LUND UNIVERSITY

Hyperparameter selection for group-sparse regression: a probabilistic approach

T. KRONVALL AND A. JAKOBSSON

Published in: Elsevier Signal Processing

doi:10.1016/j.sigpro.2018.04.021

Lund 2018

Mathematical Statistics

Centre for Mathematical Sciences

Lund University

Hyperparameter selection for group-sparse regression: a probabilistic approach

Ted Kronvall and Andreas Jakobsson

Abstract

This work analyzes the effects on support recovery for different choices of the hyper- or regularization parameter in LASSO-like sparse and group-sparse regression problems. The hyperparameter implicitly selects the model order of the solution, and is typically set using cross-validation (CV). This may be computationally prohibitive for large-scale problems, and also often overestimates the model order, as CV optimizes for prediction error rather than support recovery. In this work, we propose a probabilistic approach to select the hyperparameter, by quantifying the type I error (false positive rate) using extreme value analysis. From Monte Carlo simulations, one may draw inference on the upper tail of the distribution of the spurious parameter estimates, and the regularization level may be selected for a specified false positive rate. By solving the e group-LASSO problem, the choice of hyperparameter becomes independent of the noise variance. Furthermore, the effects on the false positive rate caused by collinearity in the dictionary is discussed, including ways of circumventing them. The proposed method is compared to other hyperparameter-selection methods in terms of support recovery, false positive rate, false negative rate, and computational complexity. Simulated data illustrate how the proposed method outperforms CV and comparable methods in both computational complexity and support recovery.

keywords *- sparse estimation, group-LASSO, regularization, model order estimation, sparsistency, extreme value analysis*

This work was supported in part by the Swedish Research Council, Carl Trygger's foundation, and the Royal Physiographic Society in Lund.

Dept. of Mathematical Statistics, Lund University, Sweden

I. Introduction

Estimating the sparse parameter support for a high-dimensional regression problem has been the focus of much scientific attention during the past two decades, as this methodology has shown its usefulness in a wide array of applications, ranging from spectral analysis [1]–[3], array- [4]–[6] and audio processing [7]–[9], to biomedical modeling [10], magnetic resonance imaging [11], [12], and more. For many of these and for other applications, the retrieved data may be well explained using a highly underdetermined regression model, in which only a small subset of the explanatory variables are required to represent the data. The approach is typically referred to as sparse regression; the individual regressors are called atoms, and the entire regressor matrix the dictionary, which is typically customized for a particular application. The common approach of inferring sparsity on the estimates is to solve a regularized regression problem, i.e., appending the fit term with a regularization term that increases as variables become active (or non-zero). Much of the work in the research area springs from extensions on the seminal work by Tibshirani et al., wherein the least absolute selection and shrinkage operator (LASSO) [13] was introduced. The LASSO is a regularized regression problem where an ℓ_1-norm on the variable vector is used as regularizer, which in signal processing is also referred to as the basis pursuit denoising (BPDN) method [14]. Another early alternative to the LASSO problem is the penalized likelihood problem, introduced in [15].

In this paper, we focus on a generalization of the sparse regression problem, wherein the atoms of the dictionary exhibit some form of grouping behavior which is defined *a priori*. This follows the notion that a particular data feature is modeled not only using a single atom, but instead by a group of atoms, such that each atom has an unknown (and possibly independent) response variable, but where the entire group is assumed to be either present or not present in the data. This is achieved in the group-LASSO [16] by utilizing an ℓ_1/ℓ_2-regularizer, but other approaches have also been successful, such as in, e.g., [9], [10]. Being a generalization of the LASSO, the group-LASSO reverts back to the standard LASSO when the group sizes in the dictionary all have size one. Typically, results which hold for the group-LASSO thus also hold for the LASSO. One reason behind the success of LASSO-like approaches is that these are typically cast as a convex optimization problems, for which there exists strong theoretical results for convergence and recovery guarantees (see, e.g., [17]–[19], and the references therein). For convex problems, there also exist user-friendly scientific software for simple experimentation and investigation of new regularizers [20].

The sparse regression problems described here, being a subset of the regularized regression problems, have in common the requirement of selecting one or several hyperparameters, which have the role of

controlling the degree of sparsity in the solution by adjusting the level of regularization in relation to the fit term. Thus, sparsity is subject to user control, and must therefore be chosen adequately for each problem. From the perspective of model order selection, one may note that there is no currently consistent approach to finding a correct model order (see, e.g., [21]). Still, as an implicit agent of model order selection in regularization problems, one may distinguish three main methodologies of selecting the regularization level. Firstly, and perhaps most commonly are the data-driven, or post-model selection, methods, where the performance of a number of candidate models are compared in some user-selected metric. To that end, the least angle regression (LARS) algorithm [22] calculates the entire (so called) path of solutions on an interval of values for the hyperparameter of a LASSO-like problem, and at a computational cost similar to solving the LASSO for a single value of the hyperparameter. However, by using warm-starts, a solution path may also be calculated quickly using some appropriate implementation of the group-LASSO. A single point on the solution path is then chosen based on user preference; most commonly prediction performance, i.e., using cross-validation (CV), as was done, for instance, in [23] for the multi-pitch estimation problem. However, due to the computationally burdensome process of CV, one often instead reverts to using heuristic data-driven approaches, or choosing the hyperparameter based on some information criteria (see, e.g., [24]). Another interesting contribution was made in [25], wherein a covariance test statistic was used to determine whether to include every new regressor along a path of regularization values. Bayesian techniques offer another common approach to the model order selection problem, wherein the joint posterior distribution of the regression variables and the hyperparameter are utilized, under the assumption on statistical priors on these, such as in, e.g., [26]. The third main group of approaches may be considered to be probabilistic in the sense that they make assumptions on only the noise statistics of the measured signal, and not the regression variables. Among these, the approach suggested in [27] might be most prominent (here, for simplicity, referred to as CDS), where in order to suppress the noise components from propagating into the estimate, an upper-endpoint of the distribution is used, such that for independent Gaussian regressors (i.e., orthogonal dictionaries), the largest interfering noise components in the limit grows in proportion to some quantity. Under these assumptions, CDS is ostensibly blind, but will by construction set the regularization level high enough to guarantee noise suppression, and might thereby also suppress the signal-of-interest. In applications containing atoms with a high degree of collinearity, thereby violating the orthogonal assumption, this will result in overshooting of the regularization level. To simplify the selection of regularization level, the scaled LASSO [28] reparametrizes the hyperparameter by introducing an auxiliary variable describing the standard deviation of the model residual. This has the effect that the regularization level may be selected (somewhat)

independently of the noise variance, which is useful for the probabilistic approaches.

Another method of selecting the regularization level that might fall into the probabilistic category is the sparse iterative covariance-based estimation (SPICE) method, which yields a relatively sparse parameter support by matching the observed covariance matrix and a covariance matrix parametrized by a dictionary. The method has been shown to work well for a variety of applications, especially those pertaining to estimation of line spectra and directions-of-arrival (see, e.g., [29]). In subsequent publications (see, e.g., [29], [30]), SPICE was shown to be equivalent to either the least absolute deviation (LAD) LASSO under a heterscedastic noise assumption, or the square root (SR) LASSO under a homoscedastic noise assumption, both for particular choices of the hyperparameter. It may be shown that the SR LASSO and the scaled LASSO are equivalent, and we conclude that SPICE is a robust (and possibly heuristic) approach of fixing the hyperparameter (somewhat) independently of the noise level. In a recent effort, the SPICE approach was extended for group sparsity [31], showing promising results, e.g., for multi-pitch estimation, but also illustrating how the fixed hyperparameter yields estimates which are not as sparse as one may typically expect. A valid argument in defence of the SPICE approach is that the measure of 'good' in sparse estimation is not entirely straightforward, and not sparse enough may still be good enough.

Borrowing some terminology from detection theory [32], one way of measuring performance is to calculate the false negatives (FNs), i.e., whether atoms pertaining to the true support of the signal (those atoms of which the data is truly composed) are estimated as zero for some choice of the hyperparameter. As the SPICE regularization level is typically set too low, the possibility of FNs is consequently also low, which for some applications may be the focus. Conversely, for some applications, the focus may be to eliminate the false positives (FPs), i.e., when noise components are falsely set to be non-zero while not being in the true support set. The FPs and FNs are also sometimes referred to as the type I and type II errors, respectively. In addition, a metric called sparsistency is sometimes used, measuring the binary output of whether the estimated and the true supports are identical, which is the complement of the union between FN and FP [33]. Sparsistency might also be unobtainable for a certain problem; avoiding FPs requires selecting the hyperparameter so large that FNs will arise, and similarly avoiding FNs will result in more FPs. Model order estimation can thus be seen as prioritizing between FPs and FNs, which is also referred to as the bias-variance trade off, and has a long history in the literature. Typically, model order estimation can be formulated as a series of hypothesis tests, subsequently tested using, e.g., F-test statistics for some specified significance level [34].

In this paper, we further this development, formulating a probabilistic method for hyperparameter

selection using hypothesis testing. By analyzing how the noise components propagate into the parameter estimates for different estimators and different choices of the hyperparameters, we seek to increase the sparsistency of the group-LASSO estimate by means of optimizing the FP rate. By making assumptions on the noise distribution and then sampling from the corresponding extreme value distribution using the Monte Carlo method, the hyperparameter is chosen as an appropriate quantile of the largest anticipated noise components. Avoiding FPs can never be guaranteed without maximizing the regularization level, thereby setting the entire solution to zero, but the risk may be quantified. By specifying the type I error, the sparsistency rate is also indirectly controlled, whenever this is feasible. Furthermore, for Gaussian noise, we show that the distribution for the maximum noise components follows a type I extreme value distribution (Gumbel), from which a parametric quantile may be obtained at a low computational cost.

For coherent dictionaries, i.e., where there is a high degree of collinearity between the atoms, many of the theoretical guarantees for sparse estimation will fail to hold, along with a few of the methods themselves. The effects on the estimates for the collinear atoms are difficulty to discern; depending on the problem either all of them, or just a few of them, become non-zero. Coherence therefore typically results in FPs, if the regularization level is not increased, which in turn might yield FNs. There exists some approaches of dealing with coherent dictionaries. The elastic net uses a combination of ℓ_1 and ℓ_2 penalties [35], with the effect of increasing the inclusion of coherent components, thereby avoiding some FNs, but still not decreasing the number of FPs. Another popular approach is the reweighted LASSO [36], which solves a series of LASSO problems where the regularization level is individually set for each atom inversely proportional to its previous estimate. This approach approximates the use of a (non-convex) logarithmic regularizer, which allows the estimates to better reallocate power to the strongest of the coherent atoms. The proposed approach does not account for the leakage of power between coherent components in the true support, but only for the coherence effects on the assumed noise. As a remedy, the proposed method instead solves the reweighted LASSO problem at the chosen regularization level.

To illustrate the achievable performance of the proposed method, numerical results show how the proposed method for selecting the hyperparameter is much less computationally demanding than both CV and the Bayesian information criterion (BIC), and at the same time being more accurate than the CV, BIC, and the probabilistic CDS method. These claims are verified for both the sparse and the group-sparse regression problems.

The remainder of this paper is organized as follows: Section II defines the mathematical notation used throughout the paper, whereafter Section III describes some background on the group-LASSO and the scaled group-LASSO problems, also including an implementation of the cyclic coordinate descent solver

for these problems. Section IV desribes the proposed method of selecting the regularization level. Section V then describes how the estimate of the noise standard deviation may be improved, and section VI how FPs due to coherence may be dealt with. Thereafter, for completeness, Sections VII and VIII describe the earlier approaches that our work mainly relates and compares with, being the CV, BIC, and CDS methods, respectively. Section IX shows some numerical results, and, finally, Section X concludes upon the presented results.

II. Notational conventions

In this paper, we use the mathematical convention of letting lower-case letters, e.g., y, denote scalars, while lower-case bold-font letters, $\mathbf{y}$, denote column vectors wheras upper-case bold-font letters, $\mathbf{Y}$, denote matrices. Furthermore, E, V, D, and P denote the expectation, variance, standard deviation, and probability of a random variable or vector, respectively. We let $|\cdot|$ denote the absolute value of a complex-valued number, while $\|\cdot\|_p$ and $\|\cdot\|_\infty$ denote the p-norm and the maximum-norm, respectively. Furthermore, $\text{diag}(\mathbf{a}) = \mathbf{A}$ denotes the diagonal matrix with diagonal vector $\mathbf{a}$, although $\text{diag}(\mathbf{A}) = \mathbf{a}$ is also denoting the diagonal vector of a square matrix. As is conventional, we let $(\cdot)^\top$ and $(\cdot)^H$ denote the transpose and hermitian transpose, respectively. Subscripts are used to denote a subgroup of a vector or matrix, and superscript typically denotes a power operation, except when the exponent is within parentheses or hard brackets, which we use to denote iteration number, e.g., $x^{(j)}$, is j:th iteration of x, and the index of a random sample, e.g., $x^{[j]}$ denotes the j:th realization of the random variable x. We also make use of the notations $(x)_+ = \max(0, x)$, $\text{sign}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$, and $x \sim F$, which states that the random variable x has distribution function F. Finally, we let $\varnothing$ denote the empty set.

III. Group-sparse regression via coordinate descent

Consider a noisy N-sample complex-valued vector $\mathbf{y}$, which may be well described using the linear regression model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \tag{1}$$

where $\mathbf{A} \in \mathbb{C}^{N\times M}$, where $M \gg N$, is the (possibly highly) underdetermined regressor matrix, or dictionary, constructed from a set of normalized regressors, i.e., $\mathbf{a}_i^H \mathbf{a}_i = 1,\ i = 1, \ldots, M$, with $\mathbf{a}_i$ denoting the i:th column of $\mathbf{A}$. The unknown parameter vector $\mathbf{x}$ is assumed to have a C/M-sparse parameter support, i.e., only $C < N$ of the parameters in $\mathbf{x}$ are assumed to be non-zero. In this paper, we consider the generalized case where the dictionary may contain groups of regressors whose components

are linked in a modeling sense, such that the model parametrizes a superposition of objects, each of which is modeled by a group of regressors rather than just one. Therefore, we let the dictionary be constructed such that the M regressors are collected into K groups with L_k regressors in the k:th group, i.e.,

$$\mathbf{A} = \left[\begin{array}{ccc} \mathbf{A}_1 & \dots & \mathbf{A}_K \end{array} \right] \tag{2}$$

$$\mathbf{A}_k = \left[\begin{array}{ccc} \mathbf{a}_{k,1} & \dots & \mathbf{a}_{k,L_k} \end{array} \right] \tag{3}$$

and where, similarly,

$$\mathbf{x} = \left[\begin{array}{ccc} \mathbf{x}_1^\top & \dots & \mathbf{x}_K^\top \end{array} \right]^\top \tag{4}$$

Furthermore, we assume that the observation noise, $\mathbf{e}$, may be well modeled as an i.i.d. multivariate random variable, such that $\mathbf{e} = \sigma\mathbf{w}$, where $\mathbf{w} \sim F$, for some sufficiently symmetric distribution F with unit variance. Let $\hat{\mathbf{x}}(\lambda)$ denote the solution to the convex optimization problem

$$\underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_{k=1}^{K} \|\mathbf{x}_k\|_2 \tag{5}$$

for some hyperparameter $\lambda > 0$. This is the group-LASSO estimate, for which we briefly outline the corresponding cyclic coordinate descent (CCD) algorithm. In its essence, CCD updates the parameters in $\mathbf{x}$ one at a time, by iteratively minimizing $f(\mathbf{x})$ for each x_i, $i = 1, \dots, M$, in random order. As $\mathbf{x}$ is complex-valued, and as $f(\mathbf{x})$ is non-differentiable for $\mathbf{x}_k = 0$, for any k, we exploit Wirtinger calculus and subgradient analysis to form derivatives of f. Let $\mathbf{r}_k = \mathbf{y} - \mathbf{A}_{-k}\mathbf{x}_{-k}$ be the residual vector where the effect of the k:th variable group has been left out, i.e., such that $\mathbf{A}_{-k}$ and $\mathbf{x}_{-k}$ omit the k:th regressor and variable group, respectively. Thus, one may find the derivative of $f(\mathbf{x})$ with respect to $\mathbf{x}_k$ and set it to zero, i.e.,

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_k} = -\mathbf{A}_k^H \left(\mathbf{r}_k - \mathbf{A}_k\mathbf{x}_k\right) + \lambda \mathbf{u}_k = \mathbf{0} \tag{6}$$

where

$$\mathbf{u}_k = \begin{cases} \text{sign}(\mathbf{x}_k) & \mathbf{x}_k \neq \mathbf{0} \\ \{\mathbf{u}_k : \|\mathbf{u}_k\|_2 \leq 1\} & \mathbf{x}_k = \mathbf{0} \end{cases} \tag{7}$$

Under the assumption that $\mathbf{A}_k^H \mathbf{A}_k = \boldsymbol{I}$, a closed-form solution may be found as

$$\hat{\mathbf{x}}_k(\lambda) = \mathcal{T}(\mathbf{A}_k^H \mathbf{r}_k, \lambda) \tag{8}$$

where $\mathcal{T}(\mathbf{z}, \alpha) = \text{sign}(\mathbf{z}) \left(\|\mathbf{z}\|_2 - \alpha\right)_+$ denotes the group-shrinkage operator. The group-LASSO estimate is thus formed by the inner product between the residual and regressor groups, albeit where the groups'

ℓ_2-norms are reduced by λ. Therefore, the estimate $\hat{\mathbf{x}}$ will be biased towards zero. Similarly, sparsity in groups is induced as the groups having an inner product with ℓ_2-norm smaller than λ are set to zero. The regularization parameter thus serves as an implicit model order selector. In particular, the zeroth model order, $\hat{\mathbf{x}} = \mathbf{0}$, is obtained for $\lambda \geq \lambda_0 = \max_k \left\|\mathbf{A}_k^H \mathbf{y}\right\|_2$. Let the true support set be denoted by $\mathcal{I} = \{k : \|\mathbf{x}_k\| \neq 0\}$. When decreasing λ, the model order grows, introducing the parameter group $k \in \hat{\mathcal{I}}(\lambda) \Leftrightarrow \left\|\mathbf{A}_k^H \mathbf{r}_k\right\|_2 > \lambda$. As a consequence, parameter groups are included in the support set in an order determined by their magnitude. In the case that $\left\|\mathbf{A}_k^H \mathbf{r}_k\right\| > \left\|\mathbf{A}_{k'}^H \mathbf{r}_{k'}\right\|_2$, then the implication $k' \in \hat{\mathcal{I}}(\lambda) \Rightarrow k \in \hat{\mathcal{I}}(\lambda)$ is always true, and a parameter group with some smaller ℓ_2-norm is never in the solution set if another one with a larger ℓ_2-norm is not. Selecting an appropriate regularization level is thus important; if set too large, the solution will have omitted parts of the sought signal, if set too small, the solution will include noise components and be too dense. Recently, the scaled LASSO was introduced, solving the optimization problem (here in group-version) [28]

$$\underset{\mathbf{x},\sigma>0}{\text{minimize}}\ g(\mathbf{x},\sigma) = \frac{1}{\sigma}\left\|\mathbf{y} - \mathbf{A}\mathbf{x}\right\|_2^2 + N\sigma + \mu\sum_{k=1}^{K}\left\|\mathbf{x}_k\right\|_2 \tag{9}$$

i.e., a modification of the group-LASSO where the auxilliary variable σ, representing the residual standard deviation, scales the least squares term, and where $\mu > 0$ is the regularization parameter. As the non-grouped scaled LASSO, (9) is jointly convex over $(\mathbf{x}, \sigma)$ for $\sigma > 0$; the first term may be identified as the convex quadratic-over-linear function (see, [17, p. 73]), and (9) is therefore a sum of convex functions, which is convex. Thus, utilizing a CCD approach, σ may be included in the cyclic optimization scheme; its resulting optimization problem with $\mathbf{x}$ fixed is strictly convex and has the unique minimizer

$$\hat{\sigma}(\mu) = \frac{\left\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}(\mu)\right\|_2}{\sqrt{N}} \tag{10}$$

Similar to the derivations above, the $\mathbf{x}_k$s may be iteratively estimated as

$$\hat{\mathbf{x}}_k(\mu) = \mathcal{T}(\mathbf{A}_k^H \mathbf{r}_k, \hat{\sigma}\mu) \tag{11}$$

which are thus regularized by $\sigma\mu$, making μ seemingly independent of the noise power. However, the estimate of σ is itself clearly affected by μ. For too low values of μ, typically $\hat{\sigma}(\mu) < \sigma$, i.e., smaller than the true noise standard deviation, as too much of the noise components will be modeled by $\hat{\mathbf{x}}(\mu)$. Similarly, when μ is too large, $\hat{\sigma}(\mu) > \sigma$ as it will also model part of the signal variability. However, even when the regularization level is chosen appropriately, one still has $\hat{\sigma}(\mu) \geq \sigma$ in general due to the estimation bias in $\hat{\mathbf{x}}(\mu)$. It is also worth noting that by inserting $\hat{\sigma}$ into $g(\mathbf{x}, \sigma)$, one obtains the equivalent

Algorithm 1 Scaled group-LASSO via cyclic coordinate descent

Intialize $\mathbf{x}^{(0)} = \mathbf{0}, \mathbf{r} = \mathbf{y}$, and $j = 1$
while $j < j_{\max}$ **do**
  $\sigma \leftarrow \|\mathbf{r}\|_2 / \sqrt{N}$
  $I_j = \mathcal{U}(1, \ldots, K)$
  for $i \in I_j$ **do**
    $\mathbf{r} \leftarrow \mathbf{r} + \mathbf{A}_i \mathbf{x}_i^{(j-1)}$
    $\mathbf{x}_i^{(j)} = \mathcal{T}(\mathbf{A}_i^H \mathbf{r}, \sigma\mu)$
    $\mathbf{r} \leftarrow \mathbf{r} - \mathbf{A}_i \mathbf{x}_i^{(j)}$
  end for
  if $\left\|\mathbf{x}^{(j)} - \mathbf{x}^{(j-1)}\right\|_2 \leq \kappa_{\text{tol}}$ **then**
    break
  end if
  $j \leftarrow j + 1$
end while

optimization problem

$$\underset{\mathbf{x}}{\text{minimize}}\ g(\mathbf{x}) = 2\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \frac{\mu}{\sqrt{N}} \sum_{k=1}^{K} \|\mathbf{x}_k\|_2 \tag{12}$$

which may be identified as the square-root group-LASSO [37]. Algorithm 1 outlines the CCD solver for the scaled group-LASSO problem at some regularization level μ, where $j_{\max}$, κ_{tol}, and $\mathcal{U}(\cdot)$ denote the maximum number of iterations, the convergence tolerance, and a random permutation of indices, respectively. Here, when comparing (11) to (8), it becomes apparent that the group-LASSO and the scaled group-LASSO will yield identical solutions when $\lambda = \sigma\mu$. The motivation behind using the scaled group-LASSO is instead that the regularization level may be chosen independently of the noise variance. In the next section, we make use of this property.

IV. A probabilistic approach to regularization

Consider the overall aim of selecting the hyperparameter in order to maximize sparsistency, i.e., selecting λ such that the estimated support coincides with the true support,

$$\lambda = \{\lambda : \hat{\mathcal{I}}(\lambda) = \mathcal{I}\} \tag{13}$$

From the perspective of detection theory, whenever the support recovery fails, at least one of the following errors have occurred:

- False positive (FP) or type-I error: the regularization level is set too low and the estimated support contains indices which are not in the true support; $\left(\hat{\mathcal{I}}(\lambda) \cap \mathcal{I}^c\right) \neq \varnothing$, where $\mathcal{I}^c$ denotes the complement of the support set.
- False negative (FN) or type-II error: the regularization level is set too high and the estimated support set does not contain all indices in the true support; $\left(\hat{\mathcal{I}}^c(\lambda) \cap \mathcal{I}\right) \neq \varnothing$.

One may therefore seek to maximize sparsistency by minimizing the FP and FN probabilities simultaneously, which for the group-LASSO means finding a regularization level which offers a compromise between FPs and FNs. To that end, let $\Lambda^* = [\lambda_{\text{min}}, \lambda_{\text{max}}]$ denote the interval for which any $\lambda \in \Lambda^*$ for the group-LASSO estimator fulfills (13), where

$$\lambda_{\text{min}} = \inf\{\lambda : \max_{i \notin \mathcal{I}} \left\|\mathbf{A}_i^H \mathbf{r}_i\right\|_2 \leq \lambda\} \tag{14}$$

$$\lambda_{\text{max}} = \sup\{\lambda : \min_{i \in \mathcal{I}} \left\|\mathbf{A}_i^H \mathbf{r}_i\right\|_2 \geq \lambda\} \tag{15}$$

Thus, λ_{min} is the smallest λ possible which does not incur FPs, whereas λ_{max} is the largest λ possible without incurring FNs in the solution. Therefore, support recovery is only possible if $\lambda_{\text{min}} \leq \lambda_{\text{max}}$. Clearly, the converse might occur, for instance, if the observations are very noisy, and the largest noise component becomes larger than the smallest signal component, and, as a result $\Lambda^* = \varnothing$.

A. Support recovery as a detection problem

The i:th parameter group is included in the estimated support if the ℓ_2-norm of the inner product between the i:th dictionary group and the residual is larger than the regularization level. The group-LASSO estimate for each group can thus can be seen as a detection problem, with λ acting as the global detection threshold. Support recovery can therefore be seen as a detection test; successful if λ can be selected such that all detection problems (for each and every group) are solved. We therefore begin by examining the statistical properties of the inner product between the i:th dictionary group and the data model. We here assume that the observations consist of a deteministic signal-of-interest and a random noise. Thus,

$$E(\mathbf{y}) = \mathbf{A}\mathbf{x}, \quad V(\mathbf{y}) = E(\mathbf{e}\mathbf{e}^H) = \sigma^2 \boldsymbol{I} \tag{16}$$

The inner product between the dictionary and the data, $\mathbf{A}^H\mathbf{y}$, yields a vector in which each element constitutes a linear combination of the data elements. Under the assumption that $M > N$, the variability

in the data vector is spread among the elements in a larger vector. Let $\mathbf{u} = \mathbf{A}^H\mathbf{y}$ denote this M element vector, which has the statistical properties

$$E(\mathbf{u}) = \mathbf{A}^H\mathbf{A}\mathbf{x}, \quad V(\mathbf{u}) = \sigma^2\mathbf{A}^H\mathbf{A} \tag{17}$$

and while the elements in $\mathbf{y}$ are statistically independent, the elements in $\mathbf{u}$ are generally not, as $\mathbf{A}^H\mathbf{A} \neq \boldsymbol{I}$ for $M > N$. By examining the i:th group in $\mathbf{u}$, one may note that

$$E(\mathbf{u}_i) = \begin{cases} \sum_{j\in\mathcal{I}} \mathbf{A}_i^H\mathbf{A}_j\mathbf{x}_j, & i \notin \mathcal{I} \\ \mathbf{x}_i, & i \in \mathcal{I} \end{cases} \tag{18}$$

where it is also assumed that the components in the true support are independent, $\mathbf{A}_i^H\mathbf{A}_j = \mathbf{0}$, for $i \neq j, (i,j) \in \mathcal{I}$. One may note how the true variables are mixed amongst the elements in $\mathbf{u}$; they appear consistently in the true support, while also leaking into the other variables, proportionally to the coherence between the groups, as quantified by $\mathbf{A}^H\mathbf{A}$. In the CCD updates for the group-LASSO, the i:th component becomes active if

$$\lambda < \left\|\mathbf{A}_i^H\mathbf{r}_i\right\|_2 \tag{19}$$

$$= \left\|\mathbf{A}_i^H\left(\mathbf{A}\mathbf{x} + \mathbf{e} - \mathbf{A}_{-i}\mathbf{x}_{-i}\right)\right\|_2 \tag{20}$$

$$= \left\|\mathbf{A}_i^H\mathbf{A}_i\mathbf{x}_i + \sum_{j\in\mathcal{I}} \mathbf{A}_i^H\mathbf{A}_j(\mathbf{x}_j - \hat{\mathbf{x}}_j) + \mathbf{A}_i^H\mathbf{e}\right\|_2 \tag{21}$$

$$= \begin{cases} \left\|\sum_{j\in\mathcal{I}} \mathbf{A}_i^H\mathbf{A}_j(\mathbf{x}_j - \hat{\mathbf{x}}_j) + \mathbf{A}_i^H\mathbf{e}\right\|_2, & i \notin \mathcal{I} \\ \left\|\mathbf{x}_i + \mathbf{A}_i^H\mathbf{e}\right\|_2, & i \in \mathcal{I} \end{cases} \tag{22}$$

This result provides some insight into choosing the regularization level; this must be set such that with high probability

$$\lambda > \left\|\sum_{j\in\mathcal{I}} \mathbf{A}_i^H\mathbf{A}_j(\mathbf{x}_j - \hat{\mathbf{x}}_j) + \mathbf{A}_i^H\mathbf{e}\right\|_2, \qquad \forall i \notin \mathcal{I} \tag{23}$$

$$\lambda < \left\|\mathbf{x}_i + \mathbf{A}_i^H\mathbf{e}\right\|_2, \qquad \forall i \in \mathcal{I} \tag{24}$$

where if (23) is not fulfilled, FPs enters the solution, whereas if (24) does not hold, FNs will enter the solution. Certainly, the true $\mathbf{x}$ is unknown, as is $\hat{\mathbf{x}}$ before the estimation starts, at which point λ must be selected. If there is coherence in the dictionary, it is not well defined how the data's variability is explained among the dependent variables, due to the bias resulting from the regularizers used in the LASSO-methods. Our proposition is thus to, when selecting the regularization level, focus on the noise part, while leaving the leakage of the $\mathbf{x}_i$:s into the other components be, dealing with them in a later

refinement step. To that end, consider a hypothesis test examining whether the observed data contains the signal-of-interest or not, i.e.,

$$\begin{aligned} \mathcal{H}_0: &\quad \mathbf{y} = \mathbf{e} \\ \mathcal{H}_1: &\quad \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \end{aligned} \tag{25}$$

Under the null hypothesis, $\mathcal{H}_0$, $\mathcal{I} = \varnothing$. In this case, (23) and (24) reduce to

$$\lambda > \left\| \mathbf{A}_i^H \mathbf{e} \right\|_2, \quad \forall i \tag{26}$$

which should be fulfilled with a high probability for all groups. Thus, one may choose the regularization level with regard to the maximum noise component, i.e.,

$$P\left(\max_i \left\| \mathbf{A}_i^H \mathbf{e} \right\|_2 \leq \lambda_\alpha \right) = 1 - \alpha \tag{27}$$

such that λ_α denotes the α-quantile of the maximum ℓ_2-norm of the inner product between the dictionary and the noise. This regularization level can be seen as a lower bound approximation of λ_{min}, where FPs due to leakage from the true support are disregarded.

B. Model selection via extreme value analysis

In order to determine λ_α, we need to find the distribution in (27), which is an extreme value distribution determined by the underlying noise distribution. To that end, let

$$z_i = \left\| \mathbf{A}_i^H \mathbf{e} \right\|_2^2 / \sigma^2 \tag{28}$$

denote the (squared) ℓ_2-norm of the inner product between the i:th dictionary group and the noise, scaled by the noise variance. For the scaled group-LASSO, where $\lambda = \mu\hat{\sigma}$, one may express the sought extreme value distribution, denoted $\bar{F}$, as

$$P\left(\max_i \left\| \mathbf{A}_i^H \mathbf{e} \right\|_2 < \mu\hat{\sigma} \right) = P\left(\max_i \sigma\sqrt{z_i} < \mu\hat{\sigma} \right) \tag{29}$$

$$= P\left(\max_i z_i < \mu^2 \left(\frac{\hat{\sigma}}{\sigma} \right)^2 \right) \tag{30}$$

$$= \bar{F}\left(\mu^2 \left(\frac{\hat{\sigma}}{\sigma} \right)^2 \right) \tag{31}$$

Thus, if $\hat{\sigma}/\sigma \approx 1$, one may seek μ instead of λ, providing a method for finding a regularization level independent of the unknown noise variance. We thus propose selecting μ as the α-quantile from the extreme value distrubution $\bar{F}$ which may be obtained as

$$\mu_\alpha = \frac{\sigma}{\hat{\sigma}} \sqrt{\bar{F}^{-1}(1-\alpha)} \approx \sqrt{\bar{F}^{-1}(1-\alpha)} \tag{32}$$

It is, however, difficult to find closed-form expressions for extremes of dependent sequences; $z_1, \ldots, z_K$ become dependent as the underlying sequence, $\mathbf{u}$ (from (17)) from which the z_i:s are formed, is dependent. As a comparison, let $\tilde{z}_1, \ldots, \tilde{z}_K$ denote a sequence of variables with the same distribution as the z_i:s, although being independent of each other. One may then form the bound

$$P\left(\max_i z_i \leq \mu^2\right) \geq P\left(\max_i \tilde{z}_i \leq \mu^2 \left(\frac{\hat{\sigma}}{\sigma}\right)^2\right) \tag{33}$$

$$= P\left(\tilde{z}_i \leq \mu^2 \left(\frac{\hat{\sigma}}{\sigma}\right)^2\right)^K \tag{34}$$

$$= G\left(\mu^2 \left(\frac{\hat{\sigma}}{\sigma}\right)^2\right)^K \tag{35}$$

$\forall \mu > 0$, where the independence of the parameters was used to form (34). One may thus form an upper bound on the sought quantile as

$$\mu_\alpha \leq \tilde{\mu}_\alpha = \frac{\hat{\sigma}}{\sigma}\sqrt{G^{-1}\left((1-\alpha)^{m^{-1}}\right)} \tag{36}$$

where G is the distribution of z_i, assumed to be equal $\forall i$. Indeed, μ_α is thus constructed such that the null hypothesis, $\mathcal{H}_0$, is falsely rejected with probability α. It does not, however, automatically mean that the probability for FPs, as described in (24), is also α. As argued above, the FP probability is typically larger than α, but, as will be illustrated below, can be shown to yield regularization levels that give high sparsistency.

C. Inference using Monte Carlo sampling

The proposed method for choosing the regularization level, as presented in this paper, only requires knowledge of the distribution family for the noise. We might then sample from the corresponding extreme value distribution, $\bar{F}$, using the Monte Carlo method. Consider $\mathbf{w}^{[j]}$ to be the j:th draw from the noise distribution F which has unit variance. A sample from the sought distribution, $\bar{F}$, is then obtained by calculating

$$\max_i z_i^{[j]} \sim \bar{F}(z) \tag{37}$$

where $z_i^{[j]} = \mathbf{w}^{[j]\,H}\mathbf{A}_i\mathbf{A}_i^H\mathbf{w}^{[j]}$. By randomly drawing N_{sim} such samples from $\bar{F}$, the quantile μ_α may be chosen either using a parametric quantile, or using the empirical distribution function, i.e.,

$$\mu_\alpha = \sqrt{\Psi_{\bar{F}}^{-1}(1-\alpha)} \tag{38}$$

where $\Psi_{\bar{F}}$ is the empirical distribution function of $\bar{F}$. For small α, the empirical approach may be computationally burdensome as

$$\mu_\alpha^2 \leq \max_{j=1,\dots N_{\text{sim}}} \left(\max_i z_i^{[j]} \right) \Rightarrow N_{\text{sim}} \geq \lfloor \alpha^{-1} \rfloor \tag{39}$$

and one might then prefer to use a parametric quantile instead. Luckily, as the noise distribution F is assumed to be known, or may be estimated using standard methods, it is often feasible to derive which distribution family $\bar{F}$ belongs to. By then estimating the parameters of the distribution using the gathered Monte Carlo samples, a parametric quantile μ_α may be obtained using much fewer samples than using the corresponding empirical quantiles.

D. The Gaussian noise case

A common assumption is to model the noise as a zero-mean circular-symmetric i.i.d. complex-valued Gaussian process with some unknown variance, σ^2, i.e., $\mathbf{e} = \sigma \mathbf{w}$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. For the i:th group, one then obtains

$$\mathbf{A}_i^H \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \Rightarrow \quad z_i \sim \chi^2(2L_i) \tag{40}$$

as it is assumed that $\mathbf{A}_i^H \mathbf{A}_i = \boldsymbol{I}, \forall i$. Thus, as z_i is a sum of L_i independent squared $\mathcal{N}(0,1)$ variables, it becomes χ^2 distributed with $2L_i$ degrees of freedom. In such a case, one may use (36) to directly find a closed-form upper bound on the regularization parameter. Alternatively, to find a more accurate quantile, one may draw inference on the Monte Carlo samples obtained when sampling from $\bar{F}$ in (37) instead. Classical extreme value theory states that the maximum domain of attraction for the Gamma distribution (of which χ^2 is a special case) is the type I extreme value distribution, i.e., the Gumbel distribution [38]. By estimating the scale and location parameters of the Gumbel distribution, one may obtain a more accurate tail estimate from the z_i:s than the empirical distribution yields. Thus, it holds that

$$\max_i z_i \sim \bar{F}(z) = \exp\left(e^{-\frac{z-\gamma}{\beta}} \right) \tag{41}$$

where the parameters γ and β are obtained using maximum likelihood estimation on the samples $\max_i z_i^{[1]}, \dots, \max_i z_i^{[N_{\text{sim}}]}$. The regularization parameter μ_α can then be calculated using (32).

V. Correcting the σ-estimate for the scaled group-LASSO

The scaled LASSO framework provides a way of choosing the regularization level independently of the noise variance. By introducing σ as an auxiliary variable in the LASSO minimization objective, it may be estimated along with $\mathbf{x}$. In the CCD solver, the estimate of the noise standard deviation is obtained

in closed-form, from (10), as the residual standard deviation estimate, i.e., $\hat{\sigma}(\mu) = \|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}(\mu)\|_2/\sqrt{N}$. There are two aspects in how well $\hat{\sigma}$ approximates the true noise standard deviation; firstly, as $\hat{\sigma}(\mu)$ models the residual, it will depend on the sparsity level of $\hat{\mathbf{x}}(\mu)$, such that

$$\hat{\sigma}(\mu) \to \sqrt{\mathbf{y}^H\mathbf{y}/N}, \quad \text{as } \mu \to \mu_0 \tag{42}$$

$$\hat{\sigma}(\mu) \to 0, \quad \text{as } \mu \to 0 \tag{43}$$

where μ_0 is the smallest μ which yields the zeroth solution, i.e.,

$$\mu_0 = \frac{\max\limits_i \left\|\mathbf{A}_i^H\mathbf{y}\right\|_2}{\sqrt{\mathbf{y}^H\mathbf{y}/N}} \tag{44}$$

Therefore, if μ is chosen too large, such that it underestimates the model order, $\hat{\sigma}$ is overestimated, whereas if μ is chosen too small, and too many components are included in the model, $\hat{\sigma}$ becomes underestimated. The second aspect is that the estimate models the residual standard deviation for the LASSO estimator, where the magnitudes of the elements in $\mathbf{x}$ are always biased towards zero, and will thus overestimate $\hat{\sigma}$ even when the regularization level is selected such that the true support is obtained, i.e.,

$$\hat{\sigma}(\mu) = \frac{\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}(\mu)\|}{N} \geq \frac{\|\mathbf{y} - \mathbf{A}\mathbf{x}\|}{N} = \sigma, \quad \mu \in M^* \tag{45}$$

where M^* is the interval over μ which yields the true support estimate. These aspects have a profound effect on the regularization level. As a result of the approximation in (32), the chosen α will not yield the actual FP rate of the hypothesis test under $\mathcal{H}_0$; let the true FP rate be denoted by α^*. The relation between the chosen quantile μ_α and the true quantile μ_{α^*} is then

$$\mu_\alpha = \sqrt{\bar{F}^{-1}(1-\alpha)} = \frac{\hat{\sigma}(\mu_\alpha)}{\sigma}\sqrt{\bar{F}^{-1}(1-\alpha^*)} = \frac{\hat{\sigma}(\mu_\alpha)}{\sigma}\mu_{\alpha^*} \tag{46}$$

and subsequently the true FP rate becomes

$$\alpha^* = 1 - \bar{F}\left(\left(\frac{\sigma}{\hat{\sigma}(\mu_\alpha)}\right)^2 \bar{F}^{-1}(1-\alpha)\right) \tag{47}$$

One may therefore deduce that when $\hat{\sigma}$ is over- or underestimated, the FP rate becomes over- or underestimated, respectively; i.e.,

$$\sigma(\mu_\alpha) > \sigma \Rightarrow \alpha^* > \alpha \tag{48}$$

$$\sigma(\mu_\alpha) < \sigma \Rightarrow \alpha^* < \alpha \tag{49}$$

while if the standard deviation is correctly estimated, $\alpha^* = \alpha$. This may be attempted by selecting α small, such that the model order reasonably reflects the true model order, and then estimate the noise

standard deviation using an unbiased method instead of via the LASSO. One may then undertake the following steps to improve the estimate of the noise standard deviation:

1) Estimate $\mathbf{x}$ and σ by solving the scaled group-LASSO problem (9) with regularization parameter μ_α, given by (32) for some α.
2) Re-estimate σ using a least squares estimate of the non-zero variables obtained in Step 1), $x_i \in \hat{\mathcal{I}}$, i.e., $\hat{\sigma}_{\text{LS}} = \left\| \left(\boldsymbol{I} - \mathbf{A}_{\hat{\mathcal{I}}} \mathbf{A}_{\hat{\mathcal{I}}}^{\dagger} \right) \mathbf{y} \right\|_2 / \sqrt{N}$.
3) Estimate $\mathbf{x}$ by solving the (regular) group-LASSO problem in (5) with the regularization parameter selected as $\lambda = \mu_\alpha \hat{\sigma}_{\text{LS}}$.

VI. Marginalizing the effect of coherence-based leakage

The proposed method calculates a regularization level by quantifying the FP error probability for the hypothesis testing of whether the noisy data observations also contain the signal-of-interest, $\mathbf{Ax}$, or not. This FP rate is used to approximate the FP rate for finding the correct support, which is a slightly different quantity. The regularization level is set by analyzing how the noise propagates into the estimate of $\mathbf{x}$, and selects a level larger than the magnitude of the maximum noise component. In relation to the hypothesis test in (25), when the signal-of-interest is present in the signal, the group-LASSO estimate suffers from spurious non-zero estimates outside of the support set, as described in (23). Thus, even if the choice of μ_α drowns out the noise part with probability $1-\alpha$, it does not necessarily zero out the spurious signal components, if the dictionary coherence is non-negligible. The true support is thereby not recovered, and the sparsistency rate is lower than $1-\alpha$.

One remedy is to set the regularization level higher, but it is inherently difficult to quantify how the variability of the signal component is divided among its coherent dictionary atoms with LASSO-like estimators, and therefore difficult to assess how much higher it should be selected. For low SNR observations, the choice of the regularization level is sensitive; if set too high, the estimate will suffer from FNs. We suggest to keep the regularization level as the proposed quantile μ_α, but instead modify the sparse regression, as to promote more sparsity among coherent components, thereby possibly increasing the sparsistency rate. One such method is to solve the reweighted group-LASSO problem, where one at the j:th iteration obtains $\hat{\mathbf{x}}^{(j)}$ by solving

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{Ax}\|_2^2 + \lambda \sum_{k=1}^{K} \frac{\|\mathbf{x}_k\|_2}{\left\|\hat{\mathbf{x}}_k^{(j-1)}\right\|_2 + \epsilon} \tag{50}$$

where ϵ is a small positive constant used to avoid numerical instability. Thus, the regularization level is iteratively updated using the ℓ_2-norm of the old estimate, which has the effect that the individual

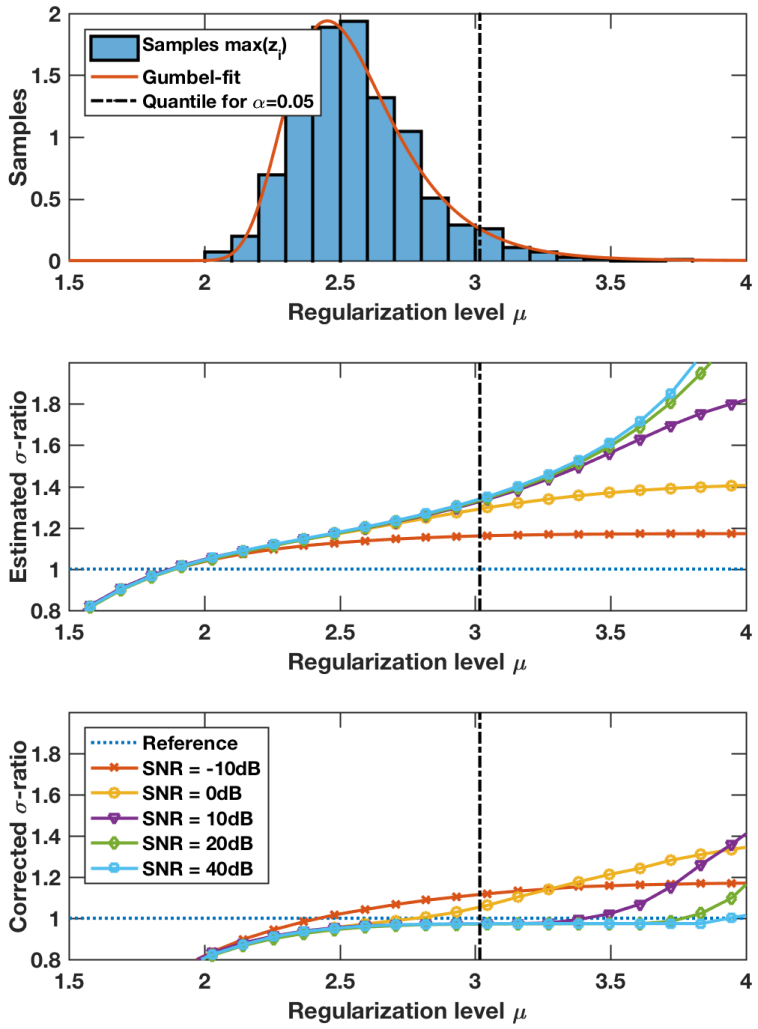

Fig. 1. Results for estimation of σ for different levels of the regularization level, μ; the top plot illustrates how the ℓ_2-norm of the maximum nuisance component is distributed, the middle and bottom plots illustrate the ratio between the estimated $\hat{\sigma}$ and the true σ, for different levels of regularizations, using the scaled LASSO estimator. The different curves show the ratio estimates for different levels of SNR, i.e., σ. In the bottom plot, the σ-correction step has been applied to the estimation.

regularizer

$$\lambda_k = \frac{\lambda}{\left\|\mathbf{x}_k^{(j-1)}\right\|_2 + \epsilon} \searrow, \quad \|\hat{\mathbf{x}}_k\|_2 > 1 \quad \text{(is large)} \tag{51}$$

$$\lambda_k = \frac{\lambda}{\left\|\mathbf{x}_k^{(j-1)}\right\|_2 + \epsilon} \nearrow, \quad \|\mathbf{x}_k\|_2 < 1 \quad \text{(is small)} \tag{52}$$

Thus, the best (largest) component among the coherent variables will be less and less regularized, while the weaker components will be more and more regularized, until they are omitted altogether. By solving (50) iteratively, the solver approximates a (non-convex) sparse regression problem with a logaritmic regularizer, which is more sparsifying than the ℓ_1-regularizer. We thus propose to modify Step 3) in the σ-corrected approach, discussed above, with the reweighted group-LASSO, using $\lambda = \mu_\alpha \hat{\sigma}_{\text{LS}}$.

VII. In comparison: Data-driven hyperparameter selection

Commonly, the data-driven approach to finding the regularization level is to solve the LASSO problem for a grid $\lambda \in \Lambda$, typically selected as N_λ points uniformly chosen on $(0,\, \lambda_0]$. Commonly, $\mathbf{x}(\Lambda)$ is then referred to as the solution path. For each point, λ_j, on the solution path, one can obtain the model order, $\hat{k}_j = \|\hat{\mathbf{x}}(\lambda_j)\|_0$ and the statistical likelihood of the observed data given the assumed distribution of the parameter estimate, $L(\mathbf{y}, \hat{\mathbf{x}}(\lambda_j))$, which when used to calculate the Bayesian Information Criteria (BIC), i.e.,

$$\text{BIC}(\lambda_j) = \log(N)\hat{k}(\lambda_j) - 2L(\mathbf{y}, \hat{\mathbf{x}}(\lambda_j)) \tag{53}$$

yields the model order estimate $\hat{\lambda}_{\text{BIC}} = \text{argmin}_j \, \text{BIC}(\lambda_j)$. Certainly, this procedure may prove costly, as it requires solving the LASSO N_λ times. Typically, BIC also tends to overestimate the model order, thereby underestimating $\hat{\lambda}_{\text{BIC}}$. Another commonly used method for selecting the regularization level is to perform cross-validation (CV) on the observed data. As there exist many different variations of CV, many of which are computationally infeasible for typical problems, this paper describes the popular R-fold CV variant [39], in which one shall:

1) Split the observed data into R disjoint random partitions.
2) Omit the r:th partition from estimation and use the remaining partitions to estimate a solution path $\mathbf{x}_r(\Lambda)$. Repeat for all partitions.
3) Calculate the prediction residual variance $\mathbf{r}(\lambda, r) = \|\mathbf{y}(r) - \mathbf{A}(r)\mathbf{x}_r(\lambda_i)\|_2^2$ and use this to calculate

$$\text{CV}(\lambda_j) = \sum_{r=1}^{R} \mathbf{r}(\lambda_i, r) n^{-1} \tag{54}$$

$$\text{SE}(\lambda_j) = D\left(\{\mathbf{r}(\lambda_j, r)\}_{r=1}^{R}\right) R^{-1/2} \tag{55}$$

4) Let $\lambda^* = \text{argmin}_j\ CV(\lambda_j)$. Utilizing the one standard error rule, one then finds $\lambda_{\text{CV}} = \sup_j \lambda_j$ such that $\text{CV}(\lambda_j) \leq \text{CV}(\lambda^*) + \text{SE}(\lambda^*)$.
5) Calculate the solution $\hat{\mathbf{x}}(\lambda_{\text{CV}})$ using the entire data set.

When R is large enough, CV has been shown to asymptotically approximate the Akaike Information Criterion (AIC) [40]. However, CV is generally computationally burdensome, requiring solving the LASSO $(N_\lambda R + 1)$ times. It should be noted that CV forms the model order estimate by selecting the λ which yields the smallest prediction error. This undoubtedly discourages overfitting, but does not specifically target support recovery, which often is the main objective of sparse estimation. Thus, CV tends to set λ too low, which reduces the bias for the correct variables of $\hat{\mathbf{x}}(\lambda)$, but which also introduces FPs.

VIII. Probabilistic hyperparameter selection

The probabilistic method by Chen et al. [27] is based on the assumption of independence; a result from [41], utilizing an extreme-value result from [42], states that for an i.i.d. Gaussian sequence of variables $\tilde{z}_1, \ldots, \tilde{z}_K$, i.e., $\tilde{z}_k \sim \mathcal{N}(0, \sigma^2)$,

$$P\left(\max_k |\tilde{z}_k| \leq \sigma\sqrt{2\log(K)}\right) \to 1, \quad \text{as } K \to \infty \tag{56}$$

which is an endpoint-of-distribution type result, thus stating that for large orthogonal dictionaries (i.e., necessitating large data sets), the maximum noise component grows proportional to $\sigma\sqrt{2\log(K)}$. The authors therefore propose that the regularization level for the LASSO should be selected as

$$\lambda = c\,\sigma\sqrt{2\log(K)} \tag{57}$$

where $c \geq 1$ is a non-critical user-defined parameter to make noise suppression even harsher. For example, for student t-distributed variables, one might use c to compensate for the heavier-than-normal tails. One may thus argue that the CDS method is a blind approach of selecting the regularization level under a Gaussian and orthogonal assumption. However, a drawback of the method is that, as shown in the following, the method guarantees noise suppression to the detriment of likely incurring FNs for both low and medium noise levels.

IX. Numerical results

To illustrate the efficacy of the proposed method, termed the PRObabilistic regularization approach for SParse Regression (PROSPR) for selecting a regularization level, we test it under a few test scenarios, in

comparison to the BIC, CV, and CDS methods. However, before doing so, we illustrate the distribution of the maximum noise component over μ (from (32)), and make analysis on how σ is estimated in the scaled LASSO for these levels of the regularization parameter, with and without the σ-correction step described in Section V. We thus simulate $N_{\text{MC}} = 200$ Monte Carlo simulations of $\mathbf{y}$, such that

$$\mathbf{y}^{[n]} = \mathbf{A}^{[n]}\mathbf{x}^{[n]} + \sigma\mathbf{w}^{[n]} \tag{58}$$

for the n:th simulation, where the elements in the dictionary consist of i.i.d. draws from the complex-valued Gaussian distribution, $\mathcal{N}(0,1)$, and wherein $\mathbf{x}$ are $S = 5$ non-zero elements with unit magnitude, at randomly selected indices. In this test scenario, we consider the standard (non-grouped) regression problem and normalize the columns of the dictionary. In each simulation, $N = 100$ data samples are retrieved, and the number of regressors is set to $M = 500$, and thus equally many groups, $K = 500$, such that $L = 1$. The signal-of-interest is here corrupted by an i.i.d. complex-valued Gaussian noise, such that $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Figure 1 illustrates the distribution of $z_i^{[n]} = \max_i \left|(\mathbf{a}_i^{[n]})^H \mathbf{w}^{[n]}\right|^2$, for $n = 1, \ldots, N_{\text{sim}}$, in the top figure, where the density function for a fitted Gumbel distribution is overlaid. The dash-dotted line illustrates the quantile value for $\alpha = 0.05$, which thus corresponds to the regularization level used with the proposed method for that α. Here, one may in comparison note that SPICE, which is shown in, e.g, [31], to have the regularization level fixed at $\mu = 1$, sets the regularization level much too low and thus very likely incurs FPs into the solution. In comparison, the CDS method, which given the correct standard deviation corresponds to $\mu = \sqrt{2\log(K)} \approx 3.53$, which is in the very tail of the estimated distribution, might also suppress the signal-of-interest. The middle plot illustrates the paths of the ratios $\hat{\sigma}(\mu)/\sigma$, when estimated using the scaled LASSO, wherein each of the four lines illustrates the estimated ratio when the true σ used in (58) is selected such that the signal-to-noise ratio (SNR) is $-10, 0, 10, 20$, and 40 dB, respectively. Depending on μ, the LASSO estimate will include either only noise, or both noise and the signal-of-interest, and the ratios thus grow as $\mu \to \mu_0$. When the SNR is low, the ratio contains much of the signal-of-interest even at a low regularization level, while, as the SNR is low, the signal-of-interest is weak. Also, one may note that the ratios level out as $\mu \to \mu_0$, at which point $\hat{\mathbf{x}} = \mathbf{0}$. The choice of α, if selected too small, in tandem with a low SNR, will therefore yield the zero solution. One sees this in the middle figure, as the ratio has leveled out for SNR = -10 dB, whereas the ratios are still growing for the other noise levels. Most important, however, is how the ratios affect the choice of μ in (32). As the method assumes that the ratio is close to one, the effect when it grows becomes, as given by (48), that the actual α becomes larger than the selected value, which decreases sparsistency as FPs enter the solution. A remedy to this may be seen in the bottom figure, where σ is re-estimated

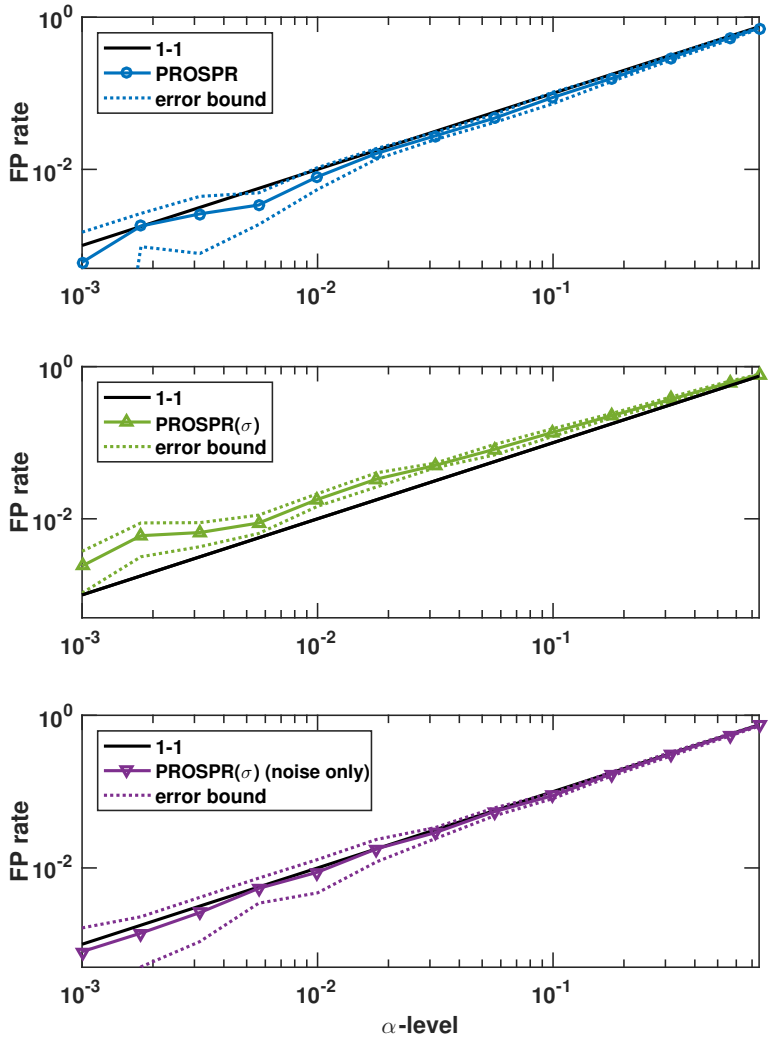

Fig. 2. The subfigures illustrate the FP rate of support recovery for different levels of α. For all figures, the filled curves shows the preferred 1-1 line, the filed lines with symbols shows the estimate, and the dashed line shows the estimates $\pm$ one standard error. The top figure illustrate the FP rates for the PROSPR method, in the middle figure PROSPR with σ-correction is used, and the bottom figure shows the FP rate when the data is noise-only.

along the lines described in Section V. By using the least squares estimate, the ratio becomes larger than one only if the components in the true support has already been excluded from the estimated support, which can be seen for SNR = -10 and 0 dB. It may be noted that for the other SNR levels, the ratio is approximately one in the upper tail of the Gumbel distribution, and α approximates the true FP rate for inclusion of components due to noise.

As noted above, this is not necessarily equal to the FP rate of support recovery, as is illustrated in Figure 2, where the estimated FP rate for support recovery,

$$\frac{1}{N_{\text{MC}}} \sum_{n=1}^{N_{\text{MC}}} 1 \left\{ \left(\hat{\mathcal{I}} \cap \mathcal{I}^c \right) \neq \varnothing \right\} \tag{59}$$

is shown, where $1\{\cdot\}$ denotes the binary indicator function, which is one if the specified condition is fullfilled (and zero otherwise); the condition being that there exists elements in estimated support which are not in the true support.

The middle plot illustrates the obtained FP rate for different choices of α, when σ-corrected PROSPR is used. The filled line with upwards pointing triangles shows the mean value, and the dashed lines shows the mean value $\pm$ one standard deviation. In this scenario, we have used the same parameter settings as above, although with $N_{\text{MC}} = 5000$ Monte Carlo simulations, at SNR = 20 dB. To select the regularization level in each simulation, we let PROSPR use $N_{\text{sim}} = 500$ simulations of the noise, $\mathbf{w}^{[j]}$, and use a parametric quantile from a Gumbel distribution fitted to the obtained draws of $\max_i z_i^{[j]}$. One note from the simulation results that in the middle figure, the estimated FP rate is consistenly higher than the selected α. This is a result of the dictionary coherence, which makes the signal power in the true support leak into the other variables, as shown in (22). To verify this claim, the bottom figure shows the same estimation scenario, when applied to the noise-only signal, i.e., $\mathbf{y} = \sigma\mathbf{w}$, where thus $\mathcal{I} = \varnothing$, and FPs occur whenever $\hat{\mathbf{x}}(\mu_\alpha) \neq \mathbf{0}$. As may be seen in the figure, the FP rate follows the α-level well for this scenario. To remedy the overestimated FP rate, one should set the regularization level somewhat higher, in order to account for the signal leakage, too. Generally, we have found that when σ-correction is not used, and the estimated σ is too large, as discussed above, this results in a regularization level that is set too high, i.e.,

$$\lambda_\alpha = \mu_\alpha \hat{\sigma}(\mu_\alpha) > \mu_\alpha \sigma = \lambda_{\alpha^*} \tag{60}$$

which may cancel out the unwanted effect of having a too large FP rate. For this estimation scenario, this indeed becomes the case, as can be seen in the top figure, where the FP rate follows the specified α-level well. Next, we compare the proposed method, with and without σ-correction, to the BIC, CV,

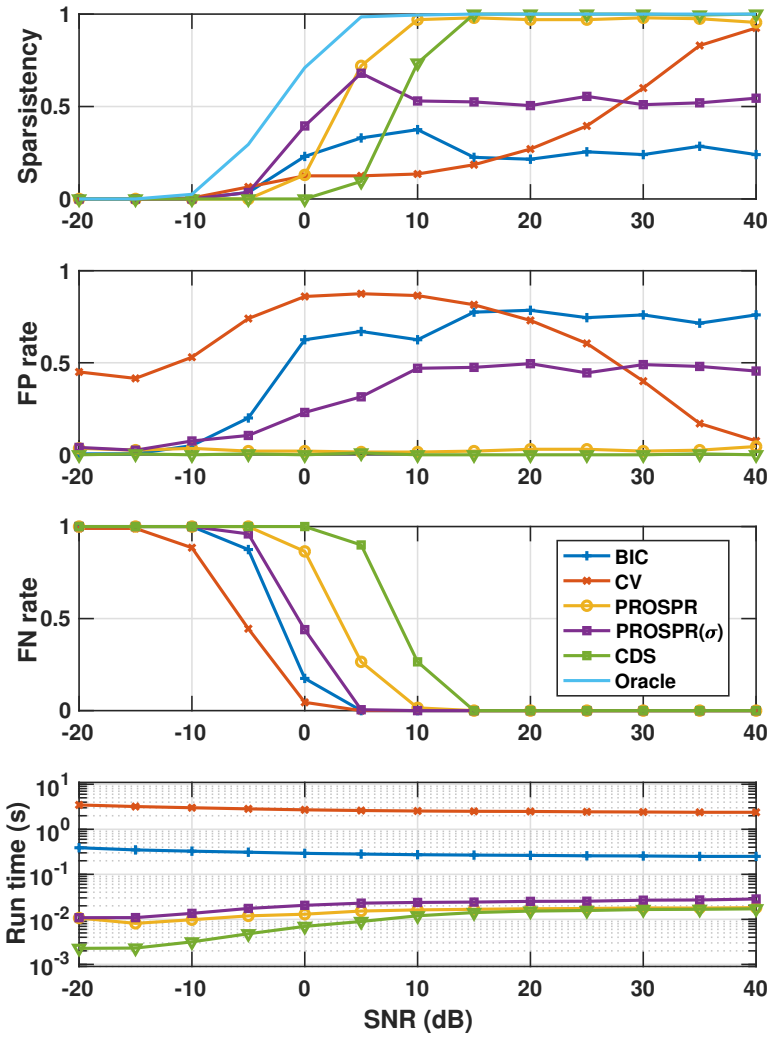

Fig. 3. Compared performance results with LASSO using PROSPR (with and without σ-correction), CV, BIC, CDS, and in the top plot, an oracle method, illustrating the best result achievable at any regularization level. The top plot shows sparsistency (or support recovery rate), the second plot shows the FP rate, the third shows the FN rate, and the bottom plot shows the average run times for each method.

and CDS methods for hyperparameter selection, in terms of sparsistency

$$\frac{1}{N_{\text{MC}}} \sum_{n=1}^{N_{\text{MC}}} 1\left\{\hat{\mathcal{I}} = \mathcal{I}\right\} \tag{61}$$

FP rate from (59), FN rate

$$\frac{1}{N_{\text{MC}}} \sum_{n=1}^{N_{\text{MC}}} 1\left\{\left(\hat{\mathcal{I}}^c \cap \mathcal{I}\right) \neq \varnothing\right\} \tag{62}$$

and average run time in seconds, when implemented in Matlab using the CCD solver on a 2013 Intel Core i7 MacBook Pro, for $N_{\text{MC}} = 200$ simulations. In the top plot in Figure 3, illustrating the sparsistency results at different levels of SNR, we have also included the oracle support recovery, which illustrates the maximum rate of support recovery achievable using an oracle choice of λ. For CV and BIC, the LASSO is solved on a grid of $N_\lambda = 50$ regularization levels, uniformly spaced on $(0,\, \lambda_0]$, and $R = 10$ folds were used for CV. For CDS, we use $c = 1$ and select $\mu = \sqrt{2\log(K)}$ for the scaled LASSO, thereby also obtaining an estimate of the standard deviation, thus equivalent to a regularization level $\lambda = \hat{\sigma}\sqrt{2\log(K)}$. For the PROSPR methods, $\alpha = 0.05$ was used. From the FP and FN results, one sees the trade-off between FPs and FNs, such that, on average, CV is the approach which selects the lowest regularization level, benefitting the FN rate, but at the cost of often incurring FPs. CDS on the other hand, selects the highest regularization level, which results in close to zero FPs, while incurring more FNs. PROSPR selects a middle group, resulting in approximately 5 % FPs, but with fewer FNs than CDS. However, if sparsistency is the focus of the regularization, the proposed methods fairs the best, outperforming CV, BIC, and CDS. An advantage with CV is that it chooses the regularization level with respect to both the signal and the noise components, and thus improves as SNR increases, whereas PROSPR yields similar FP rates independently of the SNR. One may therefore, if the SNR is high, choose an α smaller than $\alpha = 0.05$, approaching CDS. As also verified in Figure 2, the FP-rate for the σ-corrected estimate becomes smaller than α; here at most 0.5, whereas PROSPR without σ-correction performs best from SNR = 5 dB and higher. Concerning run times; CV should be at most $N_\lambda R = 2.5 \cdot 10^2$ times slower than the proposed method - a gap which is slightly narrowed as CV uses warm-starts and as PROSPR's regularization level still requires some computational effort. Still, the PROSPR methods are significantly faster than CV, and only negligibly slower than CDS. By comparison, BIC seems to fair somewhere in between; it is faster than CV, but also performing worse than CV for high levels of SNR.

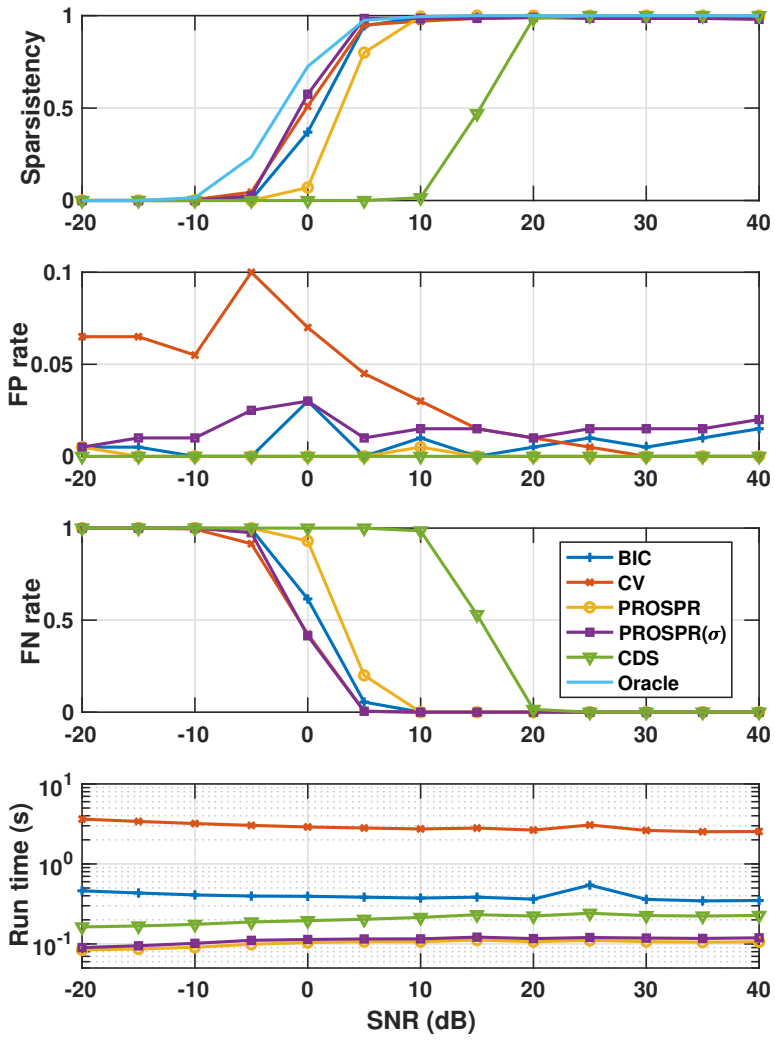

Fig. 4. Compared performance results with reweighted LASSO using PROSPR (with and without σ-correction), CV, BIC, CDS, and in the top plot, an oracle method, illustrating the best result achievable at any regularization level. The top plot shows sparsistency (or support recovery rate), the second plot shows the FP rate, the third shows the FN rate, and the bottom plot shows the average run times for each method.

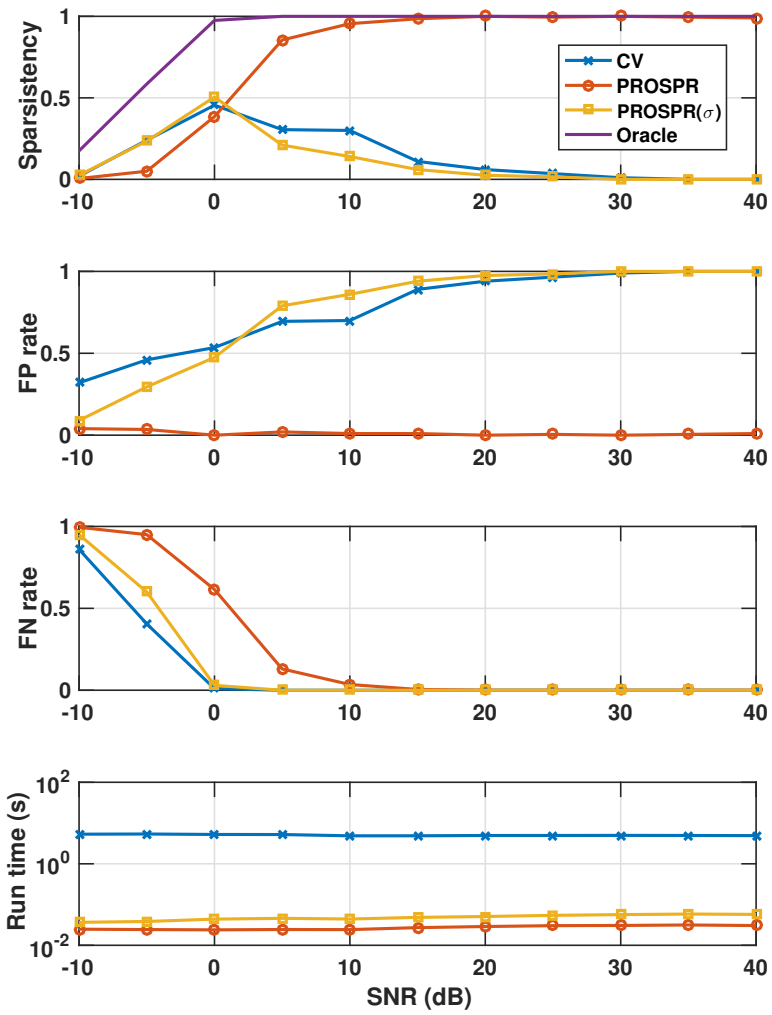

Fig. 5. Compared performance results with group-LASSO using PROSPR (with and without σ-correction), CV, and in the top plot, an oracle method, illustrating the best result achievable at any regularization level. The top plot shows sparsistency (or support recovery rate), the second plot shows the FP rate, the third shows the FN rate, and the bottom plot shows the average run times for each method.

As discussed in Section VI, the effect of coherence-based leakage from the signal components may be lessened by using a reweighted LASSO, where the (group-) LASSO problems is solved several times, with the regularization level being individually and iteratively selected for each group using the old estimate. The approach approximates a non-convex logarithmic penalty, which is sparser than the convex ℓ_1 or ℓ_2/ℓ_1 regularizers for the LASSO and group-LASSO, respectively. Typically, the reweighted (group-) LASSO handles FPs very well, pushing these towards zero, while FNs remains unchanged. As seen from Figure 3, the main error incurred in the PROSPR methods is FNs, where instead of using $\alpha = 0.05$ (i.e., using a low probability of FP), one might select a larger quantile, such that the FN rate decreases, at the expense of a higher FP rate. Figure 4 illustrates the estimation performance when the reweighted LASSO has been used at the regularization levels by the methods, where $\alpha = 0.5$ is used for the PROSPR-methods. One may note at this level, PROSPR with σ-correction follows CV well, with the FPs being dealt to a large extent. Although performing similarly to CV in terms of sparsistency, the proposed method still has a substantial computational advantage. As for CDS, the technique will by construction work poorly for the reweighted LASSO. This is due to the fact that reweighting suppressess estimates close to zero, thus reducing the FPs, while FNs, i.e., signal components which have been set to zero, cannot be recovered. The reweighting is likewise the reason why the FP rate for PROSPR is much smaller than $\alpha = 0.5$, as the small noise components have been supressed.

Next, we then analyze estimation performance for the group-sparse regression problem. We simulate $N_{\text{MC}} = 200$ Monte Carlo simulations of the signal in (58), with $N = 100$ observations in each, using a dictionary with $M = 1000$ atoms collected into $K = 200$ groups with $L = 5$ atoms in each. The true signal-of-interest consists of $S = 3$ groups, with indices randomly selected, and where $\mathbf{x}_i = \mathbf{1}$, for $i \in \mathcal{I}$.

Otherwise, the settings are identical to the standard sparse regression setup. Figure 5 shows the estimation performance for this simulation scenario, for different levels of SNR. One may note that CV does not perform as well as for the standard sparse regression model, as the FP rate does not decrease when SNR increase. As before, PROSPR performs better without σ-correction than with, for high levels of SNR. Unlike before, the BIC criterion for groups is not straightforward, as the degrees of freedom in the estimation is not well-defined; we have therefore decided to omit it from comparison in the group-sparse regression scenario. To the best of the authors' knowledge, no group-version of CDS has been developed.

Finally, similar to Figure 4, Figure 6 compares the estimation results for the reweighted group-LASSO estimator for the compared methods, for different levels of SNR. One may note that the proposed method with σ-correction now outperforms CV, approaching the oracle performance, while the proposed method

without correction performs on par with CV. It therefore seems that CV, for the group-sparse problem, sets the regularization level relatively higher than for the non-grouped case. Still selecting $\alpha = 0.5$ heuristically seems to be good for the reweighted approach; it is set low enough to avoid FNs, while the reweighting manages to lessen the FPs. Again, the computational complexity can be seen to be significantly lower than for CV.

X. Conclusions

This paper has studied the selection of regularization level for sparse and group-sparse regression problems. As an implicit model order selection, it has a profound effect on support recovery; by changing the regularization level, one obtains supports with sizes ranging from very dense to completely empty. If support recovery is the main objective, selecting the regularization level carefully is of utmost importance. The group-regression problem, includes or excludes components from the estimated support depending on how the ℓ_2-norm of the inner-product between the dictionary group and the modeling residual compares to the regularization level. Intuitively, one therefore wishes to select the regularization level larger than the noise components, as to exclude them, but smaller than the signal components, as to include these. As we propose a probabilistic approach, where the regularization level is selected prior to estimation and the signal components are still unknown, we have instead studied the effect of the unit-variance observation noise, and how it propagates into the parameter estimates. Via extreme value analysis and by virtue of Monte Carlo simulations, we sample from the distribution of the maximal noise component, and may therefore select the regularization level as a quantile from the distribution. With the implicit assumption that the signal components are larger than the noise components, the quantile level, if chosen too large, may incur FNs, and if set too small, may incur FPs.

The proposed method is thus not hyperparameter-free, and in some sense merely replaces one hyperparameter by another. However, the sparse regression model does not contain enough information to be hyperparameter-free on its own; other methods for hyperparameter-selection will also require assumptions on the model, e.g., CV assumes that the optimal regularization level is the one yielding the smallest prediction error. Similarly, SPICE simply selects $\mu = 1$ and we have shown that CDS, which selects the endpoint-of-distribution for the maximum noise component while assuming an orthonormal dictionary and Gaussian noise, is too strongly regularized. In this work, we have shown that by selecting $0 < \alpha < 1$ in μ_α, one approximately select the FP rate for support recovery. If set too generous, FPs are likely but for low SNRs, FNs become less likely, and conversely, if set too small, the solution is likely to be sparse, but might omit parts of the sought support. We argue that α is relatively easy to set heuristically,

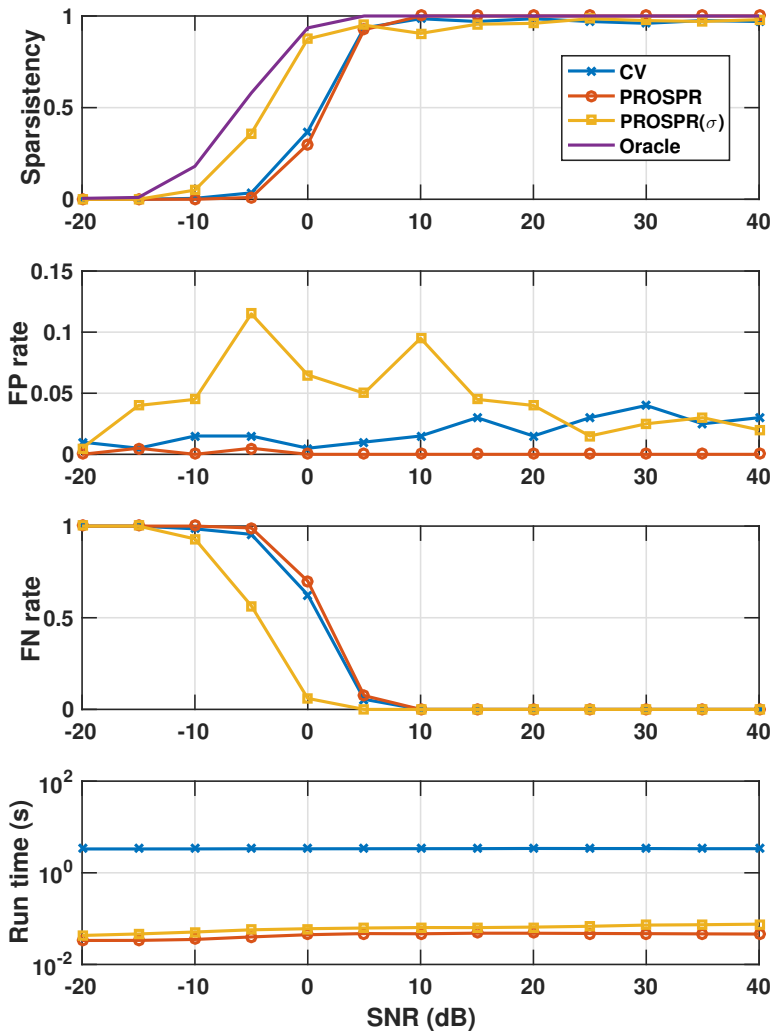

Fig. 6. Compared performance results with reweighted group-LASSO using PROSPR (with and without σ-correction), CV, and in the top plot, an oracle method, illustrating the best result achievable at any regularization level. The top plot shows sparsistency (or support recovery rate), the second plot shows the FP rate, the third shows the FN rate, and the bottom plot shows the average run times for each method.

whereas the regularization level, λ, is much more difficult to set appropriately. We have also shown that the median quantile, i.e., $\alpha = 0.5$, will approximate the CV's regularization level, which, when used for the reweighted LASSO problem, gives high rates of support recovery.

The great virtue of the probabilistic methods, including the proposed method, lies in the computational complexity; CV is often computationally burdensome, even infeasible for some applications, solving the LASSO problem again and again for different regularization levels, while the proposed method is independent of the examined data. It only requires knowing the approximate shape of the noise distribution. This may often be found using secondary noise-only data, or using some standard estimation procedure. Furthermore, the family of the noise distribution is not required to be specifically known, as it may suffice to draw samples from its empirical distribution function.

XI. Acknowledgement

We would like to express our gratitude toward Associate Professor Johan Lindström and Professor Emeritus Georg Lindgren, both at the department of Mathematical Statistics, Lund University, for sharing their insights and wisdom with us during the research work for this project.

References

[1] J. J. Fuchs, "On the Use of Sparse Representations in the Identification of Line Spectra," in *17th World Congress IFAC*, Seoul, July 2008, pp. 10 225–10 229.

[2] S. Bourguignon, H. Carfantan, and J. Idier, "A sparsity-based method for the estimation of spectral lines from irregularly sampled data," *IEEE Journal of Selected Topics in Signal Processing*, December 2007.

[3] J. Fang, F. Wang, Y. Shen, H. Li, and R. S. Blum, "Super-Resolution Compressed Sensing for Line Spectral Estimation: An Iterative Reweighted Approach," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4649–4662, September 2016.

[4] I. F. Gorodnitsky and B. D. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, March 1997.

[5] D. Malioutov, M. Cetin, and A. S. Willsky, "A Sparse Signal Reconstruction Perspective for Source Localization With Sensor Arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, August 2005.

[6] S. I. Adalbjörnsson, T. Kronvall, S. Burgess, K. Åström, and A. Jakobsson, "Sparse Localization of Harmonic Audio Sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 117–129, January 2016.

[7] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-Pitch Estimation Exploiting Block Sparsity," *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[8] T. Kronvall, M. Juhlin, S. I. Adalbjörnsson, and A. Jakobsson, "Sparse Chroma Estimation for Harmonic Audio," in *40th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Brisbane, Apr. 19-24 2015.

[9] F. Elvander, T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "An Adaptive Penalty Multi-Pitch Estimator with Self-Regularization," *Elsevier Signal Processing*, vol. 127, pp. 56–70, October 2016.

[10] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society B*, vol. 67, no. 1, pp. 91–108, January 2005.

[11] A. S. Stern, D. L. Donoho, and J. C. Hoch, "NMR data processing using iterative thresholding and minimum l1-norm reconstruction," *J. Magn. Reson.*, vol. 188, no. 2, pp. 295–300, 2007.

[12] J. Swärd, S. I. Adalbjörnsson, and A. Jakobsson, "High Resolution Sparse Estimation of Exponentially Decaying N-dimensional Signals," *Elsevier Signal Processing*, vol. 128, pp. 309–317, Nov 2016.

[13] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[14] D. Donoho, M. Elad, and V. Temlyakov, "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan 2006.

[15] J. Fan and R. Li, "Variable selection via non-concave penalized likelihood and its oracle properties," *Journal of the Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[16] M. Yuan and Y. Lin, "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006. [Online]. Available: http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[18] E. J. Candès, J. Romberg, and T. Tao, "Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[19] Y. V. Eldar, P. Kuppinger, and H. Bolcskei, "Block-Sparse Signals: Uncertainty Relations and Efficient Recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[20] I. CVX Research, "CVX: Matlab Software for Disciplined Convex Programming, version 2.0 beta," `http://cvxr.com/cvx`, Sep. 2012.

[21] H. Leeb and B. M. Pötscher, "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, vol. 21, no. 1, pp. 21–59, 2005. [Online]. Available: http://www.jstor.org/stable/3533623

[22] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, April 2004.

[23] T. Kronvall, F. Elvander, S. Adalbjörnsson, and A. Jakobsson, "Multi-Pitch Estimation via Fast Group Sparse Learning," in *24rd European Signal Processing Conference*, Budapest, Hungary, 2016.

[24] C. D. Austin, R. L. Moses, J. N. Ash, and E. Ertin, "On the Relation Between Sparse Reconstruction and Parameter Estimation With Model Order Selection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 560–570, 2010.

[25] R. Lockhart, J. Taylor, R. Tibshirani, and R. Tibshirani, "A Significance Test for the LASSO," *Ann. Statist.*, vol. 42, no. 2, pp. 413–468, 04 2014. [Online]. Available: http://dx.doi.org/10.1214/13-AOS1175

[26] M. Pereyra, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Maximum-a-posteriori Estimation With Unknown Regularisation Parameters," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 230–234.

[27] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Review*, vol. 43, pp. 129–159, 2001.

[28] T. Sun and C. H. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, p. 879, 2012. [Online]. Available: + http://dx.doi.org/10.1093/biomet/ass043

[29] P. Stoica, D. Zachariah, and L. Li, "Weighted SPICE: A Unified Approach for Hyperparameter-Free Sparse Estimation," *Digit. Signal Process.*, vol. 33, pp. 1–12, October 2014.

[30] C. R. Rojas, D. Katselis, and H. Hjalmarsson, "A Note on the SPICE Method," *IEEE Transactions on Signal Processing*, vol. 61, no. 18, pp. 4545–4551, Sept. 2013.

[31] T. Kronvall, S. I. Adalbjörnsson, S. Nadig, and A. Jakobsson, "Group-Sparse Regression Using the Covariance Fitting Criterion," *Elsevier Signal Processing*, vol. 139, pp. 116 – 130, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168417301202

[32] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Englewood Cliffs, N.J.: Prentice-Hall, 1998.

[33] Y. Li, J. Scarlett, P. Ravikumar, and V. Cevher, "Sparsistency of l1-Regularized M-Estimators," *Journal of Machine Learning Research*, vol. 38, pp. 644–652, 2015.

[34] T. Söderström and P. Stoica, *System Identification*. London, UK: Prentice Hall International, 1989.

[35] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[36] E. J. Candès, M. B. Wakin, and S. Boyd, "Enhancing Sparsity by Reweighted l_1 Minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[37] F. Bunea, J. Lederer, and Y. She, "The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms," *IEEE Trans. Inf. Theor.*, vol. 60, no. 2, pp. 1313–1325, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1109/TIT.2013.2290040

[38] P. Embrechts, T. Mikosch, and C. Klüppelberg, *Modelling extremal events for insurance and finance*. New York : Springer, 1997, formerly published in series: Applications of mathematics v 34.

[39] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

[40] M. Stone, "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 44–47, 1977. [Online]. Available: http://www.jstor.org/stable/2984877

[41] D. L. Donoho, "De-Noising by Soft-Thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[42] M. R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes*, 1st ed. Springer-Verlag New York, 1983.