Unbiased Group-Sparsity Sensing Using Quadratic Envelopes

Marcus Carlsson
Department of Mathematics
Lund University, Sweden
marcus.carlsson@math.lu.se

Jean-Yves Tourneret and Herwig Wendt
IRIT-ENSEEIHT, CNRS UMR 5505
University of Toulouse, France
firstname.lastname@irit.fr

***Abstract*—This paper investigates a new regularization of the group-sparsity estimation problem based on a quadratic envelope operator. The resulting estimator is shown to have a reduced bias when compared to the classical LASSO estimator and is characterized by a simple hyperparameter selection. Numerical results show that the quadratic envelope regularization yields estimates equal to an oracle solution with high probability. The robustness of the proposed hyperparameter selection rule is also analyzed.**

***Index Terms*—Sparse representations, group-sparsity, quadratic envelope regularization, proximal operators.**

I. Introduction and Problem formulation

Sparse representations have received an increasing interest in many signal and image processing applications. These applications include signal and image denoising [1]–[3], classification [4], [5], navigation [6] and anomaly detection [7], [8]. Sparse representations take advantage of the fact that the signals or images of interest can be generally decomposed as a linear combination of few atoms. Some additional works have shown that the estimation performance can be even more improved when the analyzed data are group-sparse [9]. Group-sparsity assumes that the unknown variables of interest can be decomposed into sub-groups of variables with many sub-groups identically equal to zero. This group-sparsity has been used successfully in several signal/image processing applications such as signal reconstruction [10], face recognition [11] or anomaly detection [8]. This section first recalls the principles of group-sparsity and introduces the problem addressed in this work.

Consider a signal which is naturally composed of segments of length k, where each segment has some structure. Mathematically, we represent this as a matrix $\mathbf{X} \in \mathbb{C}^{m\times k}$, so that each row $\mathbf{x}_j$ ($j = 1,\dots,m$) of $\mathbf{X}$ contains a segment of the original "signal". We will think of $\mathbf{X}$ as a vector where each entry is the smaller vector $\mathbf{x}_j$, and introduce the notation $|\mathbf{X}|_2 = (\|\mathbf{x}_1\|_2,\dots,\|\mathbf{x}_m\|_2)^T$ (where $\|\mathbf{x}_j\|_2$ is the ℓ_2 norm of $\mathbf{x}_j$ and ^{T} denotes transposition), which thus collapses the second dimension and reduces $\mathbf{X}$ to a vector in $\mathbb{R}^m$. We consider the classical group-sparsity problem [9]

$$\operatorname*{argmin}_{\mathbf{X}} \mu\||\mathbf{X}|_2\|_0 + \|A(\mathbf{X}) - \mathbf{b}\|_2^2 \qquad (1)$$

where $A : \mathbb{C}^{m\times k} \to \mathbb{C}^n$ is a linear operator into some measurement space, and $\|\mathbf{y}\|_0$ denotes the cardinality of the vector $\mathbf{y}$, i.e., the number of non-zero entries of $\mathbf{y}$. Of course, upon concatenating the rows of $\mathbf{X}$ into a vector, we may represent A in the usual form as a matrix in $\mathbb{C}^{(mk)\times n}$. We assume that $n < mk$ so that the least squares problem $A(\mathbf{X}) = \mathbf{b}$ is ill-posed. This is the classical compressed sensing setup except for the group structure. The parameter μ controls the trade-off between data fidelity and group sparsity. The classical approach to the NP-hard problem (1) is to relax it with the ℓ^1-type problem

$$\operatorname*{argmin}_{\mathbf{X}} \lambda\||\mathbf{X}|_2\|_1 + \|A(\mathbf{X}) - \mathbf{b}\|_2^2, \qquad (2)$$

which was suggested in [9]. Note that $\||\mathbf{X}|_2\|_1$ can be written in the more familiar form $\||\mathbf{X}|_2\|_1 = \sum_{j=1}^m \|\mathbf{x}_j\|_2$ which is often referred to as ℓ_{21} regularization.

In the scalar case $k = 1$, it has recently been shown that for reasonable values of noise, superior results are obtained by relaxing (1) by using its quadratic envelope [12], [13] instead of replacing the ℓ^0 with an ℓ^1 penalty. Indeed, this formulation has been proven to find the so called "oracle solution", i.e., the best solution assuming knowledge of the true location of the zeros in the signal. The objective of this work is to extend these results to the group sparsity problem (1) (i.e., $k > 1$).

The paper is organized as follows. Section II introduces the group-oracle solution and a relaxation of the group-sparsity problem based on a quadratic envelope operator. Section III presents some properties of the quadratic envelope regularization including its computation using proximal operators. Section IV introduces an interesting hyperparameter selection procedure for the proposed group-sparsity estimator based on a quadratic envelope regularization. Simulation results presented in Section V allow the performance of the proposed regularized group-sparsity method to be appreciated. Conclusions and future work are reported in Section VI.

II. The group oracle solution

A. Definition

Assume that $\mathbf{b} = A(\mathbf{X}_{\text{gt}})+\mathbf{e}$ where $\mathbf{e}$ is some additive noise, e.g., resulting from model errors, and the "ground truth" $\mathbf{X}_{\text{gt}}$ is group sparse, i.e., $|\mathbf{X}_{\text{gt}}|_2$ is sparse in the traditional sense. If an "oracle" would reveal the location of the true support $S = \{j : \mathbf{X}_{\text{gt}}(j,\cdot) \neq 0\}$ (where $\mathbf{X}_{\text{gt}}(j,\cdot)$ is the jth row of $\mathbf{X}_{\text{gt}}$), we would still not be able to retrieve $\mathbf{X}_{\text{gt}}$. The best we

can do is to solve the ℓ_2-problem

$$\mathbf{X}_{\text{or}} = \underset{\mathbf{X}|\mathbf{x}_j=0, j\notin S}{\operatorname{argmin}} \|A(\mathbf{X}) - \mathbf{b}\|_2^2, \tag{3}$$

and we refer to $\mathbf{X}_{\text{or}}$ as the *oracle solution* (this is well defined as long as A restricted to the constrained set $\{\mathbf{X}|\mathbf{x}_j = 0, j \notin S\}$ is injective, as we shall assume).

B. Global minimum and unbiased estimator

In the particular case $k = 1$, it is proven in [13] that the oracle solution is the *unique global minimizer* of (1), under mild assumptions on A. The proof is quite complicated, but we expect to extend it to the present group setting in a future publication. Note that the residual error $\|A(\mathbf{X}_{\text{or}}) - \mathbf{b}\|_2$ is clearly less than $\|A(\mathbf{X}_{\text{gt}}) - \mathbf{b}\|_2 = \|\mathbf{e}\|_2$. Therefore, as long as the parameter μ satisfies $\mu > \|\mathbf{e}\|_2^2$, it is quite plausible that $\mathbf{X}_{\text{or}}$ is the global minimum of (1). Indeed, a matrix $\mathbf{X}$ with $\||\mathbf{X}|_2\|_0 > \||\mathbf{X}_{\text{gt}}|_2\|_0$ can then not be a global minimum. It remains the possibility that another equally sparse matrix would provide a better data-fit than the oracle solution, but this is highly unlikely.

It is important to note that the oracle solution $\mathbf{X}_{\text{or}}$ indeed is an *unbiased* estimator for $\mathbf{X}_{\text{gt}}$, unlike the classical ℓ^1 solution (2). To see this, note that the values on the support S are obtained by solving a well posed least squares problem. It easily follows that $\mathbf{X}_{\text{or}} - \mathbf{X}_{\text{gt}}$ is a linear function of the noise $\mathbf{e}$, and hence its expectation on each coordinate equals 0. An illustration for this is provided in row 3 of Fig. 1 whereas row 4 shows the corresponding bias for the ℓ^1 solution, for the two signals in the first row (with 2 and 4 non-zero groups, respectively), where it may be useful to emphasize that different scales are used in rows 3 and 4 of Fig. 1 (details on the experimental setup are provided in Section V).

C. Computation using quadratic envelopes

Our approach to solving the NP-hard problem (1) is to regularize via the quadratic envelope $\mathcal{Q}_\gamma$, where γ is a user-parameter (c.f. [12], [13]). More precisely, we propose to solve the group-sparsity problem as follows

$$\underset{\mathbf{X}}{\operatorname{argmin}}\ \mathcal{Q}_\gamma(\mu\||\cdot|_2\|_0)(\mathbf{X}) + \|A(\mathbf{X}) - \mathbf{b}\|_2^2. \tag{4}$$

Details about this "quadratic envelope operator" $\mathcal{Q}_\gamma$ are provided in the next section. It has been shown to have the exceptional property of regularizing without moving the global minima (as long as $\|A\|^2 < \gamma$), while at the same time removing many local minima. Here the norm refers to the operator norm of $\|A\|$ (which is identical to the matrix norm of any matrix representation of A based on vectorizing $\mathbf{X}$). Our main finding here is that for reasonable values of signal to noise ratio (SNR), the solution of (4) *coincides with the group oracle solution with high probability*. A numerical illustration for this result is provided in Fig. 2 (cf. Section V for details).

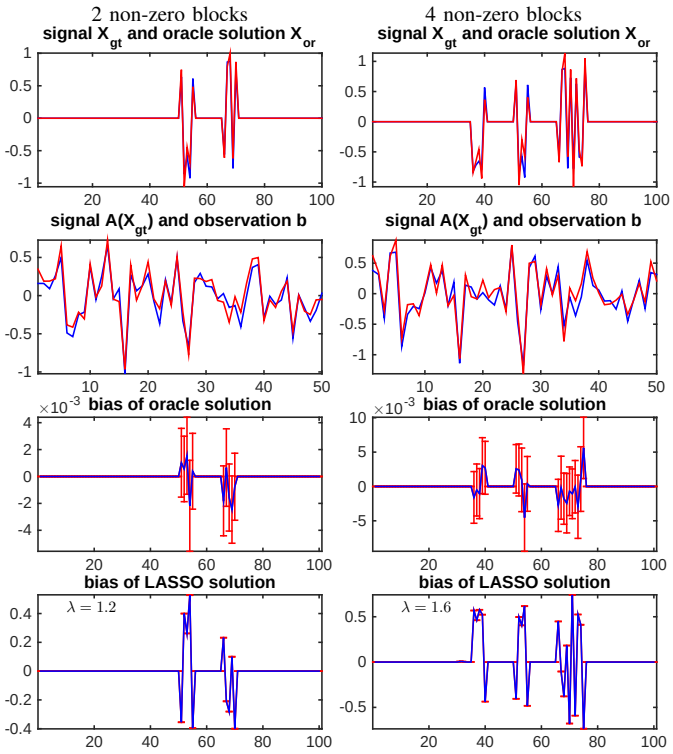

Fig. 1: **Signals and oracle solutions.** Top row: noise-free signals $\mathbf{X}_{\text{gt}}$ (blue) and oracle solutions $\mathbf{X}_{\text{or}}$ (red) for SNR=10dB; second row: noise-free and noisy observations $A(\mathbf{X}_{\text{gt}})$ (blue) and $\mathbf{b}$ (red), respectively; average bias (blue) and 95% error bars (red) for oracle solution $\mathbf{X}_{\text{or}}$ and standard ℓ^1 solution (third and bottom row, respectively).

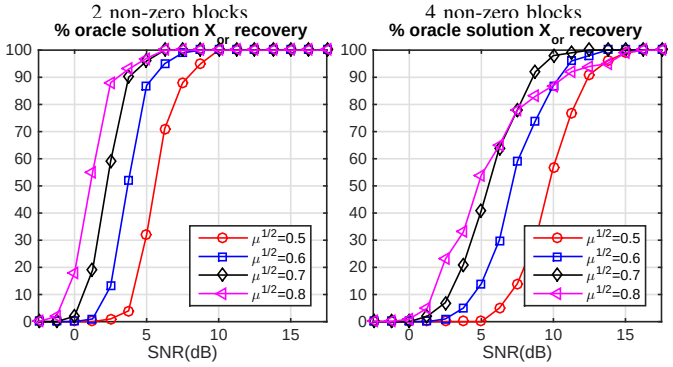

Fig. 2: **Finding the oracle solution.** Success rate of finding $\mathbf{X}_{\text{or}}$ with the proposed method (4) for different values of μ.

III. Regularization via quadratic envelope

A. The quadratic envelope

The quadratic envelope can be computed for a cost functional f on any finite dimensional Hilbert space, in particular $\mathbb{R}^{k\times m}$ or $\mathbb{C}^{k\times m}$ equipped with the Frobenius norm. In this section we represent points by $\mathbf{x}$ and $\mathbf{y}$, whatever the space may be. Given a parameter $\gamma > 0$, the quadratic envelope $\mathcal{Q}_\gamma(f)$ at a point $\mathbf{x}$ is defined as the supremum of $\{\alpha - \frac{\gamma}{2}\|\mathbf{x} - \mathbf{y}\|^2\}$ over all $\alpha \in \mathbb{R}$ and $\mathbf{y}$ such that $\alpha - \frac{\gamma}{2}\|\cdot - \mathbf{y}\|^2 \leq f$. Fig. 4 explains the idea in the simple case where $\mathbf{x} \in \mathbb{R}$. The key result of [12] reads as follows.

Theorem 1: If $2\|A\|^2 < \gamma$ and $f \geq 0$ is lower semi-continuous, then $\mathcal{Q}_\gamma(f)(\mathbf{x}) + \|A(\mathbf{x}) - \mathbf{b}\|_2^2$ has the same global

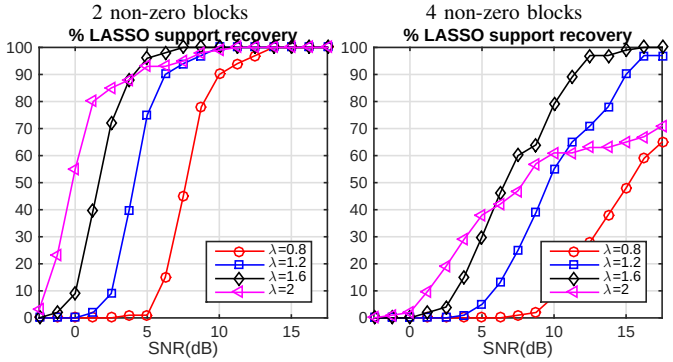

Fig. 3: ℓ^1 **support recovery.** Success rate of finding the true support with the ℓ^1 approach (2) for different values of λ.

minima as $f(\mathbf{x})+\|A(\mathbf{x})-\mathbf{b}\|_2^2$, and the set of local minimizers for the former is a subset of the latter.

A key result of [12] is that the maximum negative curvature of $\mathcal{Q}_\gamma(f)$ is γ. On the other hand, the maximum positive quadrature of $\|A(\mathbf{x})-\mathbf{b}\|_2^2$ is $2\|A\|^2$, which in a way explains the assumption $2\|A\|^2 < \gamma$[1]. Hence γ controls the tradeoff between convexity/concavity; a large value of γ thus leads to a "more" non-convex optimization problem, and better performance is often achieved by choosing γ below the theoretical bound from the above theorem (more on this will be provided in the next section).

For $k = 1$ and the particular choice of $f(\mathbf{x}) = \mu\|\mathbf{x}\|_0$, it turns out that $\mathcal{Q}_\gamma(\mu\|\mathbf{x}\|_0)$ equals the Minimax Concave Penalty (MCP) [14] as well as CEℓ0 (for normalized columns) investigated in [15], from which Theorem 1 is inspired.

B. Proximal operators and computation

We consider the computation of $\mathcal{Q}_\gamma(f(|\cdot|_2))$ and of related proximal operators. With the particular choice $f(\mathbf{X}) = \mu\||\mathbf{X}|_2\|_0$ we retrieve the situation in the previous sections, but the theory applies to any f for which $\mathcal{Q}_\gamma(f)$ is computable, see [16] for many other relevant choices. Straightforward computations lead to the following formulas

$$\mathcal{Q}_\gamma(f(|\cdot|_2))(\mathbf{X}) = \mathcal{Q}_\gamma(f)(|\mathbf{X}|_2) \tag{5}$$

and

$$((\mathbf{X}))_j = (\mathrm{prox}_{\mathcal{Q}_\gamma(f)}(|\mathbf{X}|_2))_j \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|_2} \tag{6}$$

(which hold both in the real and complex cases). Armed with these, we can now pick formulas for $\mathrm{prox}_{\mathcal{Q}_\gamma(f)}$ off the shelf from the scalar situation, and apply these to the group-sparse problem. We recall that

$$\mathrm{prox}_{\mathcal{Q}_\gamma(f)/\rho}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{x}} \mathcal{Q}_\gamma(f)(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \tag{7}$$

where it is important that $\rho \geq \gamma$ to have a convex minimization problem. In particular, for $\gamma = 2$, $\mathrm{prox}_{\mathcal{Q}_2(\mu\|\cdot\|_0)/\rho}(\mathbf{x})$ can

[1]Note that the formulation in [12] does not have the factor 2 due to the fact that the data fit term is multiplied by $\frac{1}{2}$.

Algorithm 1: FBS algorithm for solving (4)

Input : $\mathbf{y}$, A, μ, γ, ρ
Output: $\mathbf{X}$

Initialize $\mathbf{X}$
while *stopping criterion not met* **do**
  $\mathbf{X} \leftarrow \mathbf{X} - \frac{2}{\rho}(A^TA\mathbf{X} - A^T\mathbf{y})$ (gradient step)
  $\mathbf{X} \leftarrow \mathrm{prox}_{\mathcal{Q}_\gamma(\mu\|\cdot\|_0)/\rho}(\mathbf{X})$ (proximal operator)
end

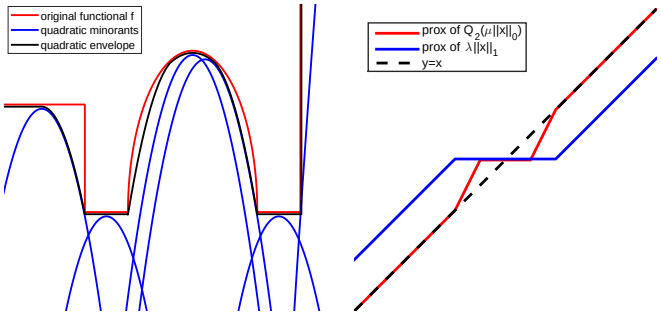

Fig. 4: Left: Illustration of the quadratic envelope. Right: Related proximal operators for the sparsity application.

be computed elementwise by applying a function to each coordinate x_j of $\mathbf{x}$. This function is defined for $t \geq 0$ by

$$t \mapsto \begin{cases} 0 & 0 \leq t < \frac{2\sqrt{\mu}}{\rho} \\ \frac{\rho t - 2\sqrt{\mu}}{\rho - 2} & \frac{2\sqrt{\mu}}{\rho} \leq t < \sqrt{\mu} \\ t & \sqrt{\mu} \leq t \end{cases}$$

and is extended to negative values by requiring it to be odd, see the red graph of Fig. 4 for an illustration.

Algorithmically, we suggest the use of forward-backward splitting (FBS), since by combining results of [12] with [17], we know that this algorithm converges to a stationary point in this non-convex setting. Having said that, we have observed numerically that ADMM works just as well. The FBS-algorithm for solving (4) is summarized in Alg. 1.

IV. Hyperparameter selection

For the particular case of $f(\mathbf{x}) = \mu\|\mathbf{x}\|_0$ ($k = 1$) it is shown in [13] that the conclusion of Theorem 1 still holds if we pick $\gamma = 2$ and only assume that the columns of A are normalized (in which case $\|A\| \geq 1$). In this work we chose $\gamma = 2$ and normalized A so that each block A_j of the matrix realization of A (related to each vector $\mathbf{x}_j$ in $\mathbf{X}$) is normalized to 1. It is our belief that we can prove that (1) and (4) have the same global minimizer under this assumption, as Fig. 2 strongly suggests, but we leave this as a conjecture for future work.

It remains to settle how to chose μ. In Section II we already derived the lower bound $\sqrt{\mu} > \|\mathbf{e}\|_2$. In Section III-B we saw that the proximal operator $\mathrm{prox}_{\mathcal{Q}_2(\mu\|\cdot\|_0)/\rho}(\mathbf{x})$ acts as the identity on any entry of $\mathbf{x}_j$ of $\mathbf{X}$ with $\|x_j\|_2 > \sqrt{\mu}$, which would suggest $\sqrt{\mu} < \min_j\{\|\mathbf{X}_{gt}(j,:)\|_2 : j \in S\}$. However, this estimate does not take into account effects of noise on the

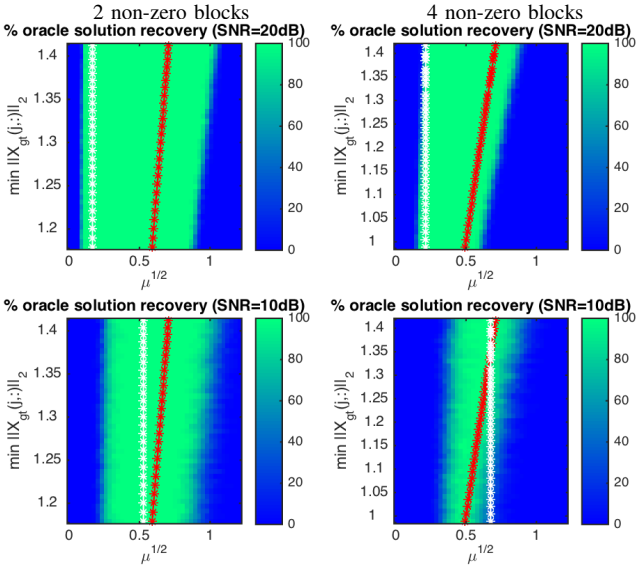

Fig. 5: **Hyperparameter selection.** Success rate of finding $\mathbf{X}_{\text{or}}$ with the proposed method (4) for different values of μ and signals, and rules of thumb (8) (white and red lines, respectively): SNR=20dB (top row), SNR=10dB(bottom row).

non-zero entries, and is therefore most likely too optimistic. In the scalar case $k = 1$, a more careful analysis leads to

$$\left(1 + \frac{1}{\beta_N^2}\right)\sqrt{\mu} < \min_j\{\|\mathbf{X}_{gt}(j,:)\|_2 : \ j \in S\}$$

where $\beta_N > 1$. Our experiments have shown that

$$\|\mathbf{e}\|_2 < \sqrt{\mu} < \frac{\min_j\{\|\mathbf{X}_{gt}(j,:)\|_2 : \ j \in S\}}{2} \tag{8}$$

is a conservative and reasonable rule of thumb.

V. Numerical experiments

Monte Carlo simulations. We test our algorithm on two group-sparse signals $\mathbf{X}_{\text{gt}}$ with $m = 20$ blocks of $k = 5$ samples, of which 2 and 4 blocks are non-zero (corresponding to 10 and 20% of samples), respectively. The samples within each block are uniform random variables on the interval $(\alpha, 1)$, with random sign. For the results reported below, $\alpha = 1$. Similar results have been obtained with $\alpha = 1/4$ and $\alpha = 0$ and are not reported here for space reasons. The resulting signals are depicted in Fig. 1 (top row). A is constructed with centered standard Gaussian entries, properly normalized. The $n = 50$ observations $\mathbf{b}$ are constructed from $A(\mathbf{X}_{\text{gt}})$ by adding centered Gaussian noise. Examples for observations $\mathbf{b}$ for the two signals are plotted in Fig. 1 (second row). The results are obtained for 100 independent realizations of noise $\mathbf{e}$, for different signal to noise ratios defined as $\text{SNR} = 10\log_{10}(||A(\mathbf{X}_{\text{gt}})||_2^2/||\mathbf{e}||_2^2)$ (in dB). We set $\gamma = 2$, $\rho = 32$ and initialize $\mathbf{X}$ with zero entries. Note that as long as $\rho > 2$, ρ essentially only affects speed of convergence, except for isolated cases for very low SNR values, for which a larger value for ρ was found to be beneficial (we tested $\rho \in (2 + \varepsilon, 64)$).

Finding the oracle solution. Fig. 2 plots the empirical rate of success (in %) for the proposed algorithm to find the oracle solution $\mathbf{X}_{\text{or}}$, for different values of μ and for a range of SNR. It demonstrates that for reasonable values of SNR ($\geq$ 10dB for the sparser signal, and $\geq$ 15dB for the signal with 20% non-zero coefficients, which is more difficult to estimate), the proposed algorithm *converges to the oracle solution every time*. Even for much smaller SNR values ($\geq$ 2dB and $\geq$ 5dB, respectively), it recovers the oracle solution with probability larger than $1/2$. As discussed above, the oracle solution that is found by our algorithm in these cases is *unbiased*, see Fig. 1 (3rd row) for an illustration for SNR = 10dB. For comparison, Fig. 3 plots the success rate for the ℓ^1 approach (2) to find the true support of the non-zero elements of $\mathbf{X}_{\text{gt}}$. It indicates that large values of λ ($\lambda = 1.2$ and $\lambda = 1.6$ for the two signals, respectively) lead to higher probability for finding the true support and that this probability remains below that of the proposed method for finding the oracle solution. Moreover, even in the cases where the true support is recovered using ℓ^1, the solution is of course strongly biased, see Fig. 1 (4th row) for an illustration for SNR = 10dB. In fact, the shown values for λ are much higher than typical recommendations found in the literature, and the corresponding reconstructions are clearly suboptimal in terms of distance to ground truth.

Robustness to hyperparameter tuning. Fig. 5 plots the empirical rate of success (in %) for the proposed algorithm to find the oracle solution $\mathbf{X}_{\text{or}}$, for different values of μ, and for different minimum block norms of the signal (for SNR= 20dB and 10dB), together with the rule (8) for picking μ. It shows that the proposed method finds the oracle solution for a relatively large range of values for μ, even in the most difficult scenario (10dB SNR, 4 non-zero blocks). More surprisingly, it indicates that the estimates (8) of μ are reasonable but overly pessimistic, so that the range of values of μ for which our algorithm finds the oracle solution is larger in practice than the theoretical estimate.

VI. Conclusions

This paper proposed and studied a novel approach to solving the group-sparse estimation problem (1). Our algorithm relies on the use of the quadratic envelope of (1) which, unlike the classical ℓ^1 relaxation, is constructed to yield unbiased estimates. We demonstrated that for reasonable signal to noise ratio, the proposed approach finds the group oracle solution with high probability, that is, it finds the true support and an estimate for the non-zero coefficients that is as good as if one knew the true positions of the zeros in the signal beforehand, hence performs significantly better than the classical ℓ^1 approach. Moreover, we showed numerically that the algorithm is quite robust to the choice of the hyperparameter, and provide rules for its selection. Future work will investigate the application of the method to anomaly detection in telemetry time series. In this application, each group of variables corresponds to correlated time series, which are for instance acquired with the same device onboard a satellite.

REFERENCES

[1] M. Elad and A. Aharon, "Image denoising via sparse and redundant representation over learned dictionaries," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 3736–3745, Dec. 2006.

[2] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[3] L. Chaari, J.-Y. Tourneret, and C. Chaux, "Sparse signal recovery using a bernoulli generalized gaussian prior," in *Proc. Euro. Conf. Signal Process. (EUSIPCO'15)*, Nice, France, Aug. 31-Sept. 4 2015.

[4] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS'06)*, Whistler, B. C., Dec. 2006.

[5] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. Conf. Comput. Vis. Pattern Recognition (CVPR'09)*, Miami, FL, June 2009.

[6] J. Lesouple, T. Robert, M. Sahmoudi, J.-Y. Tourneret, and W. Vigneau, "Multipath mitigation for GNSS positioning in urban environment using sparse estimation," *IEEE Trans. Intell. Trans. Systems*, vol. 20, no. 4, pp. 1316–1328, Apr. 2019.

[7] A. Adler, M. Elad, Y. Hel-Or, and E. Rivlin, "Sparse coding with anomaly detection," *J. Signal Process. Syst.*, vol. 79, no. 2, pp. 179–188, May 2015.

[8] B. Pilastre, L. Boussouf, S. D'Escrivan, and J.-Y. Tourneret, "Anomaly detection un mixed telemetry data using a sparse representation and dictionary learning," *submitted to Signal Process.*, 2019.

[9] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Statist. Soc. B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[10] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *J. Machine Learning Research*, vol. 212, pp. 3371–3412, Nov. 2011.

[11] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via spare representation," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 31, no. 2, pp. 1–17, Feb. 2009.

[12] M. Carlsson, "On convex envelopes and regularization of non-convex functionals without moving global minima," *Journal of Optimization Theory and Applications, to appear*, 2019.

[13] M. Carlsson, D. Gerosa, and C. Olsson, "An unbiased approach to compressed sensing," *arXiv preprint*, vol. arXiv:1806.05283, 2018.

[14] C. H. Zhang et al., "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[15] E. Soubies, L. Blanc-Féraud, and G. Aubert, "A continuous exact l0 penalty (cel0) for least squares regularized problem," *SIAM Journal on Imaging Sciences*, vol. 8, no. 3, pp. 1607–1639, 2015.

[16] M. Carlsson, "On convexification/optimization of functionals including an l2-misfit term," *arXiv preprint arXiv:1609.09378*, 2016.

[17] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.