Computational Statistics and Data Analysis 136 (2019) 59–71

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Alternating direction method of multipliers for nonconvex fused regression problems

Xianchao Xiu [a,*], Wanquan Liu [c], Ling Li [c], Lingchen Kong [b]

[a] State Key Lab for Turbulence and Complex Systems, Department of Mechanics and Engineering Science, College of Engineering, Peking University, China

[b] Department of Applied Mathematics, Beijing Jiaotong University, Beijing, PR China

[c] Department of Computing, Curtin University, Perth, WA, Australia

ARTICLE INFO

Article history:
Received 5 June 2017
Received in revised form 19 November 2018
Accepted 4 January 2019
Available online 11 January 2019

Keywords:
Fused LASSO
Alternating direction method of multipliers
Variable selection
Nonconvex optimization

ABSTRACT

It is well-known that the fused least absolute shrinkage and selection operator (FLASSO) has been playing an important role in signal and image processing. Recently, the nonconvex penalty is extensively investigated due to its success in sparse learning. In this paper, a novel nonconvex fused regression model, which integrates FLASSO and the nonconvex penalty nicely, is proposed. The developed alternating direction method of multipliers (ADMM) approach is shown to be very efficient owing to the fact that each derived subproblem has a closed-form solution. In addition, the convergence is discussed and proved mathematically. This leads to a fast and convergent algorithm. Extensive numerical experiments show that our proposed nonconvex fused regression outperforms the state-of-the-art approach FLASSO.

1. Introduction

The linear regression model has broad applications such as in astrophysics, signal and image processing, statistical inference, and optics, just to mention a few here. It assumes that there is a linear relationship between the design matrix $X \in \mathbb{R}^{n\times p}$ and the observation $y \in \mathbb{R}^n$, i.e.,

$$y = X\beta + \epsilon, \tag{1}$$

where $\beta \in \mathbb{R}^p$ is the vector of objective regression coefficients, and $\epsilon \in \mathbb{R}^n$ is an error vector which is assumed to follow the normal distribution $N(0, \sigma^2 I)$ with mean 0 and standard deviation σ.

However, when the number of variables is much larger than the number of observations, i.e., $n \ll p$, the classical statistical methods cannot work efficiently. To perform variable selection and model fitting, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) model:

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|^2 + \tau\|\beta\|_1, \tag{2}$$

where $\|\cdot\|_1$ is defined as the sum of absolute values of all entries, and τ is a tuning parameter.

* Corresponding author.
E-mail addresses: xcxiu@bjtu.edu.cn (X. Xiu), W.Liu@curtin.edu.au (W. Liu), L.Li@curtin.edu.au (L. Li), lchkong@bjtu.edu.cn (L. Kong).

https://doi.org/10.1016/j.csda.2019.01.002
0167-9473/

One drawback of (2) is the fact that it ignores ordering relationship among the coefficients. For this purpose, Tibshirani et al. (2005) considered the fused least absolute shrinkage and selection operator (FLASSO) model:

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|^2 + \tau_1\|\beta\|_1 + \tau_2 \sum_{i=1}^{p} |\beta_{i+1} - \beta_i|, \tag{3}$$

where $\tau_1 > 0$ and $\tau_2 > 0$ are two tuning parameters. The first penalty term encourages sparsity in the coefficients, and the second penalty term encourages sparsity in the differences of coefficients, which is usually called the fusion term. That is to say, this FLASSO perfectly captures the case where each coefficient is within an interval while there are jumps among these intervals. The FLASSO is widely used in gene expression in Tibshirani and Wang (2007), image denoising in Friedman et al. (2007), computer vision in Xin et al. (2015), and so on. When $\tau_2 = 0$, problem (3) degenerates into problem (2). We would also like to point out that when $\tau_1 = 0$, problem (3) degenerates into the total variation problem in Rudin et al. (1992); Rudin and Osher (1994). The total variation is widely used in image processing because it is able to improve the smoothness of boundary and trajectory that are usually the most important for image recovery, see Ng et al. (1999) for example.

Recently, the nonconvex penalty has been drawn more and more attention in sparse learning problems, because it is more efficient to extract the essential features of solutions than the ℓ_1-norm. At the same time, the nonconvex penalty has the nearly unbiased property, and could overcome the shortcomings of the ℓ_1-norm as discussed in Fan et al. (2014). An interesting question naturally comes out to us: can we extend the FLASSO model (3) in the nonconvex framework? Therefore, in this paper, we will consider the following model:

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\beta) + \sum_{i=1}^{p} \Phi_{\tau_2}(\beta_{i+1} - \beta_i). \tag{4}$$

where Φ is the nonconvex penalty functions with parameters τ_1 and τ_2, see Section 3.2. As this model is motivated by the FLASSO with nonconvex penalty functions, we call it the nonconvex fused regression in this paper.

Our contributions in this paper can be summarized as the following three parts:

1. To the best of our knowledge, we are the first to model the FLASSO in the nonconvex framework. The FLASSO is utilized to characterize the sparsity for not only the coefficients but also their successive differences, and the nonconvex penalty function is employed to improve the accuracy of variable selection.
2. We develop an efficient algorithm using the alternating direction method of multipliers to optimize our proposed model. With the help of the Kurdyka-Łojasiewicz function (see Section 3.1), we show that when the penalty parameters are chosen above a threshold and thus the sequence generated by the algorithm gives a stationary point of our proposed model.
3. Numerical experiments, including synthetic data and real-world data, are conducted to demonstrate that our proposed model can achieve better performance than the previous FLASSO.

The remaining of this paper is organized as follows. Some related work and preliminaries will be reviewed in Sections 2 and 3. In Section 4, the optimization algorithm and the convergence analysis will be presented. In Section 5, extensive experiments will be conducted to substantiate the superiority of the proposed model over the other existing one. This paper will be concluded in Section 6.

2. Related work

With the increasing prominence of large-scale data in modern science, how to solve the FLASSO model (3) efficiently becomes very important. From numerical point of view, problem (3) is hard to solve because the regularization terms in its objective function are nonsmooth. Even though some generic optimization approaches such as interior-point methods in Wright and Nocedal (1999) or subgradient descent schemes in Bertsekas (1999) and some standard convex optimization tools such as the SQOPT in Gill et al. (2008) or CVX in Glowinski and Marroco (1975) are applicable, they are usually very slow in convergence due to high computational complexity.

Alternating direction method of multipliers (ADMM) was originally proposed in Gabay and Mercier (1976) and Glowinski and Marroco (1975), and the convergence was analyzed in Gabay (1983) and He and Yuan (2012). Recently, Boyd et al. (2011) recommended to apply alternating direction method of multipliers (ADMM) to solve the LASSO model (2), and the efficiency was well demonstrated. The proposed ADMM approach takes advantage of the structure of the ℓ_1-norm instead of using a smoothing term to approximate the nonsmooth ℓ_1-norm term. Consequently, the resulting subproblems have closed-form solutions. This is the key idea to guarantee the efficiency of the ADMM when solving (2). But, when it comes to the FLASSO model (3), the ADMM cannot be applied directly. The reason behind it is the resulting subproblem related to the fusion term does not admit a closed-form solution. Thus, an inner iteration is required to pursue an approximate solution for this subproblem. Later, Li et al. (2014) embedded the linearization technique into the ADMM approach, called the LADMM, and then all the subproblems can have closed-form solutions.

However, when ADMM is applied to nonconvex optimization problems, the convergence cannot be guaranteed. Recently, Wang et al. (2015) considered a general type of nonconvex nonsmooth problems. To solve this type of problems,

they considered a variant of the ADMM approach whose subproblems are simplified by adding a Bregman proximal term. Later, Hong et al. (2016) considered the nonconvex problems with one nonconvex function as well as a bunch of convex functions. The convergence is guaranteed when the penalty parameter is chosen above a computable threshold, which is not easy to check in practice. Yang et al. (2017) studied a special nonconvex problem, including one convex function and one possibly nonconvex function. They also established the convergence of the generated sequence with mild conditions. Very recently, Guo et al. (2017) proved that if the augmented Lagrangian function is a Kurdyka-Łojasiewicz (KL) function and the separated two nonconvex functions satisfy the Lipschitz condition, then the sequence generated by ADMM converges to a KKT point. But, our optimization problem (4) contains two nonconvex penalty functions without requirement of Lipschitz property, which is different from the above convergence analysis.

In this paper, we attempt to apply ADMM to solve our nonconvex fused regression model (4). In algorithm, by introducing two new variables, all the resulting subproblems can admit closed-form solutions. In theory, as the Lipschitz property is not required, the proof is different from Guo et al. (2017). We attempt to derive the convergence result by integrating the techniques in Hong et al. (2016) and Yang et al. (2017).

3. Preliminaries

In this section, we briefly review some preliminaries. One is the notations and definitions, which will be used throughout this paper. The other one is the nonconvex penalty, which motivates us to describe the sparsity in the nonconvex framework.

3.1. Notations and definitions

In this paper, we use $\mathbb{R}^n$ to denote the n-dimensional Euclidean space. For a vector $x \in \mathbb{R}^n$, let x_i denote its ith entry. The Euclidean norm is denoted by $\|x\|$, the ℓ_1-norm is denoted by $\|x\|_1 = \sum_{i=1}^n |x_i|$, and the ℓ_p quasi-norm $(0 < p < 1)$ is denoted by $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. Furthermore, for two vectors x and y of the same size, we denote their inner product by $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$. In addition, for a matrix $X \in \mathbb{R}^{m\times n}$, we use X^T to denote its transpose. For a symmetric matrix $X \in \mathbb{R}^{n\times n}$, we use $\lambda_{\max}$ and $\lambda_{\min}$ to denote the largest and smallest eigenvalue, respectively. Finally, the domain of function f is $\mathrm{dom} f = \{x \in \mathbb{R}^n : f(x) < \infty\}$, and $\mathrm{dist}(x, \Omega)$ denotes the distance from x to Ω, i.e., $\mathrm{dist}(x, \Omega) = \inf_{y\in\Omega} \|x - y\|$.

We recall from Rockafellar and Wets (2009) that for a lower semicontinuous function $f : \mathbb{R}^n \to [-\infty, \infty]$ and a point $x \in \mathbb{R}^n$ where $f(x)$ is finite, the (limiting) subdifferential is defined as

$$\partial f = \left\{ v \in \mathbb{R}^n : \exists x^k \to x,\ v^k \to v \text{ with } v^k \in \hat{\partial} f(x^k)\ \forall k \right\},$$

where $\hat{\partial} f(x^k)$ denotes the Fréchet subdifferential of f at x^k, which is the set of all $v \in \mathbb{R}^n$ satisfying

$$\liminf_{z\to x^k} \frac{f(z) - f(x^k) - \langle v, z - x^k\rangle}{\|z - x^k\|} \geq 0.$$

From the above definition, we can easily observe that

$$\left\{ v \in \mathbb{R}^n : \exists x^k \to x,\ v^k \to v,\ v^k \in \partial f(x^k)\ \forall k \right\} \subseteq \partial f(x).$$

Notice that when f is convex, the above subdifferential coincides with the classical concept of convex subdifferential of f. Moreover, from the generalized Fermat's rule, we know that if $x \in \mathbb{R}^n$ is a local minimizer of f, then $0 \in \partial f(x)$.

At the end of this subsection, we will present the Kurdyka-Łojasiewicz (KL) function defined in Attouch et al. (2010) and the uniformized KL property defined in Bolte et al. (2014). For simplicity, we use $\Xi_\eta(\eta > 0)$ to denote the class of nonconvex functions $\varphi : [0, \eta) \to \mathbb{R}_+$ satisfying: (1) $\varphi(0) = 0$; (2) φ is a continuous differentiable on $(0, \eta)$ and continuous at 0; (3) $\varphi'(0) > 0$ for all $x \in (0, \eta)$. Then the KL function can be described as follows.

Definition 3.1. Let f be a proper lower semicontinuous function.

1. For $\bar{x} \in \mathrm{dom}\partial f = \{x \in \mathbb{R}^n : \partial f(x) \neq \emptyset\}$, if there exist an $\eta \in [0, \infty)$, neighborhood V of $\bar{x}$ and a function $\varphi \in \Xi_\eta$ such that for all $x \in V \cap \{x \in \mathbb{R}^n : f(\bar{x}) < f(x) < f(\bar{x}) + \eta\}$, it holds that
$$\varphi'(f(x) - f(\bar{x}))\,\mathrm{dist}(0, \partial f(x)) \geq 1,$$
then f is said to have the Kurdyka-Łojasiewicz (KL) property at x.
2. If f satisfies the KL property at each point of $\mathrm{dom}\partial f$, then f is called a KL function.

Based on the above KL function, one can give the following lemma, which states the uniformized KL property.

Lemma 3.2. *Suppose that f is a proper lower semicontinuous function, and Γ is a compact set. If $f = f^*$ on Γ for some constant f^* and satisfies the KL property at each point of Γ, then there exist $\varepsilon > 0$, $\eta > 0$ and $\varphi \in \Xi_\eta$ such that*

$$\varphi'(f(x) - f^*)dist(0, \partial f(x)) \geq 1$$

for all $x \in \{x \in \mathbb{R}^n : dist(x, \Gamma) < \varepsilon\} \cap \{x \in \mathbb{R}^n : f^* < f(x) < f^* + \eta\}$.

The KL function and the uniformized KL property will be used in the proof of our convergence in Section 4.2.

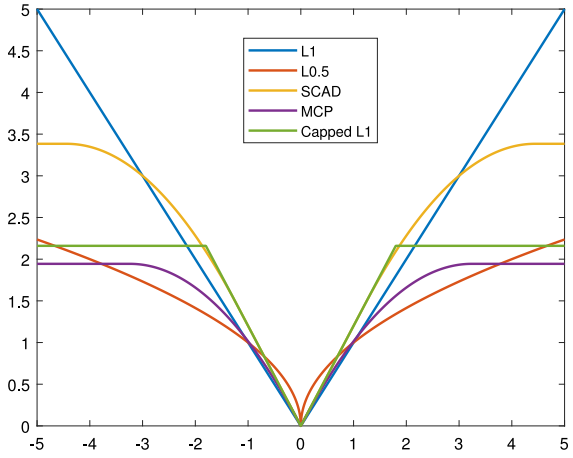

Fig. 1. Plots for ℓ_1, $\ell_{0.5}$, SCAD, MCP and capped ℓ_1.

Table 1
Nonconvex penalty functions and the associated shrinkage operators.

$\Phi_\lambda(x)$	$\text{Shrink}_\Phi(t,\lambda)$
$\lambda\|x\|^{0.5}$	$\frac{2}{3}t\left(1+\cos\left(\frac{2\pi}{3}-\frac{2}{3}\arccos\left(\frac{\lambda}{4}\left(\frac{\|t\|}{3}\right)^{-3/2}\right)\right)\right)$
$\begin{cases}\lambda\|x\| & \text{if } \|x\|\le\lambda\\ \frac{a\lambda\|x\|-0.5(\|x\|^2+\lambda^2)}{a-1} & \text{if } \lambda\le\|x\|\le a\lambda\\ \frac{\lambda^2(a+1)}{2} & \text{if } \|x\|\ge a\lambda\end{cases}$	$\begin{cases}0 & \text{if } \|t\|\le\lambda\\ \text{sgn}(t)(\|t\|-\lambda) & \text{if } \lambda\le\|t\|\le 2\lambda\\ \text{sgn}(t)\frac{(a-1)\|t\|-a\lambda}{a-2} & \text{if } 2\lambda\le\|t\|\le a\lambda\\ t & \text{if } \|t\|> a\lambda\end{cases}$
$\begin{cases}\lambda\|x\|-\frac{x^2}{2a} & \text{if } \|x\|\le a\lambda\\ \frac{a\lambda^2}{2} & \text{if } \|x\|> a\lambda\end{cases}$	$\begin{cases}0 & \text{if } \|t\|\le\lambda\\ \text{sgn}(t)\frac{a(\|t\|-\lambda)}{a-1} & \text{if } \lambda\le\|t\|\le a\lambda\\ x & \text{if } \|t\|> a\lambda\end{cases}$
$\begin{cases}\lambda\|x\| & \text{if } \|x\|\le a\lambda\\ a\lambda^2 & \text{if } \|x\|> a\lambda\end{cases}$	$\begin{cases}0 & \text{if } \|t\|\le\lambda\\ \text{sgn}(t)(\|t\|-\lambda) & \text{if } \lambda\le\|t\|\le(a+\frac{1}{2})\lambda\\ \text{sgn}(t)\frac{(2a-1)\lambda}{2} & \text{if } \|t\|=(a+\frac{1}{2})\lambda\\ t & \text{if } \|t\|>(a+\frac{1}{2})\lambda\end{cases}$

3.2. The nonconvex penalty

For a general nonconvex optimization problem

$$\min_x \ \frac{1}{2}(x-t)^2+\Phi_\lambda(x),$$

where $\Phi_\lambda(\cdot)$ is the nonconvex penalty function. Thus, the solution can be given by the following nonconvex shrinkage operator

$$x=\text{shrink}_\Phi(t,\lambda).$$

Fig. 1 plots $\ell_{0.5}$, SCAD, MCP, capped ℓ_1, and Table 1 reviews the associated closed-form solutions.

- The $\ell_{0.5}$ penalty in Xu et al. (2012) has the representiveness of ℓ_p quasi-norm with $p\in(0,1)$. Whenever $p\in[0.5,1)$, the smaller the p, the sparser the solution yielded by ℓ_p penalty, and, whenever $p\in(0,0.5]$, the performance of ℓ_p penalty has no significant difference.
- The SCAD penalty in Fan and Li (2001) is derived from the following qualitative requirements: it is singular at the origin to achieve sparsity and its derivative vanishes for large values so as to ensure unbiasedness. The condition $a>2$ ensures the well-posedness of the defined thresholding operator.
- Following the rationale of the SCAD, the MCP in Zhang et al. (2010) is derived, whose main idea is to minimize the maximum nonconvexity subject to the unbiasedness and feather selection constraints. Also, the condition $a>1$ ensures the uniqueness of the thresholding operator.
- The capped ℓ_1 penalty in Zhang (2010) is a linear approximation of the SCAD penalty. Theoretically, it can be viewed as a variant of the two-stage optimization problem: one first solves a standard LASSO problem and then solves a LASSO problem where the large coefficients are not penalized any more, thus leading to an unbiased model.

The numerical and theoretical results have shown that the nonconvex penalty performs better than the ℓ_1-norm in terms of estimation accuracy and consistency when applied to the high dimensional sparse modeling, see Chen et al. (2012), Fan et al. (2014), Liu et al. (2015) and Xiu et al. (2018) for illustrations.

4. Main results

In this section, we first describe the optimization algorithm to solve our proposed model in Section 4.1, then discuss its convergence in Section 4.2.

4.1. The optimization algorithm

For notational simplicity, we reformulate the nonconvex fused regression model (4) into

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\beta) + \Phi_{\tau_2}(D\beta) \tag{5}$$

with

$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p}.$$

It is easy to verify that if we choose the nonconvex penalty function Φ to be the ℓ_1-norm, then (5) degenerates to the classical FLASSO model (3).

By introducing the auxiliary variables α and γ, the extended FLASSO model (5) can be rewritten as

$$\begin{aligned} &\min_{\alpha,\gamma,\beta} \frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\alpha) + \Phi_{\tau_2}(\gamma) \\ &\text{s.t.} \quad \alpha = \beta \\ &\qquad \gamma = D\beta. \end{aligned} \tag{6}$$

The auxiliary variable α plays a similar role as β so that the objective is now about α and β. Also the variable γ is introduced to liberate β from D. Then the augmented Lagrangian function of (6) is

$$\begin{aligned} \mathcal{L}(\alpha, \gamma, \beta, w_1, w_2) = &\frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\alpha) + \Phi_{\tau_2}(\gamma) \\ &- w_1^T(\alpha - \beta) + \frac{\mu_1}{2}\|\alpha - \beta\|^2 \\ &- w_2^T(\gamma - D\beta) + \frac{\mu_2}{2}\|\gamma - D\beta\|^2, \end{aligned}$$

in which $w_1 \in \mathbb{R}^p$ and $w_2 \in \mathbb{R}^{p-1}$ are the Lagrange multipliers, and μ_1 and μ_2 are positive penalty parameters. It is difficult to simultaneously optimize all these variables. We therefore approximately solve this optimization problem by alternatively minimizing one variable with the others fixed. Under the framework of ADMM, the optimization problem of $\mathcal{L}$ with respect to each variable can be solved by the following subproblems. Next, we will show that each subproblem of the above framework either has a closed-form solution or can be solved by a fast solver.

Step 1. For variable α, optimizing $\mathcal{L}$ with respect to α can be written to

$$\alpha^{k+1} = \arg\min_{\alpha} \left\{ \Phi_{\tau_1}(\alpha) + \frac{\mu_1}{2}\|\alpha - \beta^k - w_1^k/\mu_1\|^2 \right\}.$$

From Section 3.2, this subproblem can be solved in element-wise by

$$\alpha^{k+1} = \text{shrink}_{\Phi}(\beta^k + w_1^k/\mu_1, \tau_1/\mu_1). \tag{7}$$

Step 2. Similarly, for variable γ, the subproblem of $\mathcal{L}$ with respect to γ can be solved by

$$\gamma^{k+1} = \text{shrink}_{\Phi}(D\beta^k + w_2^k/\mu_2, \tau_2/\mu_2). \tag{8}$$

Step 3. For variable β, the optimization subproblem of $\mathcal{L}$ with respect to β can be simplified to

$$(X^TX + \mu_1 I + \mu_2 D^T D)\beta = X^T y + \mu_1\alpha^{k+1} - w_1^k + \mu_2 D^T\gamma^{k+1} - D^T w_2^k.$$

Consequently, we have

$$\beta^{k+1} = (X^TX + \mu_1 I + \mu_2 D^T D)^{-1}(X^T y + \mu_1\alpha^{k+1} - w_1^k + \mu_2 D^T\gamma^{k+1} - D^T w_2^k). \tag{9}$$

Notice that the denominator in the equation can be pre-calculated outside the main loop, avoiding extra computational cost. Indeed, it is positive definite, which can be efficiently solved by Cholesky decomposition.

Step 4. For dual variables w_1 and w_2, according to the ADMM, the multipliers associated with $\mathcal{L}$ are updated by the following formulas

$$\begin{aligned} w_1^{k+1} &= w_1^k - \mu_1(\alpha^{k+1} - \beta^{k+1}), \\ w_2^{k+1} &= w_2^k - \mu_2(\gamma^{k+1} - D\beta^{k+1}). \end{aligned} \tag{10}$$

Thus, the proposed algorithm for (6) can now be summarized in Algorithm 1.

Algorithm 1 ADMM for solving (6).

Input: design matrix X, parameters τ_1, τ_2
Output: optimal solution (α, γ, β)
Initialize: penalty parameters $\mu_1,\ \mu_2$, primal variables $(\alpha^0, \gamma^0, \beta^0) = (0, 0, 0)$, and dual variable $(w_1^0, w_2^0) = (0, 0)$
while not converge **do**
1: Update α^{k+1} according to (7)
2: Update γ^{k+1} according to (8)
3: Update β^{k+1} according to (9)
4: Update w_1^{k+1} and w_2^{k+1} via (10)
end while

Compared with the LADMM in Li et al. (2014), our proposed Algorithm 1 introduces two auxiliary variables, and avoids the linearization. Furthermore, all the resulting subproblems have closed-form solutions, which make this algorithm efficient to solve large-scale problems. In the next subsection, we will discuss its convergence result. This leads to a fast and convergent algorithm.

4.2. Convergence analysis

In this subsection, we will discuss the convergence for Algorithm 1, and the detailed proofs can be found in the supplementary material. Now, we present the first convergence result, which characterizes the cluster point of the sequence.

Theorem 4.1. *Suppose that $\{(\alpha^k, \gamma^k, \beta^k, w_1^k, w_2^k)\}$ is a sequence generated by Algorithm 1, then any cluster point $(\alpha^*, \gamma^*, \beta^*, w_1^*, w_2^*)$ of the sequence is a stationary point of (5).*

With the help of KL function in Attouch et al. (2010) and the uniformized KL property in Bolte et al. (2014), we will show in the next theorem that the whole sequence generated by Algorithm 1 is convergent.

Theorem 4.2. *Suppose that $\{(\alpha^k, \gamma^k, \beta^k, w_1^k, w_2^k)\}$ is a sequence generated by Algorithm 1, then the sequence converges to a stationary point of (5).*

Note that our convergence analysis is different from Guo et al. (2017) since we do not need the Lipschitz condition. In fact, we integrate the techniques in Hong et al. (2016) and Yang et al. (2017), which contains one nonconvex function. In our case, we have two nonconvex functions and our functions do not satisfy the Lipschitz condition. The proof includes the following four technical parts:

1. Show that the corresponding augmented Lagrangian sequence is decreasing;
2. Prove that the sequence is also bounded;
3. Give the convergence result of the sequence;
4. With the help of KL function, develop the convergence of the whole sequence.

5. Numerical experiments

In this section, we will conduct some experiments on synthetic data and real-world data to demonstrate the superiority of our proposed nonconvex fused regression (NonFuReg for short) model (5) over the FLASSO model (3). All the experiments are performed using MATLAB (R2017a) on a desktop computer with an Intel Core i5-3570M CPU with 3.4 GHz and 8 GB of memory.

For the FLASSO, we choose the method used in Li et al. (2014). The results are generated from the source codes provided by their authors. For the nonconvex fused regression model, we choose Φ to be the SCAD with $a = 3.7$. For all test algorithms, the maximum iteration number is set as 5000, and the stopping criterion is set to be Tol $< 10^{-3}$, which is defined as

$$\text{Tol} = \frac{\|\beta^k - \beta^{k-1}\|}{\max\{\|\beta^k\|, 1\}}.$$

Table 2
Selected results for the FLASSO and the nonconvex fused regression under Gaussian noise. The best results are highlighted in bold.

σ	Methods	$\sharp\{\ \lvert\beta_i-\beta_i^*\rvert< 0.1,\ i\in G\}$	$\max_{i\in G}\lvert\beta_i-\beta_i^*\rvert$	$\sharp\{\ \lvert\beta_i\rvert< 0.1,\ i\in G^c\}$	$\max_{i\in G^c}\lvert\beta_i\rvert$
0.001	FLASSO	2233	0.0835	327	0.6821
	NonFuReg	**2238**	**0.0819**	**322**	**0.5386**
0.01	FLASSO	2230	0.0973	330	0.5158
	NonFuReg	**2240**	**0.0685**	**320**	**0.4925**
0.1	FLASSO	2231	0.0994	329	0.7047
	NonFuReg	**2238**	**0.0984**	**322**	**0.6425**
1	FLASSO	2228	0.0960	332	0.5289
	NonFuReg	**2236**	**0.0731**	**324**	**0.4307**

Table 3
The SNR (dB) under Gaussian noise.

σ	0.001	0.01	0.1	1
SNR	63.1062	44.2116	24.3822	3.3969

5.1. Synthetic data

In this subsection, we test synthetic data to compare the FLASSO and the nonconvex fused regression. According to Li et al. (2014), we first generate the Gaussian matrix $X\in\mathbb{R}^{n\times p}$ with unit column norm randomly. Then, we divide p into 80 groups and randomly select 10 groups denoted as a sample set G, whose cardinality is g. Thus we obtain a sparse solution with close relationship among the successive coefficients. In addition, the coefficient vector β^* is generated by

$$\beta_i^*=\begin{cases} U[-3,3] & \text{if } i\in G\\ 0 & \text{otherwise,}\end{cases}$$

where $U[-3,3]$ represents the uniform distribution on the interval $[-3,3]$. Finally, we can get the observation data y by $y=X\beta^*+\varepsilon$ with $\varepsilon\sim n(0,\sigma^2 I)$. We test the case $\sigma=0.001, 0.01, 0.1, 1, 10$ with $(n,p,g)=(720,\ 2560,\ 320)$. As suggested in Li et al. (2014), we choose the same sparse parameter $\tau_1=0.5$ and the fusion parameter $\tau_2=0.5$ for the FLASSO and the nonconvex fused regression.

In order to evaluate the performance of the FLASSO and the nonconvex fused regression, we define some specific measurements:

- "$\sharp\{\ \lvert\beta_i-\beta_i^*\rvert<0.1,\ i\in G\}$" refers to how many β_i satisfies $\lvert\beta_i-\beta_i^*\rvert<0.1$ for $i\in G$, where "$\sharp$" denotes the number of the elements.
- "$\max_{i\in G}\lvert\beta_i-\beta_i^*\rvert$" refers to maximum difference between β_i and β_i^* for $i\in G$;
- "$\sharp\{\ \lvert\beta_i\rvert<0.1,\ i\in G^c\}$" refers to how many β_i satisfies $\lvert\beta_i\rvert<0.1$ for $i\in G^c$, where G^c denotes the complementary space of G;
- "$\max_{i\in G^c}\lvert\beta_i\rvert$" refers to maximum value of β_i for $i\in G^c$.

On one hand, we use the first two measurements to the proximity of the approximate solution β to the true solution β^*. The higher the "$\sharp\{\ \lvert\beta_i-\beta_i^*\rvert<0.1,\ i\in G\}$" and the lower the "$\max_{i\in G}\lvert\beta_i-\beta_i^*\rvert$", the better the recovery accuracy. On the other hand, we use the last two measurements to the value of the approximate solution β for the complementary space G^c. The lower the "$\sharp\{\ \lvert\beta_i\rvert<0.1,\ i\in G^c\}$" and the lower the "$\max_{i\in G^c}\lvert\beta_i\rvert$", the better the recovery accuracy.

We summarize the results in Table 2. One can observe that the nonconvex fused regression achieves the solution with higher quality to the true solution β^*. For example, the number of the recovered coefficients by the nonconvex fused regression for $\sigma=0.001$ is 10 more than that by the FLASSO, which shows that the nonconvex fused regression can reconstruct more precise coefficients; the maximum difference between the recovered coefficient β_i and true coefficient β_i^* by the nonconvex fused regression for $\sigma=0.01$ is 0.01 lower than that by the FLASSO, which illustrates that the nonconvex fused regression can reconstruct much closer coefficients. Also, the computational results of signal-noise ratio (SNR) are listed in Table 3. In this paper, the SNR is defined as

$$\text{SNR}=10\log_{10}(V_s/V_n),$$

where V_s represents the variance of the original signal β, V_n is the variance of the noise signal ε. In Fig. 2, we visualize the results for the FLASSO and the nonconvex fused regression for $\sigma=1$.

Next, we will compare the FLASSO with the nonconvex fused regression under non-Gaussian noise, i.e., logistic distribution. Conclusions similar to those under Gaussian noise can be made based on the results in Table 4. In addition, the results of SNR are presented in Table 5, and the visual results of the FLASSO and the nonconvex fused regression for the logistic noise case where $\sigma=10$ are shown in Fig. 5. All of them illustrate that the nonconvex fused regression provides a better recovery to the true solution β^* under non-Gaussian noise.

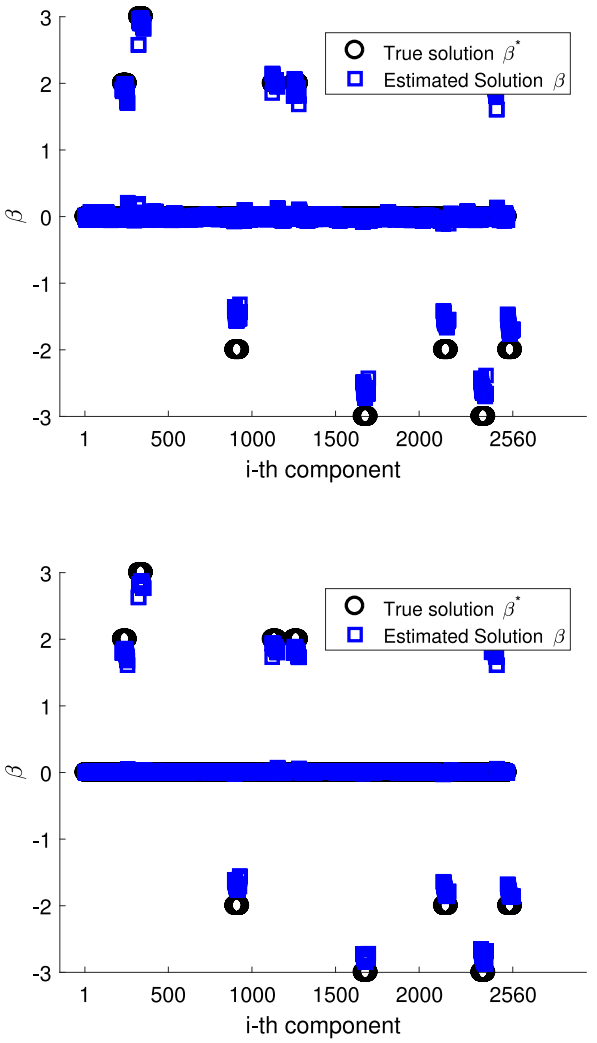

Fig. 2. Recovery results for the FLASSO (top) and the nonconvex fused regression (bottom) under Gaussian noise for $\sigma = 1$.

Table 4
Selected results for the FLASSO and the nonconvex fused regression under non-Gaussian noise. The best results are highlighted in bold.

σ	Methods	$\sharp\{\|\beta_i - \beta_i^*\| < 0.1,\ i \in G\}$	$\max_{i\in G}\|\beta_i - \beta_i^*\|$	$\sharp\{\|\beta_i\| < 0.1,\ i \in G^c\}$	$\max_{i\in G^c}\|\beta_i\|$
0.001	FLASSO	2224	0.0937	336	0.5850
	NonFuReg	**2236**	**0.0786**	**324**	**0.4588**
0.01	FLASSO	2220	0.0920	330	0.5273
	NonFuReg	**2240**	**0.0645**	**320**	**0.4931**
0.1	FLASSO	2216	0.0985	344	0.6832
	NonFuReg	**2240**	**0.0733**	**320**	**0.4413**
1	FLASSO	2210	0.0993	350	0.7166
	NonFuReg	**2236**	**0.0715**	**324**	**0.4145**
10	FLASSO	2204	0.0847	356	0.680
	NonFuReg	**2235**	**0.0523**	**325**	**0.548**

Table 5
The SNR (dB) under non-Gaussian noise.

σ	0.001	0.01	0.1	1	10
SNR	82.0378	68.4390	40.2846	24.8659	4.8134

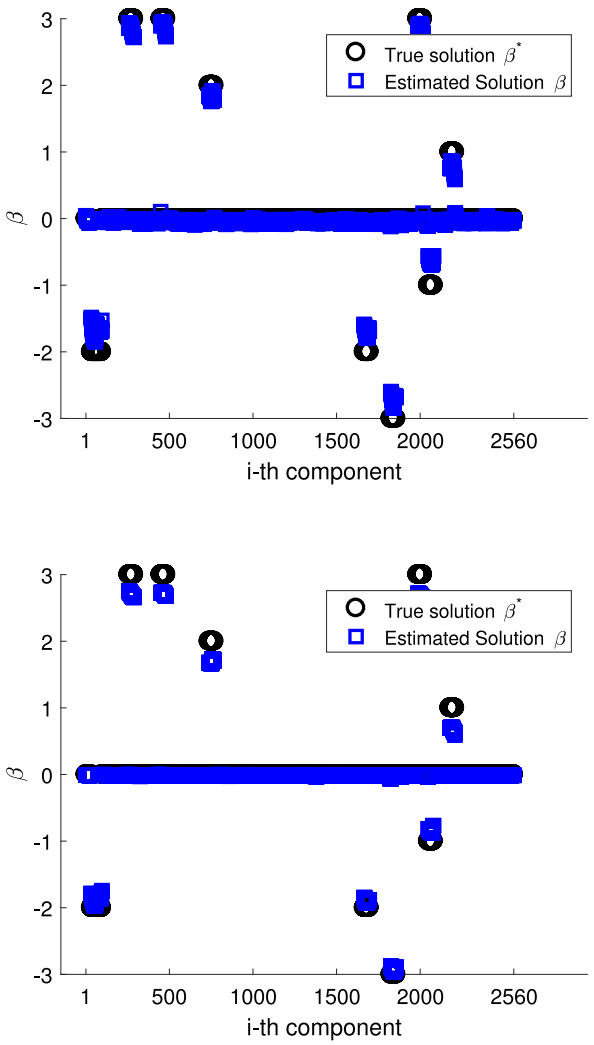

Fig. 3. Visual results for convergent case with respect to $\|\alpha\|$ (top), $\|\beta\|$ (middle) and $\|\gamma\|$ (bottom).

In summary, the nonconvex fused regression outperforms the FLASSO in terms of the recovery accuracy for a wide range of noise distributions including non-Gaussian distribution. This good reconstruction performance can be attributed to the favor of the nonconvex penalty functions, which enforce much better sparsity.

5.2. Real-world data

In this subsection, we test the colon tumor gene expression data in Alon et al. (1999), which is an abnormal growth of cells. The colon tumor data contains 40 tumor biopsies from tumors and 22 normal biopsies from healthy parts of the same patients. There are 2000 selected genes based on the confidence in the measured expression levels. Thus, this colon tumor gene expression are with $n = 62$ and $p = 2000$. Furthermore, we select $\tau_1 = 0.1$ and $\tau_2 = 0.01$ for the nonconvex fused regression. For the FLASSO, we choose the tuning parameters as suggested in Li et al. (2014).

To evaluate the performance, we give the following specific measurements:

- "$\sharp\{|\beta_i - \beta_{i-1}| > 0.01\}$" refers to how many β_i satisfies $|\beta_i - \beta_{i-1}| > 0.01$, which represents the significant successive differences;
- "$\sharp\{|\beta_i| > 0.01\}$" refers to how many β_i satisfies $|\beta_i| > 0.01$, which represents the significant coefficients.

The numerical comparison results of the FLASSO and the nonconvex fused regression are reported in Table 6. According to this table, it is not hard to conclude that, compared with the FLASSO, our proposed nonconvex fused regression can achieve a smaller number of the significant successive differences and a smaller number of the significant coefficients at the same time. The superior is due to the nonconvex penalty functions used in the nonconvex fused regression.

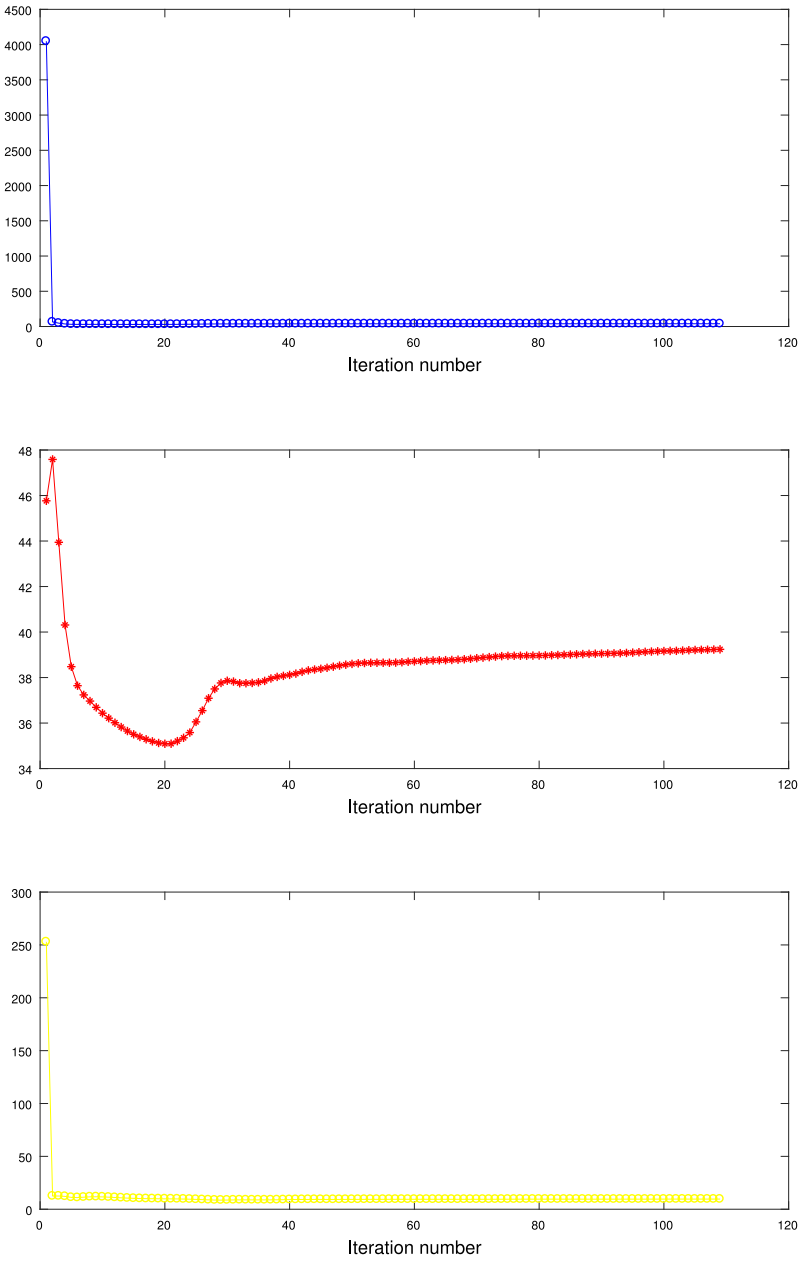

Fig. 4. Visual results non-convergent case with respect to $\|\alpha\|$ (top), $\|\beta\|$ (middle) and $\|\gamma\|$ (bottom).

5.3. Selection of parameters

In this subsection, we will illustrate how to select the parameters for our nonconvex fused regression. For parameters (τ_1, τ_2), we apply 5-fold cross-validation technique. We test 100 cases of (τ_1, τ_2) where τ_1 and τ_2 are chosen from the set $\{0.001, 0.01, 0.1, 0.2, 0.5, 1, 2, 10, 100, 1000\}$, and choose the one that can yield the smallest number of variables misclassified. The results for real-world data are listed in Table 7. Accordingly, we select $\tau_1 = 0.1$ and $\tau_2 = 0.01$ for the nonconvex fused regression.

Next, regarding the penalty parameters (μ_1, μ_2), we initialize them with small values, and then increase the μ_1 and μ_2 by a constant ratio. After finite updates, the penalty parameters μ_1 and μ_2 can be chosen to satisfy the conditions. Thus, the convergence of the resulting algorithm is guaranteed. For synthetic data under Gaussian noise, we plot plots $\|\alpha\|, \|\beta\|, \|\gamma\|$ along the iteration number in Fig. 3. In this case, the algorithm is convergent when the iteration number is around 110.

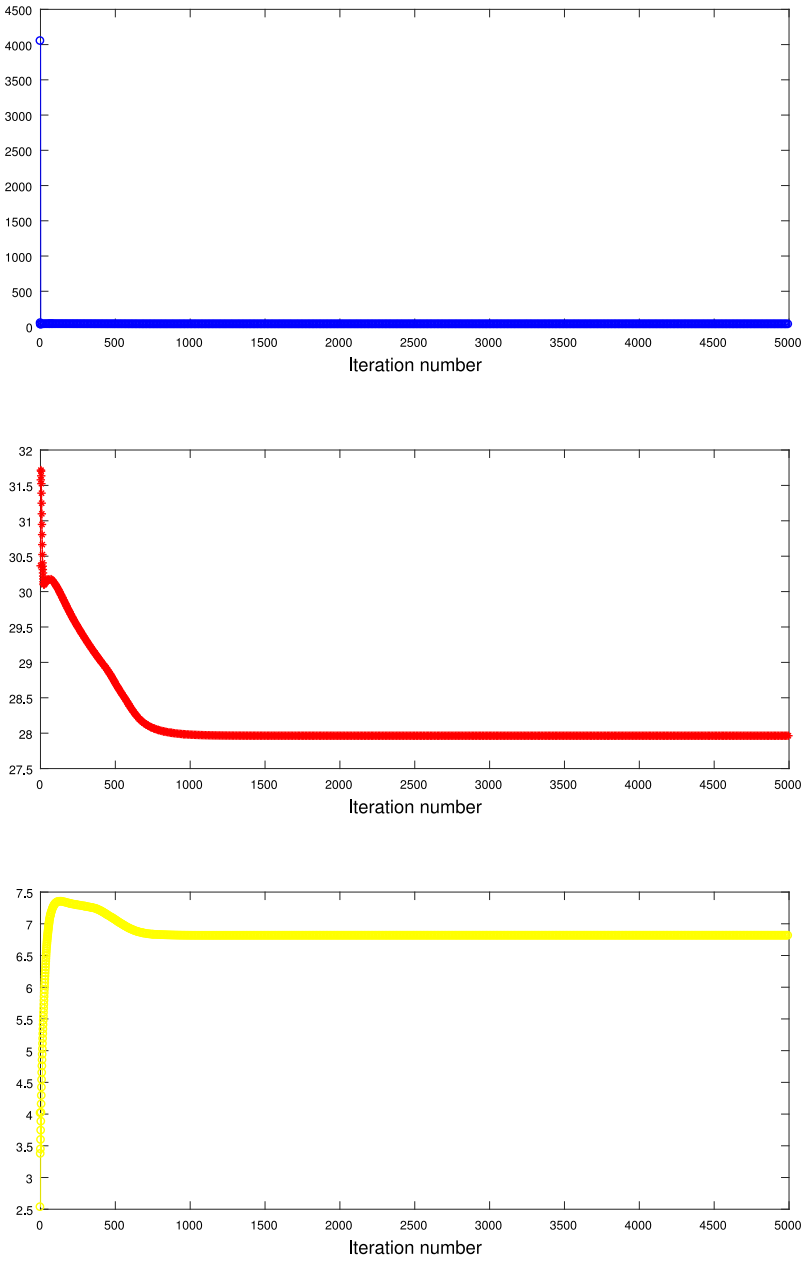

Fig. 5. Recovery results for the FLASSO (top) and the nonconvex fused regression (bottom) under non-Gaussian noise for $\sigma = 10$.

Table 6
Comparison results of the FLASSO and the nonconvex fused regression for the colon tumor data. The best results are highlighted in bold.

	$\sharp\{\|\beta_i\| > 0.01\}$	$\sharp\{\|\beta_i - \beta_{i-1}\| > 0.01\}$
FLASSO	85	158
NonFuReg	**74**	**126**

However, when μ_1 and μ_2 are too large, "Tol" could be satisfied before convergence. Fig. 4 gives an example. In this figure, Tol is satisfied when iteration number is around 150, but the problem is not convergent. How to choose μ_1 and μ_2 appropriately remains an open problem.

Table 7
Cross-validation results on (τ_1, τ_2) for the nonconvex fused regression on real-word data. The best results are highlighted in bold.

τ_1/τ_2	0.001	0.01	0.1	0.2	0.5	1	2	10	100	1000
0.001	15	16	13	14	13	12	13	20	20	20
0.01	15	15	13	13	13	12	14	22	21	21
0.1	10	**8**	8	8	10	12	19	27	26	23
0.2	9	9	10	12	12	15	18	32	32	32
1	10	9	9	10	10	13	21	28	28	28
2	9	9	9	11	11	18	25	29	29	29
10	25	25	25	25	25	25	25	25	25	25
100	25	25	25	25	25	25	25	25	25	25
1000	25	25	25	25	25	25	25	25	25	25

6. Conclusion and future work

In this paper, we extend the FLASSO model with a class of nonconvex penalty functions, in which the FLASSO is utilized to characterize the sparsity for not only the coefficients but also their successive differences, and the nonconvex penalty function is employed to enhance the sparsity. Furthermore, we develop an ADMM-based algorithm for the nonconvex fused regression model, and establish its convergence. Finally, we perform numerical experiments to demonstrate the superiority of our proposed model over the FLASSO. This is the first time to combine the FLASSO and a class of nonconvex penalty functions with an efficient algorithm.

In the future work, we are interested in the following research directions. First, we need to study the near oracle property for our proposed model as did for FLASSO in Tibshirani et al. (2005). Second, we can use our proposed model to solve other practical applications, such as image reconstruction or denoising.

Acknowledgments

The authors would like to thank Professor Xiaoming Yuan from Hong Kong Baptist University for gracefully providing the codes of FLASSO. The authors would also like to thank the anonymous referees and the associate editor for their numerous insightful comments and suggestions, which have greatly improved the paper. This work was supported in part by the National Natural Science Foundation of China (11431002, 11671029, 61633001).

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. 96 (12), 6745–6750.

Attouch, H., Bolte, J., Redont, P., Soubeyran, A., 2010. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. 35 (2), 438–457.

Bertsekas, D.P., 1999. Nonlinear Programming. Athena scientific Belmont.

Bolte, J., Sabach, S., Teboulle, M., 2014. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. 146 (1–2), 459–494.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3 (1), 1–122.

Chen, X., Ng, M.K., Zhang, C., 2012. Non-Lipschitz L_p-regularization and box constrained model for image restoration. IEEE Trans. Image Process. 21 (12), 4709–4721.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.

Fan, J., Xue, L., Zou, H., 2014. Strong oracle optimality of folded concave penalized estimation. Ann. Statist. 42 (3), 819.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al., 2007. Pathwise coordinate optimization. Ann. Appl. Stat. 1 (2), 302–332.

Gabay, D., 1983. Chapter ix applications of the method of multipliers to variational inequalities. Stud. Math. Appl. 15, 299–331.

Gabay, D., Mercier, B., 1976. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. 2 (1), 17–40.

Gill, P.E., Murray, W., Saunders, M.A., 2008. User's Guide for SQOPT Version 7: Software for Large-Scale Linear and Quadratic Programming.

Glowinski, R., Marroco, A., 1975. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. Rev. Fr. Autom. Inform. Rech. Oper. Anal. Numer. 9 (R2), 41–76.

Guo, K., Han, D., Wu, T., 2017. Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints. Int. J. Comput. Math. 94 (8), 1653–1669.

He, B., Yuan, X., 2012. On the o(1/n) convergence rate of the Douglas–Rachford alternating direction method. SIAM J. Numer. Anal. 50 (2), 700–709.

Hong, M., Luo, Z.Q., Razaviyayn, M., 2016. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. SIAM J. Optim. 26 (1), 337–364.

Li, X., Mo, L., Yuan, X., Zhang, J., 2014. Linearized alternating direction method of multipliers for sparse group and fused LASSO models. Comput. Statist. Data Anal. 79, 203–221.

Liu, Y.F., Dai, Y.H., Ma, S., 2015. Joint power and admission control: Non-convex lq approximation and an effective polynomial time deflation approach. IEEE Trans. Signal Process. 63 (14), 3641–3656.

Ng, M.K., Chan, R.H., Tang, W.C., 1999. A fast algorithm for deblurring models with Neumann boundary conditions. SIAM J. Sci. Comput. 21 (3), 851–866.

Rockafellar, R., Wets, R.J.B., 2009. Variational Analysis, vol. 317. Springer Science & Business Media.

Rudin, L.I., Osher, S., 1994. Total variation based image restoration with free local constraints. In: Image Processing, 1994. Proceedings. ICIP-94. IEEE International Conference, vol. 1. IEEE, pp. 31–35.

Rudin, L.I., Osher, S., Fatemi, E., 1992. Nonlinear total variation based noise removal algorithms. Physica D 60 (1–4), 259–268.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 267–288.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (1), 91–108.
Tibshirani, R., Wang, P., 2007. Spatial smoothing and hot spot detection for CGH data using the fused lasso. Biostatistics 9 (1), 18–29.
Wang, F., Cao, W., Xu, Z., 2015. Convergence of multi-block Bregman ADMM for nonconvex composite problems. ArXiv preprint arXiv:1505.03063.
Wright, S.J., Nocedal, J., 1999. Numerical optimization. Springer Sci. 35 (67–68), 7.
Xin, B., Tian, Y., Wang, Y., Gao, W., 2015. Background subtraction via generalized fused lasso foreground modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4676–4684.
Xiu, X., Kong, L., Li, Y., Qi, H.D., 2018. Iterative reweighted methods for L1- Lp minimization. Comput. Optim. Appl.
Xu, Z., Chang, X., Xu, F., Zhang, H., 2012. *L*_{1/2} regularization: A thresholding representation theory and a fast solver. IEEE Trans. Neural Netw. Learn. Syst. 23 (7), 1013–1027.
Yang, L., Pong, T.K., Chen, X., 2017. Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction. SIAM J. Imaging Sci. 10 (1), 74–110.
Zhang, T., 2010. Analysis of multi-stage convex relaxation for sparse regularization. J. Mach. Learn. Res. 11 (Mar), 1081–1107.
Zhang, C.H., et al., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38 (2), 894–942.