Multi-dimensional sparse structured signal approximation using split bregman iterations

Yoann Isaac, Quentin Barthélemy, Cedric Gouy-Pailler, Jamal Atif, Michèle Sebag

MULTI-DIMENSIONAL SPARSE STRUCTURED SIGNAL APPROXIMATION USING SPLIT BREGMAN ITERATIONS

Yoann Isaac[1,2]*, Quentin Barthélemy*[1]*, Jamal Atif*[2]*, Cédric Gouy-Pailler*[1]*, Michèle Sebag*[2]

[1] CEA, LIST
Data Analysis Tools Laboratory
91191 Gif-sur-Yvette CEDEX, FRANCE

[2] TAO, CNRS − INRIA − LRI
Université Paris-Sud
91405 Orsay, FRANCE

ABSTRACT

The paper focuses on the sparse approximation of signals using overcomplete representations, such that it preserves the (prior) structure of multi-dimensional signals. The underlying optimization problem is tackled using a multi-dimensional extension of the split Bregman optimization approach. An extensive empirical evaluation shows how the proposed approach compares to the state of the art depending on the signal features.

***Index Terms*—** Sparse approximation, Regularization, Fused-LASSO, Split Bregman, Multidimensional signals

1. INTRODUCTION

Dictionary-based representations proceed by approximating a signal via a linear combination of dictionary elements, referred to as atoms. Sparse dictionary-based representations, where each signal involves but a few atoms, have been thoroughly investigated for 1D and 2D signals for their good properties, as they enable robust transmission (compressed sensing [1]) or image in-painting [2]. The dictionary is either given, based on the domain knowledge, or learned from the signals [3].

The so-called sparse approximation algorithm aims at finding a sparse approximate representation of the considered signals using this dictionary, by minimizing a weighted sum of the approximation loss and the representation sparsity (see [4] for a survey). When available, prior knowledge about the application domain can also be used to guide the search toward "plausible" decompositions.

This paper focuses on sparse approximation enforcing a structured decomposition property, defined as follows. Let the signals be structured (e.g. being recorded in consecutive time steps); the structured decomposition property then requires that the signal structure is preserved in the dictionary-based representation (e.g. the atoms involved in the approximation of consecutive signals have "close" weights). The structured decomposition property is enforced through adding a total variation (TV) penalty to the minimization objective.

In the 1D case, the minimization of the above overall objective can be tackled using the fused-LASSO approach first introduced in [5]. In the case of multi-dimensional signals[1] however, the minimization problem presents additional difficulties. The first contribution of the paper is to show how this problem can be handled efficiently, by extending the (1D) split Bregman fused-LASSO solver presented in [6], to the multi-dimensional case. The second contribution is a comprehensive experimental study, comparing state-of-the-art algorithms to the presented approach referred to as Multi-SSSA and establishing their relative performance depending on diverse features of the structured signals.

The section 2 introduces the formal background. The proposed optimization approach is described in section 3.1. Section 4 presents our experimental setting and reports on the results. The presented approach is discussed w.r.t. related work in section 5 and the paper concludes with some perspectives for further research.

2. PROBLEM STATEMENT

Let $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_T] \in \mathbb{R}^{C \times T}$ be a matrix made of T C-dimensional signals, $\Phi \in \mathbb{R}^{C \times N}$ an overcomplete dictionary of N atoms ($N > C$). We consider the linear model

$$\mathbf{y}_t = \Phi \mathbf{x}_t + \mathbf{e}_t, \;\; t \in 1, \ldots, T \;\; , \tag{1}$$

in which $X = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{N \times T}$ stands for the decomposition matrix and $E = [\mathbf{e}_1, \ldots, \mathbf{e}_T] \in \mathbb{R}^{C \times T}$ is a zero-mean Gaussian noise matrix.

The sparse structured decomposition problem consists of approximating the $\mathbf{y_i}, i \in \{1, \ldots, T\}$ by decomposing them on the dictionary Φ, such that the structure of the decompositions $\mathbf{x}_i$ reflects that of the signals $\mathbf{y}_i$. This goal is formalized as

[1] Our motivating application considers electro-encephalogram (EEG) signals, where the number of sensors ranges up to a few hundreds.

the minimization[2] of the objective function

$$\min_X \|Y - \Phi X\|_2^2 + \lambda_1 \|X\|_1 + \lambda_2 \|XP\|_1 \quad , \tag{2}$$

where λ_1 and λ_2 are regularization coefficients and P encodes the signal structure (provided by the prior knowledge) as in [7]. In the remainder of the paper, the considered structure is that of the temporal ordering of the signals, *i.e.* $\|XP\|_1 = \sum_{t=2}^{T} \|X_t - X_{t-1}\|_1$.

3. OPTIMIZATION STRATEGY

3.1. Algorithm description

Bregman iterations have shown to be very efficient for ℓ_1 regularized problem [8]. For convex problems with linear constraints, the split Bregman iteration technique is equivalent to the method of multipliers and the augmented Lagrangian one [9]. The iteration scheme presented in [6] considers an augmented Lagrangian formalism. We have chosen here to present ours with the initial split Bregman formulation.

First, let us restate the sparse approximation problem

$$\begin{array}{ll} \min_X & \|Y - \Phi X\|_2^2 + \lambda_1 \|A\|_1 + \lambda_2 \|B\|_1 \\ \text{s.t.} & A = X \\ & B = XP \end{array} \quad . \tag{3}$$

This reformulation is a key step of the split Bregman method. It decouples the three terms and allows to optimize them separately within the Bregman iterations. To set-up this iteration scheme, Eq.(3) must be transform to an unconstrained problem

$$\begin{array}{ll} \min_{X,A,B} & \|Y - \Phi X\|_2^2 + \lambda_1 \|A\|_1 + \lambda_2 \|B\|_1 \\ & + \frac{\mu_1}{2} \|X - A\|_2^2 + \frac{\mu_2}{2} \|XP - B\|_2^2 \end{array} \quad . \tag{4}$$

The split Bregman iterations could then be expressed as [8]

$$\begin{aligned} (X^{i+1}, A^{i+1}, B^{i+1}) = & \ \operatorname{argmin}_{X,A,B} \|Y - \Phi X\|_2^2 \\ & + \lambda_1 \|A\|_1 + \lambda_2 \|B\|_1 \\ & + \tfrac{\mu_1}{2} \|X - A + D_A^i\|_2^2 \\ & + \tfrac{\mu_2}{2} \|XP - B + D_B^i\|_2^2 \end{aligned} \tag{5}$$

$$D_A^{i+1} = \ D_A^i + (X^{i+1} - A^{i+1}) \tag{6}$$

$$D_B^{i+1} = \ D_B^i + (X^{i+1}P - B^{i+1}) \tag{7}$$

Thanks to the split of the three terms realized above, the minimization of Eq.(5) could be realized iteratively by alternatively updating variables in the system

$$\begin{aligned} X^{i+1} = & \operatorname{argmin}_X \|Y - \Phi X\|_2^2 + \tfrac{\mu_1}{2} \|X - A^i + D_A^i\|_2^2 \\ & + \tfrac{\mu_2}{2} \|XP - B^i + D_B^i\|_2^2 \end{aligned} \tag{8}$$

$$A^{i+1} = \operatorname{argmin}_A \lambda_1 \|A\|_1 + \tfrac{\mu_1}{2} \|X^{i+1} - A + D_A^i\|_2^2 \tag{9}$$

$$B^{i+1} = \operatorname{argmin}_B \lambda_2 \|B\|_1 + \tfrac{\mu_2}{2} \|X^{i+1}P - B + D_B^i\|_2^2 \tag{10}$$

[2] $\|A\|_p = (\sum_i \sum_j |A_{i,j}|^p)^{\frac{1}{p}}$. The case $p = 2$ corresponds to the classical Frobenius norm

Only few iterations of this system are necessary for convergence. In our implementation, this update is only performed once at each iteration of the global optimization algorithm.

Eq.(9), Eq.(10) could be resolved with the soft-thresholding operator

$$A^{i+1} = \ \operatorname{SoftThreshold}_{\frac{\lambda_1}{\mu_1}\|.\|_1}(X^{i+1} + D_A^i) \tag{11}$$

$$B^{i+1} = \ \operatorname{SoftThreshold}_{\frac{\lambda_2}{\mu_1}\|.\|_1}(X^{i+1}P + D_B^i) \quad . \tag{12}$$

Solving Eq.(8) requires the minimization of a convex differentiable function which can be performed via classical optimization methods. We proposed here to solve it deterministically. The main difficulty in extending [6] to the multi-dimensional signals case rely on this step. Let us define H from Eq.(8) such as

$$X^{i+1} = \ \operatorname{argmin}_X H(X) \quad . \tag{13}$$

Differentiating this expression with respect to X yields

$$\frac{d}{dX}H = \ (2\Phi^T\Phi + \mu_1 I)X + X(\mu_2 PP^T) - 2\Phi Y \tag{14}$$

$$+\mu_1(D_A^i - A^i) + \mu_2(D_B^i - B^i)P^T \quad , \tag{15}$$

where I is the identity matrix. The minimum $\hat{X} = X^{i+1}$ of Eq.(8) is obtained by solving $\frac{d}{dX}H(\hat{X}) = 0$ which is a Sylvester equation

$$W\hat{X} + \hat{X}Z = C^i \quad , \tag{16}$$

with $W = 2\Phi^T\Phi + \mu_1 I$, $Z = \mu_2 PP^T$ and $C = -U^i + 2\Phi Y + \mu_1 A^i + (\mu_2 B^i - V^i)P^T$. Fortunately, in our case, W and Z are real symmetric matrix. Thus, they can be diagonalized as follow:

$$W = FD_wF^T \tag{17}$$

$$Z = GD_zG^T \tag{18}$$

and Eq.(16) can then be rewritten

$$D_w\hat{X}' + \hat{X}'D_z = C^{i\prime} \quad , \tag{19}$$

with $\hat{X}' = F^T\hat{X}G$ and $C^{i\prime} = F^TC^iG$. $\hat{X}'$ is then obtained by

$$\forall s \in \{1, \ldots, S\} \ \ \hat{X}'(:, s) = (D_w + D_z(s, s)I)^{-1}C^{i\prime}(:, s)$$

where the notation $(:, s)$ indices the column s of matrices. Going back to $\hat{X}$ could be performed with: $\hat{X} = F\hat{X}'G^T$. W and Z being independent of the iteration (i) considered, theirs diagonalization is done only once and for all as well as the computation of the terms $(D_w + D_z(s, s)I)^{-1} \ \ \forall s \in \{1, \ldots, S\}$. Thus, this update does not require heavy computation. The full algorithm is summarized below.

3.2. Algorithm sum up

Input: Y, Φ, P
Parameters: λ_1, λ_2, μ_1, μ_2, ϵ, $iterMax$, $kMax$
Init D_A^0, D_B^0 and X^0
$A^0 = X^0 P,\ B^0 = X^0, Y = 2\Phi^T\Phi + \mu_1 I,\ Z = \mu_2 PP^T$
Compute D_w, D_z, F and G.
$i = 0$
while $i \leq iterMax$ and $\frac{\|X^i - X^{i-1}\|_2}{\|X^i\|_2} \geq \epsilon$ **do**
  $k = 0$
  $X^{temp} = X^i$; $A^{temp} = A^i$; $B^{temp} = B^i$
  while $k \leq kMax$ **do**
    $C = F^T(2\Phi^T Y - \mu_1(D_A^i - A^{temp}) - \mu_2(D_B^i - B^{temp})P^T)G$
    for $s \to S$ **do**
      $X^{temp}(:, s) = (D_y + D_z(s, s)I)^{-1}C(:, s)$
    end for
    $X^{temp} = FX^{temp}G^T$
    $A^{temp} = \text{SoftThreshold}_{\frac{\lambda_1}{\mu_1}\|.\|_1}(X^{temp} + D_B^i)$
    $B^{temp} = \text{SoftThreshold}_{\frac{\lambda_2}{\mu_2}\|.\|_1}(X^{temp}P + D_B^i)$
    $k = k + 1$
  end while
  $X^{i+1} = X^{temp}$; $A^{i+1} = A^{temp}$; $B^{i+1} = B^{temp}$
  $D_A^{i+1} = D_A^i + (X^{i+1} - B^{i+1})$
  $D_B^{i+1} = D_B^i + X^{i+1}P - A^{i+1})$
  $i = i + 1$;
end while

4. EXPERIMENTAL EVALUATION

The following experiment aims at assessing the efficiency of our approach in decomposing signals built with particular regularities. We compare it both to algorithms coding each signal separately, the Orthogonal Matching Pursuit [10] and the LARS [11] (a LASSO solver) and to methods performing the decomposition simultaneously, the simultaneous OMP and FISTA [12] a proximal method solving a group-LASSO problem only composed of a $l_{1,2}$ penalty.

4.1. Data generation

From a fixed random overcomplete dictionary Φ, a set of K signals having piecewise constant structures have been created. Each signal is synthesized from the dictionary and a pre-determined decomposition matrix.
The TV penalization of the fused-LASSO regularization makes him more suitable to deal with data having abrupt changes. Thus, the decomposition matrices of signals have been built as linear combinations of activities. This writes as follows:

$$P_{ind,m,d}(i,j) = \begin{cases} 0 & \text{if } i \neq ind \\ \mathcal{H}(j - (m - \frac{d \times T}{2})) & \\ -\mathcal{H}(j - (m + \frac{d \times T}{2})) & \text{if } i = k. \end{cases} \quad (20)$$

where $P \in \mathbb{R}^{N \times T}$, $\mathcal{H}$ is the Heaviside function, $ind \in \{1, \ldots, N\}$ is the index of an atoms, m is the center of the activity and d it's duration. A decomposition matrix X could then be written:

$$X = \sum_{i=1}^{M} a_i P_{ind_i, m_i, d_i} \quad (21)$$

where M is the number of activities appearing in one signal and a_i stands for an activation weight. An example of such signal is given in the figure (4.1) below.

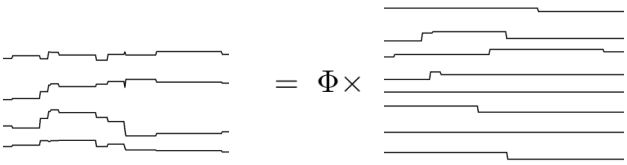

Fig. 1. Built signal

4.2. Experimental setting

Each method has been applied to the previously created signals. Then the distance between the decomposition matrices obtained and the real ones have been computed as follow:

$$dist(X, \hat{X}) = \frac{\|X - \hat{X}\|_F}{\|X\|_F} \quad (22)$$

The goal was to understand the influence of the number of activities (M) and the range of durations (d) on the efficiency of the fused-LASSO regularization compared to others sparse coding algorithms. The scheme of experiment described above has been carried out with the following grid of parameters:

- $M \in \{20, 30, \ldots, 110\}$,
- $d \sim \mathcal{U}(d_{min}, d_{max})$
  $(d_{min} d_{max}) \in \{(0.1, 0.15), (0.2, 0.25), \ldots, (1, 1)\}$

For each point in the above parameters grid, two set of signals has been created: a train set allowing to determine for each method the best regularization coefficients and a test set designed for evaluate them with these coefficients.
Other parameters have been chosen as follows:

Model	Activities
$C = 20$	$m \sim \mathcal{U}(0, T)$
$T = 300$	$a \sim \mathcal{N}(0, 2)$
$N = 40$	$ind \sim \mathcal{U}(1, N)$
$K = 100$	

Dictionaries have been randomly generated using Gaussian independent distributions on individual elements.

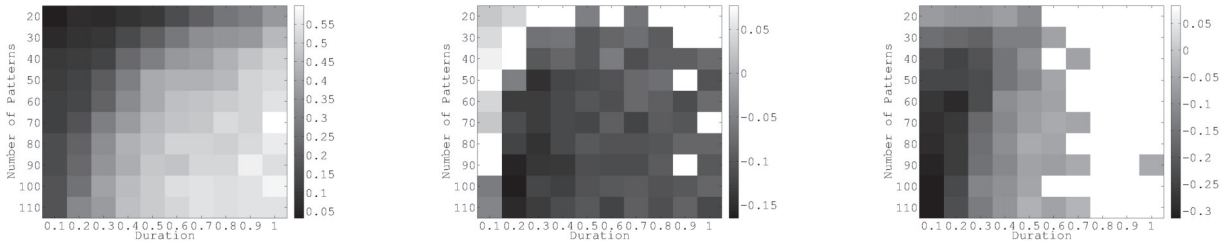

Fig. 2. Mean distances on the grid of parameters. On the left: Fused Lasso, in the middle: Fused Lasso vs LARS, on the right: Fused Lasso vs Group Lasso-Solver . The white mask corresponds to non-significant values.

4.3. Results and discussion

In order to evaluate the proposed algorithm, for each point (i, j) in the above grid of parameters, the mean of the previously defined distance has been computed for each method and compared to the mean obtained by our algorithm. A paired t-test ($p < 0.05$) has then been performed to check the significance of these results.
Results are displayed in Figure 4.3. In the ordinate axis, the number of patterns increases from the top to the bottom and in the abscissa axis, the duration grows from left to right. The left image displays the mean distances obtained with our algorithm. The middle and right one present its performance compared to other methods by displaying the difference (point to point) of mean distances in grayscale. This difference is performed such that, negative values (darker blocks) means that our method outperform the other one. The white mask corresponds to zone where the difference of mean distances is not significant and methods have similar performances. Results of the OMP and the LARS are very similar as well as those of the SOMP and the group-Lasso solver. Thus, we only display here the matrix comparing the our method to the LARS and the group-LASSO solver.

Compared to the OMP and the LARS, our method obtains same results as them when only few atoms are active at the same time. It happens in our artificial signals when only few patterns have been added to create decomposition matrices and/or when the pattern durations are small. On the contrary, when many atoms are active simultaneously, the OMP and LARS are outperformed by the above algorithm which use inter-signal prior informations to find better decompositions.
Compared to the SOMP and the group-LASSO solver, results depends more on the duration of patterns. When patterns are long and not too numerous, theirs performances is similar to the fused-LASSO one. The SOMP is outperformed in all other cases. On the contrary, the group-LASSO solver is outperformed only when patterns are short.

5. RELATION TO PRIOR WORKS

The simultaneous sparse approximation of multi-dimensional signals has been widely studied during these last years [13] and numerous methods developed [14, 15, 16, 17, 4]. More recently, the concept of structured sparsity has considered the encoding of priors in complex regularization [18, 19]. Our problem belongs to this last category with a regularisation combining a classical sparsity term and a Total Variation one. This second term has been studied intensively for image denoising as the in the ROF model [20, 21].
The combination of these terms has been introduced as the fused-LASSO [5]. Despite its convexity, the two ℓ_1 non-differentiable terms make it difficult to solve. The initial paper [5] transforms it to a quadratic problem and uses standard optimization tools (SQOPT). Increasing the number of variables, this approach can not deal with large-scale problem. A path algorithm has been developed but is limited to the particular case of the fused-LASSO signal approximator [22]. More recently, scalable approaches based on proximal subgradient methods [23], ADMM [24] and split Bregman iterations [6] have been proposed for the general fused-LASSO.
To the best or our knowledge, the multi-dimensional fused-LASSO in the context of overcomplete representations has never been studied. One attempt of multi-dimensional fused-LASSO has been found in an arxiv version [7] for regression task, but the journal published version does not contain any mention of the multi-dimensional fused-LASSO anymore.

6. CONCLUSION AND PERSPECTIVES

This paper has shown the efficiency of the proposed Multi-SSSA based on a split Bregman approach, in order to achieve the sparse structured approximation of multi-dimensional signals, under general conditions. Specifically, the extensive validation has considered different regimes in terms of the signal complexity and dynamicity (number of patterns simultaneously involved and average duration thereof), and it has established a *relative competence map* of the proposed Multi-SSSA approach comparatively to the state of the art. Further work will apply the approach to the motivating application domain, namely the representation of EEG signals.

7. REFERENCES

[1] D.L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *Image Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 53–69, 2008.

[3] I. Tošić and P. Frossard, "Dictionary learning: What is the right representation for my signal?," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 27–38, 2011.

[4] A. Rakotomamonjy, "Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms," *Signal processing*, vol. 91, no. 7, pp. 1505–1526, 2011.

[5] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[6] G.B. Ye and X. Xie, "Split bregman method for large scale fused lasso," *Computational Statistics & Data Analysis*, vol. 55, no. 4, pp. 1552–1569, 2011.

[7] X. Chen, S. Kim, Q. Lin, J.G. Carbonell, and E.P. Xing, "Graph-structured multi-task regression and an efficient optimization method for general fused lasso," *arXiv preprint arXiv:1005.3579*, 2010.

[8] T. Goldstein and S. Osher, "The split bregman method for ℓ_1 regularized problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 323–343, 2009.

[9] C. Wu and X.C. Tai, "Augmented lagrangian method, dual methods, and split bregman iteration for rof, vectorial tv, and high order models," *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 300–339, 2010.

[10] Y.C. Pati, R. Rezaiifar, and PS Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 1993, pp. 40–44.

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

[12] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[13] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *Signal Processing, IEEE Transactions on*, vol. 54, no. 12, pp. 4634–4643, 2006.

[14] J.A. Tropp, A.C. Gilbert, and M.J. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, 2006.

[15] J.A. Tropp, "Algorithms for simultaneous sparse approximation. part ii: Convex relaxation," *Signal Processing*, vol. 86, no. 3, pp. 589–602, 2006.

[16] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms," *Journal of Fourier analysis and Applications*, vol. 14, no. 5, pp. 655–687, 2008.

[17] S.F. Cotter, B.D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *Signal Processing, IEEE Transactions on*, vol. 53, no. 7, pp. 2477–2488, 2005.

[18] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *Journal of Machine Learning Research*, vol. 12, pp. 3371–3412, 2011.

[19] R. Jenatton, J.Y Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, pp. 2777–2824, 2011.

[20] L.I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[21] J. Darbon and M. Sigelle, "A fast and exact algorithm for total variation minimization," *Pattern recognition and image analysis*, pp. 717–765, 2005.

[22] H. Hoefling, "A path algorithm for the fused lasso signal approximator," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 984–1006, 2010.

[23] J. Liu, L. Yuan, and J. Ye, "An efficient algorithm for a class of fused lasso problems," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 323–332.

[24] B. Wahlberg, S. Boyd, M. Annergren, and Y. Wang, "An admm algorithm for a class of total variation regularized estimation problems," *arXiv preprint arXiv:1203.1828*, 2012.