Digital Signal Processing 31 (2014) 124–135

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Solving OSCAR regularization problems by fast approximate proximal splitting algorithms

Xiangrong Zeng [a,b,*], Mário A.T. Figueiredo [a,**]

[a] Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001, Lisboa, Portugal
[b] College of Information System and Management, National University of Defense Technology, Changsha 410073, Hunan, China

ARTICLE INFO

Article history:
Available online 18 April 2014

Keywords:
Proximal splitting algorithms
Alternating direction method of multipliers
Group sparsity
Proximity operator
Signal recovery

ABSTRACT

The OSCAR (*octagonal selection and clustering algorithm for regression*) regularizer consists of an ℓ_1 norm plus a pairwise ℓ_∞ norm (responsible for its grouping behavior) and was proposed to encourage group sparsity in scenarios where the groups are a priori unknown. The OSCAR regularizer has a non-trivial proximity operator, which limits its applicability. We reformulate this regularizer as a weighted sorted ℓ_1 norm, and propose its *grouping proximity operator* (GPO) and *approximate proximity operator* (APO), thus making state-of-the-art proximal splitting algorithms (PSAs) available to solve inverse problems with OSCAR regularization. The GPO is in fact the APO followed by additional grouping and averaging operations, which are costly in time and storage, explaining the reason why algorithms with APO are much faster than those with GPO. The convergences of PSAs with GPO are guaranteed since GPO is an exact proximity operator. Although convergence of PSAs with APO is may not be guaranteed, we have experimentally found that APO behaves similarly to GPO when the regularization parameter of the pairwise ℓ_∞ norm is set to an appropriately small value. Experiments on synthetic data and real-world data show the robustness and efficiency of APO, respectively, and experiments on recovery of group-sparse signals (with unknown groups) show that PSAs with APO are very fast and accurate.

1. Introduction

In the past few decades, linear inverse problems have attracted a lot of attention in a wide range of areas, such as statistics, machine learning, signal processing, and compressive sensing, to name a few. The typical forward model is

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

where $\mathbf{y} \in \mathbb{R}^m$ is the measurement vector, $\mathbf{x} \in \mathbb{R}^n$ the original signal to be recovered, $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a known sensing matrix, and $\mathbf{n} \in \mathbb{R}^m$ is noise (usually assumed to be white and Gaussian). In most cases of interest, $\mathbf{A}$ is not invertible (*e.g.*, because $m < n$), making (1) an ill-posed problem (even in the absence of noise), which can only be addressed by using some form of regularization or prior knowledge about the unknown $\mathbf{x}$. Classical regularization formulations seek solutions of problems of the form

$$\min_{\mathbf{x}} f(\mathbf{x}) + \lambda r(\mathbf{x}) \tag{2}$$

or one of the equivalent (under mild conditions) forms

$$\min_{\mathbf{x}} r(\mathbf{x}) \text{ s.t. } f(\mathbf{x}) \le \varepsilon \quad \text{or} \quad \min_{\mathbf{x}} f(\mathbf{x}) \text{ s.t. } r(\mathbf{x}) \le \epsilon, \tag{3}$$

where, typically, $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is the data-fidelity term (under a white Gaussian noise assumption), $r(\mathbf{x})$ is the regularizer that enforces certain properties (such as sparsity) on the target solution, and λ, ε, and ϵ are non-negative parameters.

1.1. Group-sparsity-inducing regularizers

In recent years, much attention has been paid not only to the sparsity of solutions but also to the structure of this sparsity, which may be relevant in some problems and which provides another avenue for inserting prior knowledge into the problem. In particular, considerable interest has been attracted by group sparsity [49], block sparsity [18], or more general structured sparsity [3,24,29]. A classic model for group sparsity is the *group LASSO* (gLASSO) [49], where the regularizer is the so-called $\ell_{1,2}$ norm [37,27] or the $\ell_{1,\infty}$ norm [37,28], defined as $r_{\text{gLASSO}}(\mathbf{x}) = \sum_{i=1}^{s} \|\mathbf{x}_{g_i}\|_2$

* Corresponding author at: Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001, Lisboa, Portugal.
** Principal corresponding author.
E-mail addresses: Xiangrong.Zeng@lx.it.pt (X. Zeng), mario.figueiredo@lx.it.pt (M.A.T. Figueiredo).

http://dx.doi.org/10.1016/j.dsp.2014.03.010

and $\sum_{i=1}^{s} \|\mathbf{x}_{g_i}\|_\infty$, respectively,[1] where $\mathbf{x}_{g_i}$ represents the subvector of $\mathbf{x}$ indexed by g_i, and $g_i \subseteq \{1, ..., n\}$ denotes the index set of the i-th group. Different ways to define the groups lead to overlapping or non-overlapping gLASSO. Notice that if each group above is a singleton, then r_{gLASSO} reduces to r_{LASSO}, whereas if $s = 1$ and $g_1 = \{1, ..., n\}$, then $r_{\text{gLASSO}}(\mathbf{x}) = \|\mathbf{x}\|_2$. Recently, the *sparse gLASSO* (sgLASSO) regularizer was proposed as $r_{\text{sgLASSO}}(\mathbf{x}) = \lambda_1 r_{\text{LASSO}}(\mathbf{x}) + \lambda_2 r_{\text{gLASSO}}(\mathbf{x})$, where λ_1 and λ_2 are non-negative parameters [41]. In comparison with gLASSO, sgLASSO not only selects groups, but also individual variables within each group. Note that one of the costs of the possible advantages of gLASSO and sgLASSO over standard LASSO is the need to define *a priori* the structure of the groups.

In some problems, the components of $\mathbf{x}$ are known to be similar in value to its neighbors (assuming that there is some natural neighborhood relation defined among the components of $\mathbf{x}$). To encourage this type of solution (usually in conjunction with sparsity), several proposals have appeared, such as the elastic net [53], the fused LASSO (fLASSO) [43], *grouping pursuit* (GS) [40], and the *octagonal shrinkage and clustering algorithm for regression* (OSCAR) [7]. The elastic net regularizer is $r_{\text{elast-net}}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{x}\|_2^2$, encouraging both sparsity and grouping [53]. The fLASSO regularizer is given by $r_{\text{fLASSO}}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_i |x_i - x_{i+1}|$, where the *total variation* (TV) term (sum of the absolute values of differences) encourages consecutive variables to be similar; fLASSO is thus able to promote both sparsity and smoothness. The GS regularizer is defined as $r_{\text{GS}}(\mathbf{x}) = \sum_{i<j} G(x_i - x_j)$, where $G(\mathbf{z}) = |\mathbf{z}|$, if $|\mathbf{z}| \le \lambda$, and $G(\mathbf{z}) = \lambda$, if $|\mathbf{z}| > \lambda$ [40]; however, r_{GS} is neither sparsity-promoting nor convex.

Finally, r_{OSCAR} [7] has the form $r_{\text{OSCAR}}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}$; due to the ℓ_1 term and the pairwise ℓ_∞ penalty, the components are encouraged to be sparse and pairwise similar in magnitude. The OSCAR has been extensively applied in various feature grouping tasks [7,36,47,51]. Notice that OSCAR does not promotes group-sparsity in the same sense as other methods; what it does is promote standard sparsity (via the presence of an ℓ_1 term) and a grouping behavior, that is, components of similar magnitude (regardless of their position in the vector/signal) are encouraged to take the same value. This grouping encouragement makes OSCAR close to standard group-sparsity regularizers, but with a key difference: the found groups depend exclusively on the magnitudes, not on the position of the involved signal components.

Other recently proposed group-sparsity regularizers include the adaptive LASSO (aLASSO) [52], where the regularizer is $r_{\text{aLASSO}}(\mathbf{x}) = \lambda \sum_i |x_i|/|\tilde{x}_i|^\gamma$, where $\tilde{\mathbf{x}}$ is an initial consistent estimate of $\mathbf{x}$, and λ and γ are positive parameters. The *pairwise* fLASSO (pfLASSO [35]) is a variant of fLASSO, given by $r_{\text{pfLASSO}}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} |x_i - x_j|$, is related to OSCAR, and extends fLASSO to cases where the variables have no natural ordering. Another variant is the *weighted* fLASSO (wfLASSO [14]), given by $r_{\text{wfLASSO}}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} w_{ij}(x_i - \text{sign}(\rho_{ij}) x_j)^2$, where ρ_{ij} is the sample correlation between the i-th and j-th predictors, and w_{ij} is a non-negative weight. Finally, the recent *graph-guided* fLASSO (ggfLASSO [25]) regularizer is based on a graph $G = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is the set of variable nodes and $\mathbf{E} \subseteq \mathbf{V}^2$ the set of edges: $r_{\text{ggfLASSO}}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{(i,j)\in\mathbf{E}, i<j} w_{ij} |x_i - \text{sign}(r_{ij}) x_j|$, where r_{ij} represents the weight of the edge $(i, j) \in \mathbf{E}$; if $r_{ij} = 1$, r_{ggfLASSO} reduces to r_{fLASSO}, and the former can group variables with different signs through the assignment of r_{ij}, while the latter cannot. Some other graph-guided group-sparsity-inducing regularizers have been proposed in [47], and all this kind of regularizers needs a strong requirement – the prior information on an undirected graph.

[1] Recall that $\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, ..., |x_n|\}$.

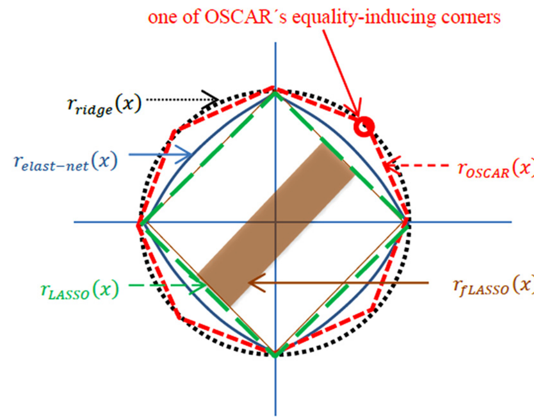

Fig. 1. Illustration of r_{fLASSO}, $r_{\text{elast-net}}$, r_{LASSO}, r_{OSCAR} and r_{ridge} which is the regularizer of ridge regression.

For the sake of comparison, several of the above mentioned regularizers are illustrated in Fig. 1, where the corresponding level curves (balls) are depicted; we also plot the level curve of the classical ridge regularizer $r_{\text{ridge}} = \|\mathbf{x}\|_2^2$. We can see that $r_{\text{OSCAR}}(\mathbf{x})$, $r_{\text{elast-net}}(\mathbf{x})$, and $r_{\text{fLASSO}}(\mathbf{x})$ promote both sparsity and grouping, but their grouping behaviors are clearly different: 1) OSCAR encourages equality (in magnitude) of each pair of variables, as will be discussed in detail in Section 2; 2) elastic net is strictly convex, but doesn't promote strict equality like OSCAR; 3) the total variation term in fLASSO can be seen to encourage sparsity in the differences between each pair of successive variables, thus its recipe of grouping is to guide variables into the shadowed region shown in Fig. 1, which corresponds to $\sum_i |x_i - x_j| \le \varsigma$ (where ς is a function of λ_2).

As seen above, fLASSO, the elastic net, and OSCAR have the potential to be used as regularizers when it is known that the components of the unknown signal exhibit structured sparsity, but a group structure is not *a priori* known. However, as pointed out in [51], OSCAR outperforms the other two models in feature grouping. Moreover, fLASSO is not suitable for grouping according to magnitude, since it cannot group positive and negative variables together, even if their magnitudes are similar; fLASSO also relies on a particular ordering of the variables, for which there may not always be a natural choice. Consequently, we will focus on the OSCAR regularizer in this paper.

In [7], a costly quadratic programming approach was adopted to solve the optimization problem corresponding to OSCAR. More recently, [36] solved OSCAR in a generalized linear model context; the algorithm therein proposed solves a complicated constrained maximum likelihood problem in each iteration, which is also costly. An efficient algorithm was proposed in [51], by reformulating OSCAR as a quadratic problem and then applying FISTA (*fast iterative shrinkage-thresholding algorithm*) [5]. To the best of our knowledge, this is the currently fastest algorithm for OSCAR. In this paper, we propose reformulating r_{OSCAR} as a weighted and sorted ℓ_1 norm, and present an exact grouping proximity operator (termed GPO) of r_{OSCAR} that is based on a projection step proposed in [51] and an element-wise approximate proximal step (termed APO). We show that GPO consists of APO and an additional grouping and averaging operation. Furthermore, we use alternative state-of-the-art *proximal splitting algorithms* (PSAs, briefly reviewed next) to solve the problems involved by the OSCAR regularization.

1.2. Proximal splitting and augmented Lagrangian algorithms

In the past decade, several special purpose algorithms have been proposed to solve optimization problems of the form (2)