FUSED LASSO WITH A NON-CONVEX SPARSITY INDUCING PENALTY

İlker Bayram[⋆] *Po-Yu Chen*[†] *Ivan W. Selesnick*[†]

[⋆] Istanbul Technical University, Istanbul, Turkey
[†] NYU Polytechnic School of Engineering, Brooklyn, NY, USA

ABSTRACT

The fused lasso problem involves the minimization of the sum of a quadratic, a TV term and an ℓ_1 term. The solution can be obtained by applying a TV denoising filter followed by soft-thresholding. However, soft-thresholding introduces a certain bias to the non-zero coefficients. In order to prevent this bias, we propose to replace the ℓ_1 penalty with a non-convex penalty. We show that the solution can similarly be obtained by applying a modified thresholding function to the result of the TV-denoising filter.

Index Terms— Fused lasso, non-convex penalty, thresholding, total variation denoising, audio denoising.

1. INTRODUCTION

Suppose x is a piecewise constant signal, many samples of which are also known to be zero. Given a noisy observation of x, namely y, the fused lasso formulation [12] proposes to reconstruct x as

$$\hat{x} = \underset{x}{\operatorname{argmin}} \ \frac{1}{2}\|y-x\|_2^2 + \lambda_1 \ \mathrm{TV}(x) + \lambda_2 \|x\|_1, \qquad (1)$$

where $\mathrm{TV}(x)$ denotes the total variation of x, $\|x\|_1$ denotes the ℓ_1 norm of x, defined as,

$$\mathrm{TV}(x) = \sum_{i=1}^{N-1} |x_i - x_{i+1}|, \quad \|x\|_1 = \sum_{i=1}^{N} |x_i|, \qquad (2)$$

for $x \in \mathbb{R}^N$. Friedman et al. show in [7] that the solution to this problem can be obtained in two steps.

(i) Set $\hat{z} = \operatorname{argmin}_z \ \frac{1}{2}\|y-z\|_2^2 + \lambda_1 \ \mathrm{TV}(z)$,

(ii) $\hat{x} = \mathrm{ST}_{\lambda_2}(\hat{z}, \lambda_2)$, where $\mathrm{ST}_{\lambda_2}(\cdot)$ denotes the soft-thresholding operator with threshold λ_2.

Existence of finite-terminating algorithms (see e.g. [9, 5, 4, 1]) for the TV denoising problem renders this procedure attractive from a computational point of view.

One objection to the formulation in (1), also made explicit by the two step procedure above, is the use of the ℓ_1 norm as a sparsity inducing prior (or the soft-threshold as a sparsity inducing operator). The problem is that, although the low-magnitude coefficients are thresholded to zero, in line with the sparsity requirement, the remaining non-zero estimates are biased because of the shrinkage towards zero. In sparse denoising/reconstruction applications, this last feature has been addressed by employing non-convex penalty functions (see e.g. [11, 6]). An example of such a penalty function is shown in Fig. 1a. Here, the non-convex function, referred to as the log-penalty is $\phi(z) = \ln(1+|z|)$. This penalty function has a discontinuity in its derivative, like the absolute value function, at the origin. However, it penalizes high values of $|x|$ less than the absolute value function. The 'threshold function' associated with $\phi(z)$ is denoted by $T(\cdot)$ and is defined so that $\hat{x} = T(z)$ for

$$\hat{x} = \underset{x}{\operatorname{argmin}} \ \frac{1}{2}(z-x)^2 + \phi(x). \qquad (3)$$

The threshold function as defined above is also referred to as the 'proximity operator' of ϕ [3][1]. In contrast to the soft threshold function, this threshold function converges to the identity asymptotically. This in turn ensures that high-valued estimates obtained by this threshold function are less biased, compared to the estimates given by the soft-threshold.

In view of the forgoing discussion, we propose to replace the ℓ_1 term in the fused lasso cost function in (1) with a coordinate-

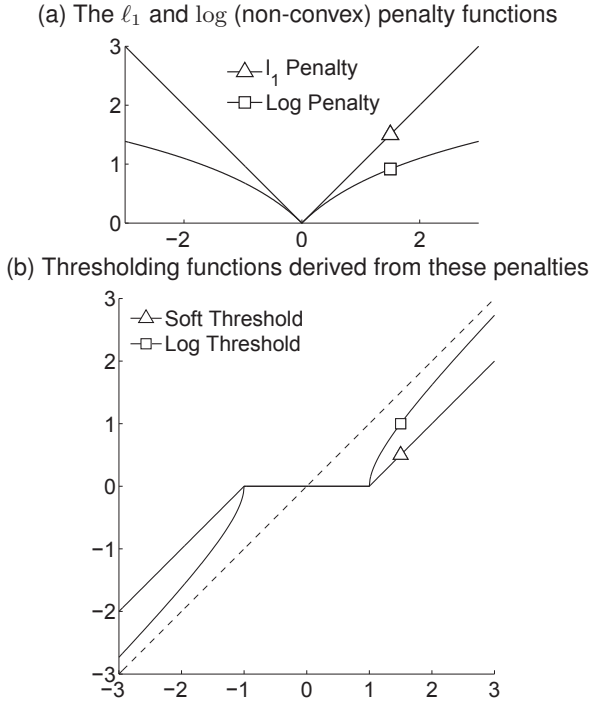

Fig. 1: (a) The absolute value function (ℓ_1 penalty) and the log penalty as an instance of a non-convex penalty. (b) The denoising, or thresholding operators that solve a denoising problem when the functions in (a) are used as regularizers. Note that the bias of the log-threshold decreases with increasing values of the input.

E-mail : ibayram@itu.edu.tr, poyupaulchen@gmail.com, selesi@poly.edu
This work was supported by the NSF under Grant No. CCF-1018020.

[1] Actually, the definition in [3] requires that ϕ be lower semi-continuous, convex. We use the term rather formally here.

wise non-convex penalty ϕ as above. More precisely, we propose to estimate x as,

$$\hat{x} = \underset{x}{\operatorname{argmin}} \frac{1}{2} \|y - x\|_2^2 + \lambda_1 \ \mathrm{TV}(x) + \lambda_2 \sum_{i=1}^{N} \phi(x_i). \quad (4)$$

Although this replacement is motivated from a theoretical point of view, the two step procedure (TV denoise + soft threshold) could still be a reason to favor the ℓ_1 penalty, from a computational point of view. A natural question is then, whether a similar two step procedure exists for non-convex penalties as well. That is, if we first TV-denoise and then apply to the result the threshold function associated with ϕ, do we obtain the same result as $\hat{x}$ in (4)? We show in this paper that this is indeed the case under certain conditions. The two step procedure outlined above for the ℓ_1-penalty/soft-threshold pair also applies for ϕ and its threshold function (proximity operator) T, under mild assumptions on ϕ.

Outline

In Section 2 we show that the two step procedure above extends to non-convex penalties under some conditions. Experiments demonstrating the utility of the proposed formulation are given in Section 3. Section 4 contains some concluding remarks.

2. EXTENSION TO COMPONENTWISE MONOTONE DENOISING OPERATORS

In this section, we show that the two step procedure, consisting of TV denoising followed by soft thresholding can be modified to handle cases where the ℓ_1 penalty is replaced by a nonconvex penalty term. More precisely, we consider the problem in (4). Let the TV-denoised input be denoted as $\hat{z}$, that is

$$\hat{z} = \underset{z}{\operatorname{argmin}} \frac{1}{2} \|y - z\|_2^2 + \lambda_1 \ \mathrm{TV}(z) \quad (5)$$

Suppose now we apply to $\hat{z}$ the threshold function associated with $\lambda\,\phi(\cdot)$, that is,

$$\tilde{x} = \underset{x}{\operatorname{argmin}} \frac{1}{2} \|\hat{z} - x\|_2^2 + \lambda_2 \sum_i \phi(x_i) \quad (6)$$

In this setting, we have the following result.

Proposition 1. If $\phi : \mathbb{R} \to \mathbb{R}$ is a function and $\lambda_2 \in \mathbb{R}$ is a constant such that

$$\frac{\alpha}{2} z^2 + \lambda_2\, \phi(z) \quad (7)$$

is convex for some $0 < \alpha < 1$, then, $\tilde{x}$ in (6) and $\hat{x}$ in (4) are equal. □

In the rest of this section, we provide a proof of this result. For this, we need a few definitions from convex analysis. We refer to [8, 10] for a more detailed account.

2.1. Preliminaries

Definition 1. The *subdifferential* of a convex function $f : \mathbb{R}^n \to \mathbb{R}$, is a set-valued map, denoted by ∂f, and is defined for $x \in \mathbb{R}^n$ as,

$$\partial f(x) = \{t \in \mathbb{R}^n : f(z) \geq f(x) + \langle t, z - x\rangle, \ \forall z \in \mathbb{R}^n\}. \quad (8)$$

We note that for differentiable f, the subdifferential ∂f is in 1-1 correspondence with the usual derivative or gradient. For $f(x) = \|x\|_1$, $\partial f(x)$ consists of vectors u such that,

$$u_i \in \begin{cases} \{-1\}, & \text{if } x_i < 0, \\ [-1, 1], & \text{if } x_i = 0, \\ \{1\}, & \text{if } x_i > 0. \end{cases} \quad (9)$$

Proposition 2. If f and g are convex functions, then $\partial(f + g) = \partial f + \partial g$. □

Proposition 3. A point $x \in \mathbb{R}^n$ minimizes f if and only if '$0 \in \partial f(x)$'. □

Definition 2. A set valued mapping $T(\cdot)$ defined on $\mathbb{R}^n$ is said to be *monotone* if

$$\langle x - y, u - v\rangle \geq 0 \quad (10)$$

for all x, y in $\mathbb{R}^n$ and $u \in T(x)$, $v \in T(y)$.

For $n = 1$, the monotonicity of the operator implies that if $x < y$ and $u \in T(x)$, $v \in T(y)$, then the inequality $u \leq v$ holds.

Proposition 4. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a lower semi-continuous, convex function. Also, let T_β denote the operator defined as,

$$T_\beta(y) = \underset{z}{\operatorname{argmin}} \frac{\beta}{2} \|y - z\|_2^2 + f(z). \quad (11)$$

Then, for $\beta > 0$, T_β is a (single-valued) monotone operator. □

We note that the operator defined in Prop. 4 is also known as the proximity operator of f [3].

2.2. Proof of Proposition 1

As required by the hypothesis of Prop. 4, suppose now that $\phi : \mathbb{R} \to \mathbb{R}$ is a penalty function such that

$$\frac{\alpha}{2} z^2 + \lambda_2\, \phi(z) \quad (12)$$

is convex for some $0 < \alpha < 1$. In that case,

$$F(x) = \frac{1}{2} \|x\|_2^2 + \lambda_2 \sum_i \phi(x_i) \quad (13)$$

will be strictly convex. This in turn implies that the cost function in (4) is strictly convex (and guarantees the existence of a unique minimum). Moreover, by Prop. 4, the operator which maps $y \in \mathbb{R}$ to $\hat{z} \in \mathbb{R}$ through

$$\hat{z} = \underset{z}{\operatorname{argmin}} \frac{1}{2} (y - z)^2 + \lambda_2\, \phi(z) \quad (14)$$

is single-valued and monotone. Suppose now that for $y \in \mathbb{R}^n$, we define an operator as,

$$T(y) = \underset{t}{\operatorname{argmin}} \sum_i \left(\frac{1}{2}(y_i - t_i)^2 + \lambda_2\, \phi(t_i) \right). \quad (15)$$

Note that from (5), (6), we have that $\tilde{x} = T(\hat{z})$. Thanks to the separability of the problem and the componentwise monotonicity of T, we obtain,

Lemma 1. Suppose $x = T(t)$ for some t. In that case,

(i) if $x_i < x_{i+1}$ then $t_i < t_{i+1}$,

(ii) if $x_i > x_{i+1}$ then $t_i > t_{i+1}$.

In words, the output of the threshold function,

(i) increases only if its input increases,

(ii) decreases only if its input decreases.

Consequently,

$$\partial\big(\|\cdot\|_1\big)\big|_{D\,x} \supset \partial\big(\|\cdot\|_1\big)\big|_{D\,t}\,, \tag{16}$$

where D is the $(N-1)\times N$ matrix defined as,

$$D = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ & & \ddots & & \\ 0 & \dots & 0 & 1 & -1 \end{bmatrix}. \tag{17}$$

(That is, the subdifferential of $\|\cdot\|_1$ evaluated at '$D\,t$' is a subset of the subdifferential of $\|\cdot\|_1$ evaluated at '$D\,x$'.)

Proof. Parts (i) and (ii) follow directly from the preceding discussion. Now let $d = D\,x$, $e = D\,t$. Parts (i), (ii), imply that if d_i is non-zero then e_i is non-zero and both share the same sign. By (9) then, if $u \in \partial\big(\|\cdot\|_1\big)\big|_{D\,x}$ we also have $u \in \partial\big(\|\cdot\|_1\big)\big|_{D\,t}$. □

Let us now turn to (5). By Props. 2, 3 and (9), we have in this case,

$$0 = \hat{z} - y + \lambda_1\, D^T\, u \tag{18}$$

for some $u \in \partial\big(\|\cdot\|_1\big)\big|_{D\,\hat{z}}$.

Similarly, by (6), we can write

$$0 = \tilde{x} - \hat{z} + \lambda_2\, v \tag{19}$$

for some $v \in \partial\big(\sum_i \Phi(\cdot)\big)\big|_{\tilde{x}}$, where $\Phi(x) = \sum_i \phi(x_i)$.

Adding (18) and (19), and recalling that $\partial\big(\|\cdot\|_1\big)\big|_{D\,\tilde{x}} \supset \partial\big(\|\cdot\|_1\big)\big|_{D\,\hat{z}}$, which follows by Lemma 1, we obtain

$$0 = \tilde{x} - y + \lambda_1\, D^T\, u + \lambda_2\, v \tag{20}$$

for some $u \in \partial\big(\|\cdot\|_1\big)\big|_{D\,\tilde{x}}$ and $v \in \partial\big(\sum_i \Phi(\cdot)\big)\big|_{\tilde{x}}$. In words, $\tilde{x}$ minimizes (4). Since $F(x)$ in (13) and hence the cost in (4) is strictly convex, $\tilde{x}$ is actually the unique minimizer of (4).

2.3. The Condition on ϕ

Note that the condition on ϕ in Prop. 1 restricts the family of non-convex penalties to which this result applies. For instance, $\phi(x) = \sqrt{|x|}$ violates this condition, independently of λ_2. Nevertheless, there are useful and interesting candidates in the set of allowed functions. Two examples, for $\lambda_2 \leq 1$ are,

$$\phi_1(x) = \ln(1+|x|), \tag{21}$$

$$\phi_2(x) = \frac{2}{\sqrt{3}}\left(\operatorname{atan}\left(\frac{1+2|x|}{3}\right)\right), \tag{22}$$

which have been previously used in [11]. Both of these functions are non-convex and symmetric. They both have a singularity at the origin, which leads to the dead zone in the associated threshold function. Also, on the positive part of the real line, their derivative monotonely decreases, which in turn leads to convergence of the threshold function to the identity asymptotically. A curious difference is that

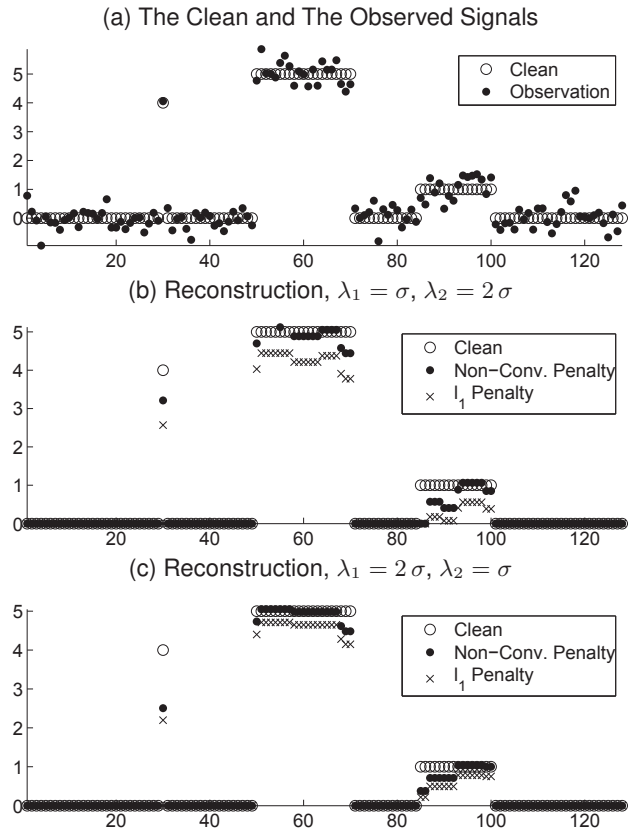

Fig. 2: (a) Clean piece-wise constant, sparse signal and the observed noisy signal, (b,c) Reconstruction with the fused lasso and the non-convex fused lasso using different regularization parameters.

$\phi_1(x)$ increases without bound whereas $\phi_2(x)$ is bounded. This in turn leads to a threshold function for $\phi_2(x)$ that converges to the identity faster than the threshold function of $\phi_1(x)$. However, the threshold function for $\phi_1(x)$ has a closed form expression, but that is not the case for the threshold function of $\phi_2(x)$ [11].

The condition in Prop. 1 allows non-symmetric non-convex functions to be used as penalty functions. Such choices could be of interest if positive values are more probable than negative values, for instance. Other than this, the function may have more than one discontinuity in its first derivative, as in the SCAD function, proposed by Fan and Li [6].

3. EXPERIMENTS

Experiment 1. In the Introduction, we noted that the soft threshold causes the high-magnitude estimates to be biased towards zero, and one way to avoid this bias is to employ non-convex penalties instead of the ℓ_1 penalty. In order to test this claim, we conducted an experiment. The clean signal x which is piece-wise constant and sparse is shown in Fig. 2a. We add Gaussian noise to this signal to obtain the noisy observation with an SNR of 15dB. As the non-convex penalty, we use the function (recall (21))

$$\phi(x) = a^{-1}\ln\big(1+a|x|\big) \tag{23}$$

with $a = 2$. The threshold function associated with this function is given by [11],

$$T_\lambda(y) = \begin{cases} \left[\frac{|y|}{2} - \frac{1}{2a} + \sqrt{\left(\frac{|y|}{2} + \frac{1}{2a}\right)^2 - \frac{\lambda}{a}}\right] \mathrm{sgn}(y), & |y| \geq \lambda, \\ 0, & |y| < \lambda. \end{cases}$$

We experimented with different choices of the regularization parameters. These choices are $(\lambda_1, \lambda_2) = (\sigma, 2\sigma)$, and $(\lambda_1, \lambda_2) = (2\,\sigma, \sigma)$, where σ is the standard deviation of the noise.

The estimates obtained by solving the fused lasso problem in (1) and the proposed modified problem in (4) are shown in Fig. 2b for $(\lambda_1, \lambda_2) = (\sigma, 2\sigma)$. We observe that the estimate with the ℓ_1 penalty is biased towards zero, whereas the bias is seen to be reduced for the non-convex penalty. This is expected, since the threshold function for the non-convex penalty is indeed closer to the identity as the magnitude of the inputs increase. In addition, we can see that the difference in bias between the two methods increases as the magnitude of the signals increase. For high values, the soft-threshold introduces a clear bias, which is prevented significantly by employing the non-convex penalty.

The estimates with $(\lambda_1, \lambda_2) = (2\sigma, \sigma)$ are shown in Fig. 2c. Especially for the nonconvex penalty, we see that the constant pieces are estimated with a low bias. However, the estimate of the isolated component is rather poor, primarily because TV denoising pulls this value towards zero. In fact, this effect could be reduced by employing a non-convex penalty on the differences of the neighboring coefficients, instead of the regular total variation as used in this paper, which computes the ℓ_1 norm of these differences. Although the resulting problem can be shown to be convex, as in the modification proposed in this paper, the simple two step procedure does not extend to that case – an iterative algorithm is necessary to solve such a problem.

Experiment 2. In a second experiment, we employ the fused lasso for audio denoising. The spectrogram of the noisy signal is shown in Fig. 3a. Given this spectrogram, with an SNR of 5 dB, we applied fused lasso denoising on the magnitudes of each subband using $(\lambda_1, \lambda_2) = (\sigma, \sigma)$. The resulting modified spectrogram magnitude is shown in Fig. 3b. To obtain the time-domain signal, we added back the noisy phase as is usual in audio denoising [2]. The resulting SNR is 9.24 dB. Then we replaced the ℓ_1 penalty with the non-convex penalty '$\lambda_2 \ln(1 + \lambda_2^{-1}|x|)$' (see Experiment 1 for the associated threshold/proximity operator) and repeated the same procedure. The resulting spectrogram is shown in Fig. 3c. Although the two spectrograms look very similar, the resulting SNR is 11.13 dB, about two decibels higher than that obtained by the regular fused lasso. This is because the ℓ_1 penalty suppresses the magnitudes of harmonics, just as in the previous experiment, which is not easy to observe directly from the spectrograms. However, a close look reveals that there is a change in the gray levels. To make this more apparent, we show in Fig. 3d the ratios of the two denoised spectrograms (the one obtained with the modified formulation being on the numerator). Note that, if the two spectrograms were the same, the ratio would be unity (and the figure black). However, the suppression of the ℓ_1 fused lasso is such that the ratio is below -1 dB for most of the time-frequency plane, hence the difference in the resulting SNRs.

4. CONCLUSION

The fused lasso penalty consists of the sum of a TV term and an ℓ_1 penalty. Although the ℓ_1 penalty leads to sparse estimates, it introduces bias for the non-zero coefficients. This bias can be reduced by replacing the ℓ_1 penalty with a non-convex penalty. We show in this case that the resulting cost function for the denoising case can be solved with fast, finite-terminating algorithms, as is the case for the original fused lasso.

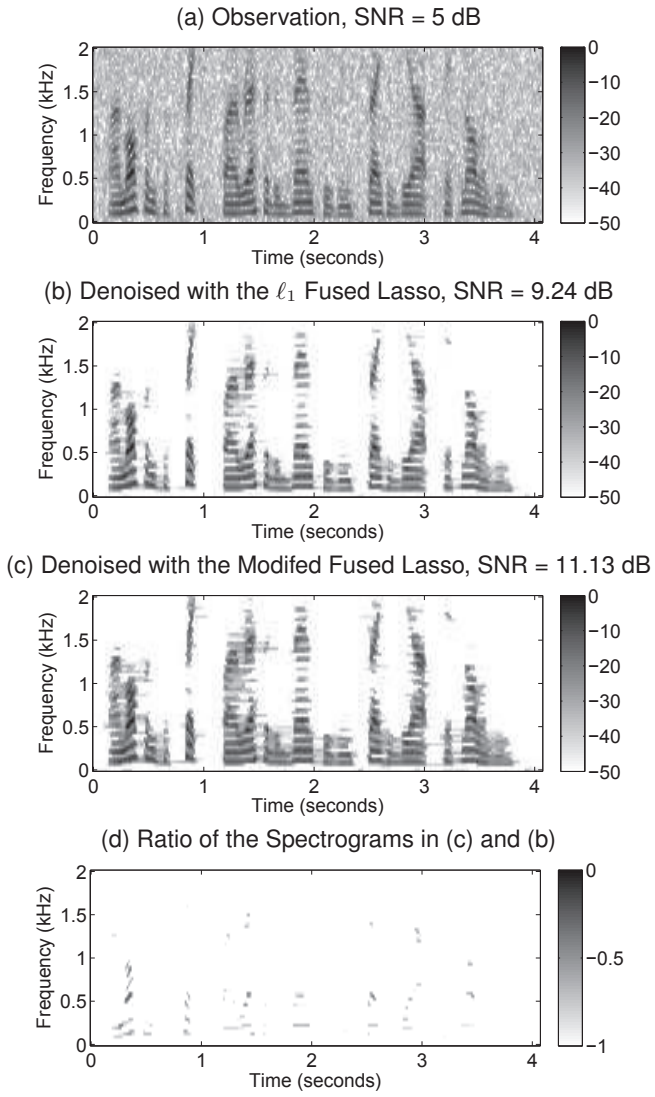

Fig. 3: Spectrograms from Experiment 2. (a) Observed noisy spectrogram. Spectrogram denoised with the fused lasso, using (b) the ℓ_1 penalty, (c) the non-convex 'log'-penalty. (d) the ratio of the two spectrograms in (c) and (b).

The TV penalty may be regarded as the ℓ_1 norm of the derivative of the input. Due to the existence of the ℓ_1 norm also in the TV term, we could similarly argue in favor of modifying the definition of TV so as to reduce the bias of the non-zero piecewise constant segments. However, in that case, the two step procedure that leads to a finite-terminating denoising scheme does not apply. Nevertheless, it would be of interest to employ such a penalty in more general problems that require iterative solutions.

5. REFERENCES

[1] İ. Bayram. A divide-and-conquer algorithm for 1D total variation denoising. `http://web.itu.edu.tr/ibayram/TVDenoise/TVDenoise.pdf`. Manuscript, 2013.

[2] I. Cohen and S. Gannot. Spectral enhancement methods. In J. Benesty, M. M. Sondhi, and Y. Huang, editors, *Springer Handbook of Speech Processing*. Springer, 2008.

[3] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, New York, 2011.

[4] L. Condat. A direct algorithm for 1D total variation denoising. *IEEE Signal Processing Letters*, 20(11):1054–1057, November 2013.

[5] P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *The Annals of Statistics*, 29(1):1–48, February 2001.

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

[7] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, December 2007.

[8] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2004.

[9] E. Mammen and S. van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, February 1997.

[10] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

[11] I. W. Selesnick and İ. Bayram. Sparse signal estimation by maximally sparse convex optimization. *IEEE Trans. Signal Processing*, 62(5):1078–1092, March 2014.

[12] 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*, 67(1):91–108, 2005.