. $\|a\| = \|b\| = 1$, $a^\top \frac{P_\perp^\top G^{(i)}}{\sqrt{n}} b \geq k$, $k = \sqrt{1 - \cos^2(\theta_{min})} \geq \frac{c_{min}}{c_{max}} > 0$. This, using (28), guarantees that

$$\begin{aligned} D_n^{(i)} &= \left(U_n^{(i)}\right)^\top \frac{\Sigma_{\bar{v}}^{(i)\,-1/2} G^{(i)}}{\sqrt{n}} V_n^{(i)} = \left(U_n^{(i)}\right)^\top \left(P(nS + \sigma^2)^{-1/2} P^\top + P_\perp \sigma^{-1} P_\perp^\top\right) \frac{G^{(i)}}{\sqrt{n}} \\ &\geq \left(U_n^{(i)}\right)^\top \left(P_\perp \sigma^{-1} P_\perp^\top\right) \frac{G^{(i)}}{\sqrt{n}} \\ &\geq k\sigma^{-1} I \end{aligned}$$

and therefore $D_n^{(i)}$ is bounded away from zero. It is then a matter of simple calculations to show that with the definitions (30) then (27) can be rewritten in the equivalent form (29). □

A.5 Preliminary Lemmas

This part of the Appendix contains some preliminary lemmas which will be used in the proof of Theorem 14. This first focuses on the estimator (31). We show that when the hypotheses of Lemma 13 hold, the estimate (31) satisfies the key assumptions of the forthcoming Lemma 20. We begin with a detailed study of the objective (4).

Let

$$I_1 := \left\{ j \,:\, \bar{\theta}^{(j)} \neq 0 \right\}, \quad I_0 := \left\{ j \,:\, \bar{\theta}^{(j)} = 0 \right\}.$$

Note that these are analogous to $I_1^{(i)}$ and $I_0^{(i)}$ defined in (33), but do not depend on any specific index i. We now state the following lemma.

Lemma 18 *Writing the objective in* (4) *in expanded form gives*

$$g_n(\lambda) = \log \sigma^2 + \underbrace{\frac{1}{2n} \log\det(\sigma^{-2}\Sigma_y(\lambda))}_{S_1} + \underbrace{\frac{1}{2n} \sum_{j \in I_1} \frac{\|\hat{\theta}^{(j)}(\lambda)\|^2}{k_j \lambda_j}}_{S_2} + \underbrace{\frac{1}{2n} \sum_{j \in I_0} \frac{\|\hat{\theta}^{(j)}(\lambda)\|^2}{k_j \lambda_j}}_{S_3}$$
$$+ \underbrace{\frac{1}{n}\gamma\|\lambda\|_1}_{S_4} + \underbrace{\frac{1}{2n\sigma^2}\|y - \sum_j G^j \hat{\theta}^{(j)}(\lambda)\|^2}_{S_5},$$

where $\hat{\theta}(\lambda) = \Lambda G^T \Sigma_y^{-1} y$ *(see* (5)*),* k_j *is the size of the* j*th block, and dependence on* n *has been suppressed. For any minimizing sequence* λ^n*, we have the following results:*

1. $\hat{\theta}_n \to_p \bar{\theta}$.
2. $S_1, S_2, S_3, S_4 \to_p 0$.
3. $S_5 \to_p \frac{1}{2}$.
4. $n\lambda_j^n \to_p \infty$ *for all* $j \in I_1$.

Proof First, note that $0 \leq S_i$ for $i \in \{1, 2, 3, 4\}$. Next,

$$\begin{aligned} S_5 &= \frac{1}{2n\sigma^2}\|y - \sum_j G^j \bar{\theta}^{(j)}(\lambda) + \sum_j G^j \left(\bar{\theta}^{(j)}(\lambda) - \hat{\theta}^{(j)}(\lambda) \right) \|^2 \\ &= \frac{1}{2n\sigma^2}\|\nu + \sum_j G^j \left(\bar{\theta}^{(j)}(\lambda) - \hat{\theta}^{(j)}(\lambda) \right) \|^2 \\ &= \frac{1}{2n\sigma^2}\|\nu\|^2 + \frac{1}{2n\sigma^2}\nu^T \sum_j G^j \left(\bar{\theta}^{(j)}(\lambda) - \hat{\theta}^{(j)}(\lambda) \right) + \frac{1}{2n\sigma^2}\| \sum_j G^j \left(\bar{\theta}^{(j)}(\lambda) - \hat{\theta}^{(j)}(\lambda) \right) \|^2. \end{aligned} \tag{43}$$

The first term converges in probability to $\frac{1}{2}$. Since ν is independent of all G^j, the middle term converges in probability to 0. The third term is the bias incurred unless $\hat{\theta} = \bar{\theta}$. These facts imply that, $\forall \epsilon > 0$,

$$\lim_{n \to \infty} P\left[S_5(\lambda(n)) > \frac{1}{2} - \epsilon \right] = 1\,. \tag{44}$$

Next, consider the particular sequence $\bar{\lambda}_j^n = \frac{\|\bar{\theta}_j\|^2}{k_j}$. For this sequence, it is immediately clear that $S_i \to_p 0$ for $i \in \{2,3,4\}$. To show $S_1 \to_p 0$, note that $\sum \lambda_i G_i G_i^T \leq \max\{\lambda_i\} \sum G_i G_i^T$, and that the nonzero eigenvalues of GG^T are the same as those of $G^T G$. Therefore, we have

$$S_1 \leq \frac{1}{2n} \sum_{i=1}^{m} \log(1 + n\sigma^{-2} \max\{\lambda\} c_{max}) = O_P\left(\frac{\log(n)}{n}\right) \to_p 0 \, .$$

Finally $S_5 \to_p \frac{1}{2}$ by (43), so in fact, $\forall \epsilon > 0$,

$$\lim_{n\to\infty} P\left[\left| g_n(\bar{\lambda}(n)) - \frac{1}{2} - \log(\sigma^2) \right| < \epsilon \right] = 1 \, . \tag{45}$$

Since (45) holds for the deterministic sequence $\bar{\lambda}_n$, any minimizing sequence $\hat{\lambda}_n$ must satisfy, $\forall \epsilon > 0$,

$$\lim_{n\to\infty} P\left[g_n(\hat{\lambda}(n)) < \frac{1}{2} + \log(\sigma^2) + \epsilon \right] = 1$$

which, together with (44), implies (45)

Claims $1, 2, 3$ follow immediately. To prove claim 4, suppose that for a particular minimizing sequence $\check{\lambda}(n)$, we have $n\check{\lambda}_j^n \not\to_p \infty$ for $j \in I_1$. We can therefore find a subsequence where $n\check{\lambda}_j^n \leq K$, and since $S_2(\check{\lambda}(n)) \to_p 0$, we must have $\|\hat{\theta}^{(j)}(\check{\lambda})\| \to_p 0$. But then there is a nonzero bias term in (43), since in particular $\bar{\theta}^{(j)}(\lambda) - \hat{\theta}^{(j)}(\lambda) = \bar{\theta}^{(j)}(\lambda) \neq 0$, which contradicts the fact that $\check{\lambda}(n)$ was a minimizing sequence. ■

We now state and prove a technical Lemma which will be needed in the proof of Lemma 20.

Lemma 19 *Assume* (26) *holds; then the following conditions hold*

(i) Consider $I = [I(1), \ldots, I(p_I)]$ of size p_I to be any subset of the indices $[1, \ldots, p]$, so $p \geq p_I$ and define

$$G^{(I)} = \left[G^{(I(1))} \ldots G^{(I(p_I))} \right] \, ,$$

obtained by taking the subset of blocks of columns of G indexed by I. Then

$$c_{min} I \leq \frac{(G^{(I)})^T G^{(I)}}{n} \leq c_{max} I \, . \tag{46}$$

(ii) Let I^c be the complementary set of I in $[1, \ldots, p]$, so that $I^c \cap I = \emptyset$ and $I \cup I^c = [1, \ldots, p]$. Then the minimal angle θ_{min} between the spaces

$$\mathcal{G}^I := col\, span\{G^{(i)}/\sqrt{n}, \;\; i \in I\} \quad \text{and} \quad \mathcal{G}^{I^c} := col\, span\{G^{(j)}/\sqrt{n} : j \in I^c\}$$

satisfies:

$$\theta_{min} \geq \operatorname{acos}\left(\sqrt{1 - \frac{c_{min}}{c_{max}}} \right) > 0.$$

Proof Result (46) is a direct consequence of Horn and Johnson (1994), see Corollary 3.1.3. As far as condition (ii) is concerned we can proceed as follows: let U_I and U_{I^c} be orthonormal matrices whose columns span $\mathcal{G}^I$ and $\mathcal{G}^{I^c}$, so that there exist matrices T_I and T_{I^c} so that

$$G^{(I)}/\sqrt{n} = U_I T_I,$$
$$G^{(I^c)}/\sqrt{n} = U_{I^c} T_{I^c}$$

where $G^{(I^c)}$ is defined analogously to $G^{(I)}$. The minimal angle between $\mathcal{G}^I$ and $\mathcal{G}^{I^c}$ satisfies

$$\cos(\theta_{min}) = \left\| U_I^\top U_{I^c} \right\|.$$

Now observe that, up to a permutation of the columns which is irrelevant, $G/\sqrt{n} = [U_I T_I \quad U_{I^c} T_{I^c}]$, so that

$$U_I^\top G/\sqrt{n} = [T_I \quad U_I^\top U_{I^c}\ T_{I^c}] = [I \quad U_I^\top U_{I^c}] \begin{bmatrix} T_I & 0 \\ 0 & T_{I^c} \end{bmatrix}.$$

Denoting with $\sigma_{min}(A)$ and $\sigma_{max}(A)$ the minimum and maximum singular values of a matrix A, it is a straightforward calculation to verify that the following chain of inequalities holds:

$$\begin{aligned} c_{min} = \sigma_{min}(G^\top G/n) \leq \sigma^2_{min}\left(U_I^\top G/\sqrt{n}\right) &= \sigma^2_{min}\left([I \quad U_I^\top U_{I^c}] \begin{bmatrix} T_I & 0 \\ 0 & T_{I^c} \end{bmatrix}\right) \\ &\leq \sigma^2_{min}\left([I \quad U_I^\top U_{I^c}]\right) \sigma^2_{max}\left(\begin{bmatrix} T_I & 0 \\ 0 & T_{I^c} \end{bmatrix}\right) \\ &= \sigma^2_{min}\left([I \quad U_I^\top U_{I^c}]\right) \max\left(\sigma^2_{max}(T_I), \sigma^2_{max}(T_{I^c})\right) \\ &\leq \sigma^2_{min}\left([I \quad U_I^\top U_{I^c}]\right) c_{max}. \end{aligned}$$

Observe now that $\sigma^2_{min}\left([I \quad U_I^\top U_{I^c}]\right) = 1 - \cos^2(\theta_{min})$ so that

$$c_{min} \leq (1 - \cos^2(\theta_{min}))c_{max}$$

and, therefore,

$$\cos^2(\theta_{min}) \leq 1 - \frac{c_{min}}{c_{max}}$$

from which the thesis follows. ■

Lemma 20 *Assume that the spectrum of G satisfies* (25)*. For any index i, let $I_1^{(i)}$ and $I_0^{(i)}$ be as in* (33)*. Finally, assume $a\lambda_j^n$, which may depend on n, are bounded and satisfy:*

$$\lim_{n\to\infty} f_n = +\infty \quad \text{where} \quad f_n := \min_{j \in I_1^{(i)}} n\lambda_j^n. \tag{47}$$

Then, conditioned on θ, $\epsilon_n^{(i)}$ in (30) *and* (29) *can be decomposed as*

$$\epsilon_n^{(i)} = m_{\epsilon_n}(\theta) + v_{\epsilon_n}.$$

The following conditions hold:

$$\mathbb{E}_v\left[\epsilon_n^{(i)}\right] = m_{\epsilon_n}(\theta) = O_P\left(\frac{1}{\sqrt{f_n}}\right) \qquad v_{\epsilon_n} = O_P\left(\frac{1}{\sqrt{n}}\right) \tag{48}$$

so that $\epsilon_n^{(i)}|\theta$ *converges to zero in probability (as* $n \to \infty$*). In addition*

$$Var_v\{\epsilon_n^{(i)}\} = \mathbb{E}_v\left[v_{\epsilon_n} v_{\epsilon_n}^\top\right] = O_P\left(\frac{1}{n}\right). \tag{49}$$

If in addition[5]

$$n^{1/2}\frac{\left(G^{(i)}\right)^\top G^{(j)}}{n} = O_P(1) \;;\; j \in I_1^{(i)} \tag{50}$$

then

$$m_{\epsilon_n}(\theta) = O_P\left(\frac{1}{\sqrt{n f_n}}\right). \tag{51}$$

Proof Consider the Singular Value Decomposition

$$\bar{P}_1 \bar{S}_1 \bar{P}_1^\top := \frac{1}{n}\sum_{j\in I_1} G^{(j)}\left(G^{(j)}\right)^\top \lambda_j^n. \tag{52}$$

Using (47), there exist $\bar{n}$ so that, $\forall\ n > \bar{n}$ we have $0 < \lambda_j^n \leq M < \infty$, $j \in I_1^{(i)}$. Otherwise, we could find a subsequence n_k so that $\lambda_j^{n_k} = 0$ and hence $n_k \lambda_j^{n_k} = 0$, contradicting (47). Therefore, the matrix $\bar{P}_1$ in (52) is an orthonormal basis for the space $\mathcal{G}_1 := col\ span\{G^{(j)}/\sqrt{n} : j \in I_1^{(i)}\}$. Let also $T^{(j)}$ be such that $G^{(j)}/\sqrt{n} = \bar{P}_1 T^{(j)}$, $j \in I_1^{(i)}$. Note that by assumption (25) and lemma 19

$$\|T^{(j)}\| = O_P(1) \quad \forall\, j \in I_1^{(i)}. \tag{53}$$

Consider now the Singular Value Decomposition

$$\begin{aligned}\begin{bmatrix} P_1 & P_0 \end{bmatrix}\begin{bmatrix} S_1 & 0 \\ 0 & S_0 \end{bmatrix}\begin{bmatrix} P_1^\top \\ P_0^\top \end{bmatrix} &:= \underbrace{\frac{1}{n}\sum_{j\in I_1} G^{(j)}\left(G^{(j)}\right)^\top \lambda_j^n}_{\bar{P}_1\bar{S}_1\bar{P}_1^\top} + \underbrace{\frac{1}{n}\sum_{j\in I_0} G^{(j)}\left(G^{(j)}\right)^\top \lambda_j^n}_{\Delta.} \\ &= \bar{P}_1\bar{S}_1\bar{P}_1^\top + \Delta.\end{aligned} \tag{54}$$

For future reference note that $\exists T_{\bar{P}_1} \;:\; \bar{P}_1 = \begin{bmatrix} P_1 & P_0 \end{bmatrix} T_{\bar{P}_1}$. Now, from (28) we have that

$$\frac{\Sigma_{\bar{v}}^{(i)\,-1} G^{(i)}}{\sqrt{n}} V_n^{(i)}\left(D_n^{(i)}\right)^{-1} = \Sigma_{\bar{v}}^{(i)\,-1/2} U_n^{(i)}. \tag{55}$$

Using (55) and defining

$$P := \begin{bmatrix} P_1 & P_0 \end{bmatrix} \quad S := \begin{bmatrix} S_1 & 0 \\ 0 & S_0 \end{bmatrix},$$

5. This is equivalent to say that the columns of $G^{(j)}$, $j = 1, .., k$, $j \neq i$ are asymptotically orthogonal to the columns of $G^{(i)}$.

Equation (30) can be rewritten as:

$$\begin{aligned}
\epsilon_n^{(i)} &= \left(U_n^{(i)}\right)^\top \frac{{\Sigma_{\bar{v}}^{(i)}}^{-1/2}\bar{v}}{\sqrt{n}} \\
&= \left(D_n^{(i)}\right)^{-1}\left(V_n^{(i)}\right)^\top \frac{\left(G^{(i)}\right)^\top}{\sqrt{n}}{\Sigma_{\bar{v}}^{(i)}}^{-1}\frac{\bar{v}}{\sqrt{n}} \\
&= \left(D_n^{(i)}\right)^{-1}\left(V_n^{(i)}\right)^\top \frac{\left(G^{(i)}\right)^\top}{\sqrt{n}}\left[\begin{array}{cc} P & P_\perp \end{array}\right]\left[\begin{array}{cc} (nS+\sigma^2 I)^{-1} & 0 \\ 0 & \sigma^{-2}I \end{array}\right]\left[\begin{array}{c} P^\top \\ P_\perp^\top \end{array}\right]\frac{\bar{v}}{\sqrt{n}} \\
&= \left(D_n^{(i)}\right)^{-1}\left(V_n^{(i)}\right)^\top \frac{\left(G^{(i)}\right)^\top}{\sqrt{n}}\left[\begin{array}{cc} P & P_\perp \end{array}\right]\left[\begin{array}{cc} (nS+\sigma^2 I)^{-1} & 0 \\ 0 & \sigma^{-2}I \end{array}\right]\left[\begin{array}{c} P^\top \\ P_\perp^\top \end{array}\right] \\
&\times \left[\sum_{j\in I_1^{(i)}}\frac{G^{(j)}}{\sqrt{n}}\theta^{(j)}+\frac{v}{\sqrt{n}}\right] \\
&= \underbrace{\left(D_n^{(i)}\right)^{-1}\left(V_n^{(i)}\right)^\top \frac{\left(G^{(i)}\right)^\top}{\sqrt{n}}P(nS+\sigma^2 I)^{-1}\left[\begin{array}{c} P_1^\top P_1 \\ P_0^\top P_1 \end{array}\right]\sum_{j\in I_1}T^{(j)}\theta^{(j)}}_{m_{\epsilon_n}(\theta)} + \\
&\quad \underbrace{+\left(D_n^{(i)}\right)^{-1}\left(V_n^{(i)}\right)^\top \frac{\left(G^{(i)}\right)^\top}{\sqrt{n}}\left[\begin{array}{cc} P & P_\perp \end{array}\right]\left[\begin{array}{cc} (nS+\sigma^2 I)^{-1} & 0 \\ 0 & \sigma^{-2}I \end{array}\right]\frac{v_{\bar{P}}}{\sqrt{n}}}_{v_{\epsilon_n}}
\end{aligned}$$

where the last equation defines $m_{\epsilon_n}(\theta)$ and v_{ϵ_n}, the noise

$$v_{\bar{P}} := \left[\begin{array}{c} P^\top \\ P_\perp^\top \end{array}\right] v$$

is still a zero mean Gaussian noise with variance $\sigma^2 I$ and $\frac{G^{(j)}}{\sqrt{n}} = \bar{P}_1 T^{(j)}$ provided $j \neq i$. Note that m_{ϵ_n} does not depend on v and that $\mathbb{E}_v v_{\epsilon_n} = 0$. Therefore $m_{\epsilon_n}(\theta)$ is the mean (when only noise v is averaged out) of ϵ_n. As far as the asymptotic behavior of $m_{\epsilon_n}(\theta)$ is concerned, it is convenient to first observe that

$$(nS+\sigma^2 I)^{-1}\left[\begin{array}{c} P_1^\top \bar{P}_1 \\ P_0^\top \bar{P}_1 \end{array}\right] = \left[\begin{array}{c} (nS_1+\sigma^2 I)^{-1}P_1^\top \bar{P}_1 \\ (nS_0+\sigma^2 I)^{-1}P_0^\top \bar{P}_1 \end{array}\right]$$

and that the second term on the right hand side can be rewritten as

$$(nS_0+\sigma^2 I)^{-1}P_0^\top \bar{P}_1 = \left[\begin{array}{c} (n[S_0]_{1,1}+\sigma^2)^{-1}P_{0,1}^\top \bar{P}_1 \\ (n[S_0]_{22}+\sigma^2)^{-1}P_{0,2}^\top \bar{P}_1 \\ \vdots \\ (n[S_0]_{m-k,m-k}+\sigma^2)^{-1}P_{0,m-k}^\top \bar{P}_1 \end{array}\right] \tag{56}$$

where $[S_0]_{ii}$ is the $i-th$ diagonal element of S_0 and $P_{0,i}$ is the $i-th$ column of P_0. Now, using Equation (54) one obtains that

$$\begin{aligned}
n[S_0]_{ii} = P_{0,i}^\top P n S P^\top P_{0,i} &= P_{0,i}^\top\left(\bar{P}_1 n\bar{S}_1\bar{P}_1^\top + n\Delta\right)P_{0,i} \\
&\geq P_{0,i}^\top \bar{P}_1 n\bar{S}_1\bar{P}_1^\top P_{0,i} \\
&\geq \sigma_{min}(n\bar{S}_1)P_{0,i}^\top\bar{P}_1\bar{P}_1^\top P_{0,i} \\
&= \sigma_{min}(n\bar{S}_1)\|P_{0,i}^\top\bar{P}_1\|^2.
\end{aligned}$$

An argument similar to that used in (42) shows that

$$\sigma_{min}(n\bar{S}_1) \geq c_{min}\min\{n\lambda_j^n,\ j \in I_1^{(i)}\} = c_{min} f_n \tag{57}$$

also holds true; denoting $\|P_{0,i}^\top \bar{P}_1\| = g_n$, the generic term on the right hand side of (56) satisfies

$$\begin{array}{rcl} \|(n[S_0]_{ii} + \sigma^2)^{-1} P_{0,i}^\top \bar{P}_1\| & \leq & \frac{\|P_{0,i}^\top \bar{P}_1\|}{n\sigma_{min}(\bar{S}_1)\|P_{0,i}^\top \bar{P}_1\|^2 + \sigma^2} \\ & \leq & k\ \min(g_n, (f_n g_n)^{-1}) \\ & = & \frac{k}{\sqrt{f_n}} \min(\sqrt{f_n} g_n, (\sqrt{f_n} g_n)^{-1}) \\ & \leq & \frac{k}{\sqrt{f_n}} \end{array} \tag{58}$$

for some positive constant k. Now, using Lemma 13, $D_n^{(i)}$ is bounded and bounded away from zero in probability, so that $\|D_n^{(i)}\| = O_P(1)$ and $\| \left(D_n^{(i)}\right)^{-1} \| = O_P(1)$. In addition, $V_n^{(i)}$ is an orthonormal matrix and $\|\frac{G^{(i)}}{\sqrt{n}}\| = O_P(1)$. Last, using (57) and (25), we have $\|(nS_1 + \sigma^2 I)^{-1}\| = O_P(1/n)$. Combining these conditions with (53) and (58), we obtain the first expression in (48). As far as the asymptotics on v_{ϵ_n} are concerned, it suffices to observe that

$$w_n^\top v_{\bar{P}}/\sqrt{n} = O_P(1/\sqrt{n}) \ \text{ if : } \ \|w_n\| = O_P(1).$$

The variance (w.r.t. noise v) $Var_v\{\epsilon_n\} = \mathbb{E}_v\left[v_{\epsilon_n} v_{\epsilon_n}^\top\right]$ satisfies

$$Var_v\{\epsilon_n\} = \frac{\sigma^2}{n} \left(U_n^{(i)}\right)^\top \Sigma_{\bar{v}}^{(i)^{-1}} \left(U_n^{(i)}\right)$$

so that, using the condition $\left\| \Sigma_{\bar{v}}^{(i)^{-1}} \right\| = \sigma^{-2}$ derived in Lemma 13, and the fact that $U_n^{(i)}$ has orthonormal columns, the condition $Var_v\{\epsilon_n\} = O_P\left(\frac{1}{n}\right)$ in (49) follows immediately.

If, in addition, (50) holds then (53) becomes

$$\|T^{(j)}\| = O_P(1/\sqrt{n}) \quad j = 1, ..., k; \quad j \neq k$$

so that an extra $\sqrt{n}$ appears in the denominator in the expression of $m_\epsilon(\theta)$ yielding (51). This concludes the proof. ■

Before we proceed, we review a useful characterization of convergence. While it can be stated for many types of convergence, we present it specifically for convergence in probability, since this is the version we will use.

Lemma 21 *The sequence a^n converges in probability to a (written $a^n \to_p a$) if and only if every subsequence $a^{n(j)}$ of a^n has a further subsequence $a^{n(j(k))}$ with $a^{n(j(k))} \to_p a$.*

Proof If $a^n \to_p a$, this means that for any $\epsilon > 0$, $\delta > 0$ there exists some $n_{\epsilon,\delta}$ such that for all $n \geq n_{\epsilon,\delta}$, we have $P(|a^n - a| > \epsilon) \leq \delta$. Clearly, if $a^n \to_p a$, then $a^{n(j)} \to_p a$ for every subsequence $a^{n(j)}$ of a^n. We prove the other direction by contrapositive.

Assume that $a^n \not\to_p a$. That means precisely that there exist some $\epsilon > 0, \delta > 0$ and a subsequence $a^{n(j)}$ so that $P(|a - a^{n(j)}| > \epsilon) \geq \delta$. Therefore the subsequence $a^{n(j)}$ cannot have further subsequences that converge to a in probability, since every term of $a^{n(j)}$ stays ϵ-far away from a with positive probability δ. ■

Lemma 21 plays a major role in the proof of the main result.

A.6 Proof of Theorem 14

Since the hypotheses of Lemma 13 hold, we know $w_{k,n} \to 0$ in (32). Then Lemma 18 guarantees that condition (47) holds true (in probability) so that Lemma 20 applies, and therefore $v_{k,n} \to_p 0$ in (32). We now give the proofs of results 1-4 in Theorem 14.

1. The reader can quickly check that $\frac{d}{d\gamma}\bar{\lambda}_1^\gamma < 0$, so $\bar{\lambda}_1^\gamma$ is decreasing in γ. The limit calculation follows immediately from L'Hopital's rule yielding $\lim_{\gamma\to 0^+} \lambda_1^\gamma = \bar{\lambda}_1$.

2. We use the convergence characterization given in Lemma 21. Pick any subsequence $\hat{\lambda}_1^{n(j)}$ of $\hat{\lambda}_1^n$. Since $\{V_{n(j)}\}$ is bounded, by Bolzano-Weierstrass it must have a convergent subsequence $V_{n(j(k))} \to V$, where V satisfies $V^TV = I$ by continuity of the 2-norm. The first-order optimality conditions for $\hat{\lambda}_1^n > 0$ are given by

$$0 = f_1(\lambda, w, v, \eta) = \frac{1}{2}\sum_{k=1}^{k_1} \frac{-\eta_k^2 - v_k}{(\lambda + w_k)^2} + \frac{1}{\lambda + w_k} + \gamma \,, \tag{59}$$

   and we have $f_1(\lambda, 0, 0, V^T\bar{\theta}^{(1)}) = 0$ if and only if $\lambda = \bar{\lambda}_1^\gamma$. Taking the derivative we find

$$\frac{d}{d\lambda} f_1(\lambda, 0, 0, V^T\bar{\theta}^{(1)}) = \frac{\|\bar{\theta}^{(1)}\|^2}{\lambda^3} - \frac{k_1}{2\lambda^2} \,,$$

   which is nonzero at λ_1^γ for any γ, since the only zero is at $2\frac{\|\bar{\theta}^{(1)}\|^2}{k_1} = 2\bar{\lambda}_1 \geq 2\bar{\lambda}_1^\gamma$.

   Applying the Implicit Function Theorem to f at $\left(\lambda_1^\gamma, 0, 0, V^\top\bar{\theta}^{(1)}\right)$ yields the existence of neighborhoods $\mathcal{U}$ of $(0, 0, V^\top\bar{\theta}^{(1)})$ and $\mathcal{W}$ of λ_1^γ such that

$$f(\phi(w, v, \eta), w, v, \eta) = 0 \qquad \forall\, (w, v, \eta) \in \mathcal{U} \,.$$

   In particular, $\phi(0, 0, V^\top\bar{\theta}^{(1)}) = \lambda_1^\gamma$. Since $(w_{n(j(k))}, v_{n(j(k))}, \eta_{n(j(k))}) \to_p (0, 0, V^\top\bar{\theta}^{(1)})$, we have that for any $\delta > 0$ there exist some k_δ so that for all $n(j(k)) > n(j(k_\delta))$ we have $P((w_{n(j(k)}, v_{n(j(k))}, \eta_{n(j(k))}) \notin \mathcal{U}) \leq \delta$. For anything in $\mathcal{U}$, by continuity of ϕ we have

$$\hat{\lambda}_1^{n(j(k))} = \phi(w_{n(j(k))}, v_{n(j(k))}, \eta_{n(j(k))}) \to_p \phi(0, 0, V^\top\bar{\theta}^{(1)}) = \lambda_1^\gamma \,.$$

   These two facts imply that $\hat{\lambda}_1^{n(j(k))} \to_p \lambda_1^\gamma$. We have shown that every subsequence $\hat{\lambda}_1^{n(j)}$ has a further subsequence $\hat{\lambda}_1^{n(j(k))} \to_p \lambda_1^\gamma$, and therefore $\hat{\lambda}_1^n \to_p \lambda_1^\gamma$ by Lemma 21.

3. In this case, the only zero of (59) with $\gamma = 0$ is found at $\bar{\lambda}_1$, and the derivative of the optimality conditions is nonzero at this estimate, by the computations already given. The result follows by the implicit function theorem and subsequence argument, just as in the previous case.

4. Rewriting the derivative (59)

$$\frac{1}{2}\sum_{k=1}^{k_1} \frac{\lambda - v_k - \eta_k^2 + w_k}{(\lambda + w_k)^2} + \gamma \,,$$

   we observe that for any positive λ, the probability that the derivative is positive tends to one. Therefore the minimizer λ_1^γ converges to 0 in probability, regardless of the value of γ.

References

A. Aravkin, J. Burke, A. Chiuso, and G. Pillonetto. On the estimation of hyperparameters for empirical bayes estimators: Maximum marginal likelihood vs minimum MSE. In *Proc. IFAC Symposium on System Identification (SysId 2012)*, 2012.

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*, page 4148, 2004.

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

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

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.

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. 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*, 48:1553–1565, 2012.

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

B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23: 122, 2008.

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.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, december 2001.

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.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

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.

H. Leeb and B. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:2159, 2005.

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).

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.

M. Schmidt, E. Van Den Berg, M. 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.

C.M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals 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.

M.K. Titsias and M. Lzaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. *Advances in Neural Information Processing Systems 25 (NIPS 2011)*, 2011.

R. Tomioka and T. Suzuki. Regularization strategies and empirical bayesian learning for MKL. *Journal of Machine Learning Research*, 2011.

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*, 57(9):6236–6255, 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.

H. Zou. The adaptive Lasso and it oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.