Pure Mathematics 理论数学, 2022, 12(11), 2021-2032
Published Online November 2022 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2022.1211219

Hans汉斯

基于Huber损失和Capped-L1正则的线性不等式约束稀疏优化问题研究

田梦达，彭定涛*，张　弦

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2022年10月23日；录用日期：2022年11月22日；发布日期：2022年11月30日

摘　要

对多元线性回归中回归系数的估计问题，本文考虑了基于 Huber 损失和线性不等式约束的稀疏优化模型。首先，给出了稀疏优化的原问题、基于 Capped-L1 正则的松弛问题和基于约束惩罚的无约束问题三种模型。其次，借助惩罚模型方向稳定点的下界性质，在一定条件下分析了三种模型全局最优解的等价性。最后，提出了光滑化惩罚算法，并证明了该算法的收敛性。本文为求解线性不等式约束稀疏优化问题提供了理论和方法基础。

关键词

线性不等式约束稀疏优化问题，Huber损失，Capped-L1正则，方向稳定点，光滑化惩罚算法

On Sparse Optimization Problems with Linear Inequality Constraints Based on Huber Loss and Capped-L1 Regularization

Mengda Tian, Dingtao Peng*, Xian Zhang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Oct. 23rd, 2022; accepted: Nov. 22nd, 2022; published: Nov. 30th, 2022

Abstract

For the estimation of regression coefficients in multivariate linear regression, a sparse optimization model based on Huber loss and linear inequality constraints is considered in this paper.

*通讯作者。

文章引用: 田梦达, 彭定涛, 张弦. 基于 Huber 损失和 Capped-L1 正则的线性不等式约束稀疏优化问题研究[J]. 理论数学, 2022, 12(11): 2021-2032. DOI: 10.12677/pm.2022.1211219

Firstly, three models of the original sparse optimization problem, the relaxation problem based on Capped-L1 regularization and the unconstrained problem based on the penalty of constraint are given. Secondly, by use of the lower bound property of the directional stationary point of the penalized model, the equivalence of the global optimal solutions of the three models is analyzed under certain conditions. Finally, a smoothing penalty algorithm is proposed and its convergence is proved. This paper provides a theoretical and methodological basis for solving sparse optimization problems with linear inequality constraints.

Keywords

Sparse Optimization Problem with Linear Inequality Constraints, Huber Loss, Capped-L1 Regularization, Directional Stationary Point, Smoothing Penalty Algorithm

1. 引言

在多元线性回归中，常用的最小二乘法是通过极小化残差平方和来估计回归系数，即：

$$\min_{x\in\mathbb{R}^n} F(x) := \|Ax-b\|_2^2,$$

其中，$A\in\mathbb{R}^{m\times n}$，$b\in\mathbb{R}^m$。最小二乘法也常用于曲线拟合。为避免数据的多重共线性和可能出现的欠定和过拟合现象，并实现变量选择，研究者引入了 ℓ_0 正则的稀疏优化模型：

$$\min_{x\in\mathbb{R}^n} F(x) := \|Ax-b\|_2^2 + \lambda\|x\|_0, \tag{1}$$

其中 $\|x\|_0$ 表示向量 x 非零分量的个数，$\lambda>0$ 是正则化参数。

由于 ℓ_0 正则是非凸、非光滑甚至是不连续的。因此，求解问题(1)是 NP 难的[1]。于是一些研究学者考虑用 ℓ_1 正则来松弛 ℓ_0 正则[2] [3] [4]，即 LASSO 回归模型：

$$\min_{x\in\mathbb{R}^n} F(x) := \|Ax-b\|_2^2 + \lambda\|x\|_1,$$

其中 $\|x\|_1 = \sum_{i=1}^n |x_i|$。LASSO 回归模型具有子集选择和岭回归的一些优点，它能够产生可解释的模型并且具有岭回归的稳定性。但是 Fan 和 Li 证明了 LASSO 回归模型得到的解是有偏估计[5]，并指出一个好的正则函数应当使得产生的估计量具有下述四个性质：1) 无偏性，2) 稀疏性，3) 连续性，4) Oracle 性质：所得估计量与 Oracle 解具有相同的渐进分布，其中 Oracle 解定义为：

$$x^{\text{Oracle}} \in \arg\min_{x:\operatorname{supp}(x)\subset\operatorname{supp}(x^*)} L(x),$$

$L(x)$ 是损失函数，$\operatorname{supp}(x^*)$ 是真实解 x^* 的支撑集。研究表明，SCAD [5]，MCP [6]和 Capped-L1 等几类折叠凹正则函数[7] [8] [9] [10]可产生满足无偏性、稀疏性、连续性和 Oracle 性质的估计量[11] [12] [13]。因此研究者们考虑用折叠凹函数来松弛 ℓ_0 正则，即考虑以下折叠凹正则模型：

$$\min_{x\in\mathbb{R}^n} F(x) := \|Ax-b\|_2^2 + \Phi(x), \tag{2}$$

其中 $\Phi(x)=\sum_{i=1}^{n}\varphi(x_i)$ 是折叠凹函数， $\varphi(t)$ 可取如下几类函数：

i) Capped-L1： $\varphi(t)=\lambda\cdot\min\left\{1,\frac{|t|}{\gamma}\right\},\lambda>0,\gamma>0$

ii) MCP： $\varphi(t)=\begin{cases}\lambda|t|-\frac{t^2}{2\gamma}, & \text{if } 0\le|t|\le\gamma\lambda,\\ \frac{\gamma\lambda^2}{2}, & \text{if } |t|>\gamma\lambda,\end{cases},\lambda>0,\gamma>1;$

iii) SCAD： $\varphi(t)=\begin{cases}\lambda|t|, & \text{if } 0\le|t|\le\lambda,\\ \lambda|t|-\frac{(|t|-\lambda)^2}{2(\gamma-1)}, & \text{if } \lambda<|t|\le\gamma\lambda,,\lambda>0,\gamma>2.\\ \frac{(\gamma+1)\lambda^2}{2}, & \text{if } |t|>\gamma\lambda,\end{cases}$

因为 $\Phi(x)$ 是非凸正则函数，所以问题(2)是非凸优化。研究者们已经发展了多种有效算法，例如：凸差算法[9] [14] [15] [16]，信赖域算法[17]，迭代重加权算法[18]等。

文献[19]研究了 Capped-L1 正则模型(2)与原 ℓ_0 正则模型(1)解的关系，在一定条件下证明了 Capped-L1 正则模型与原 ℓ_0 模型全局解的等价性和局部解的包含关系，并给出了邻近梯度算法。文献[20]研究了损失函数为最小一乘时，MCP 正则模型与原 ℓ_0 正则模型解的关系，证明了两模型全局解的等价性。文献[21]对损失函数为一般凸函数的组稀疏优化问题，研究了组 Capped-L1 正则模型与 $\ell_{2,0}$ 正则模型解的关系，证明了两模型全局解的等价性和局部解的包含关系。文献[22]研究了带线性约束的组稀疏优化问题及其折叠凹松弛问题解的等价性和求解算法。文献[23]进一步对带一般凸约束的组稀疏优化问题，研究了组 Capped-L1 正则模型与 $\ell_{2,0}$ 正则模型解的关系，并利用组 Capped-L1 正则模型给出了组光滑化邻近梯度算法。

由于模型(1)中的最小二乘损失缺乏鲁棒性，对异常值的容忍度不高[3]，而 Huber 函数不仅对异常值具有鲁棒性，而且结合了最小一乘和最小二乘的优点，既光滑又不会放大误差，因此，将 Huber 函数作为损失函数具有非常大的优点。另一方面，模型(1)没有考虑约束条件，这也在很大程度上限制了它的应用范围。

基于上述分析，本文考虑如下带 Huber 损失和线性不等式约束的稀疏优化模型：

$$\begin{aligned}&\min_{x\in\mathbb{R}^n}\frac{1}{m}\sum_{i=1}^{m}H\left(A_i^{\mathrm{T}}x-b_i\right)+\lambda\|x\|_0\\&\text{s.t. } Bx\le h,\end{aligned}\tag{3}$$

其中

$$H(t)=\begin{cases}\frac{1}{2}t^2, & |t|\le\delta,\\ \delta\left(|t|-\frac{1}{2}\delta\right), & \text{其他},\end{cases}$$

是 Huber 函数， $A\in\mathbb{R}^{m\times n}$， $A_i^{\top}$ 是 A 的第 i 个行向量， $i=1,\cdots,m$， $B\in\mathbb{R}^{q\times n}$， $h\in\mathbb{R}^q$， $\delta>0$。不等式约束 $Bx\le h$ 可以刻画真实解(信号)的某些先验信息，如非负性[22]、有界性[19]等，但增加了约束使得问题的分析和求解变得更加复杂。

为便于分析和求解，本文将模型(3)松弛为如下 Capped-L1 正则模型：

$$\begin{aligned}&\min_{x\in\mathbb{R}^n}\frac{1}{m}\sum_{i=1}^{m}H\left(A_i^{\mathrm{T}}x-b_i\right)+\Phi(x)\\&\text{s.t. } Bx\le h.\end{aligned}\tag{4}$$

其中 $\Phi(x)=\sum_{i=1}^{n}\varphi(x_i)$，而 $\varphi(t)=\lambda\cdot\min\left\{1,\frac{|t|}{\gamma}\right\}$ 是 Capped-L1 函数。

模型(4)是约束优化，为方便研究，将其不等式约束作为惩罚项加罚到目标函数上去，从而转化为如下无约束优化：

$$\min_{x\in\mathbb{R}^n}\frac{1}{m}\sum_{i=1}^{m}H\left(A_i^{\mathrm{T}}x-b_i\right)+\Phi(x)+\alpha\left\|(Bx-h)_+\right\|_1,\tag{5}$$

其中 $z_+\in\mathbb{R}^n$，其定义为 $(z_+)_i:=\max\{0,z_i\}$，$\alpha>0$ 为惩罚因子。

这里重要且有趣的问题是，模型(3)、模型(4)和模型(5)之间解的关系如何，特别是它们的解是否具有某种等价性，以及如何对其求解的问题。

本文主要工作如下：

i) 定义了问题(5)的方向稳定点(d-稳定点)，分析了问题(5)的一阶最优性条件，并探讨了问题(3)，(4)与(5)之间解的关系，证明了等价性。

ii) 因为问题(4)是非光滑优化问题，本文使用光滑化惩罚方法来计算其 d-稳定点。通过对约束惩罚函数的光滑化逼近来得到近似问题，并证明了该算法产生的任意聚点都是松弛问题的 d-稳定点，为使用光滑化方法求解该问题提供了理论和方法保证。

在接下来的讨论中，为了简便，记：

$$\begin{aligned}F(x)&=\frac{1}{m}\sum_{i=1}^{m}H\left(A_i^{\mathrm{T}}x-b_i\right)+\Phi(x)+\alpha\left\|(Bx-h)_+\right\|_1\\f(x)&=\frac{1}{m}\sum_{i=1}^{m}H\left(A_i^{\mathrm{T}}x-b_i\right),\\Q(x)&=\left\|(Bx-h)_+\right\|_1,\\\Omega&=\left\{x\in\mathbb{R}^n:Bx\le h\right\}.\end{aligned}$$

对任意闭集 $\Omega\in\mathbb{R}^n$，$\operatorname{dist}(x,\Omega)=\inf_{y\in\Omega}\|x-y\|$ 表示 x 到闭集 Ω 的距离，$P_\Omega(x)$ 表示 $x\in\mathbb{R}^n$ 在 Ω 上投影点的集合。

记向量 x 的支撑集为：

$$\Gamma(x)=\left\{i\in\{1,\cdots,n\}:x_i\ne0\right\}=\Gamma_1(x)\cup\Gamma_2(x),$$

其中 $\gamma>0$，

$$\Gamma_1(x)=\left\{i:0<|x_i|<\gamma\right\},\Gamma_2(x)=\left\{i:|x_i|\ge\gamma\right\}.$$

符号函数 $\operatorname{sgn}(t)$ 定义为：

$$\operatorname{sgn}(t):=\begin{cases}1, & t>0,\\0, & t=0,\\-1, & t<0.\end{cases}$$

本文结构如下：第二节首先借助方向导数给出问题(5) d-稳定点的定义，然后得出其 d-稳定点的下界

性质。第三节探讨问题(3)，问题(4)与问题(5)解的等价性。第四节给出求解问题(5) d-稳定点的光滑化惩罚方法，并证明该算法的收敛性。

2. 最优性条件

2.1. 问题(5)的 d-稳定点

首先给出问题(5)的 d-稳定点的定义[16] [24]。

定义 2.1：设 $F:\mathbb{R}^n\to\mathbb{R}$ 在点 $x\in\mathbb{R}^n$ 处方向可微，则函数 F 在点 x 处沿方向 $w\in\mathbb{R}^n$ 的方向导数定义为：

$$F'(x;w):=\lim_{\tau\downarrow 0}\frac{F(x+\tau w)-F(x)}{\tau}.$$

定义 2.2：称 $x^*\in\mathbb{R}^n$ 是问题(5)的 d-稳定点，如果：

$$F'(x^*;x-x^*)=f'(x^*;x-x^*)+\Phi'(x^*;x-x^*)+\alpha Q'(x^*;x-x^*)\geq 0,\quad \forall x\in\mathbb{R}^n.$$

Peng 和 Chen [24] [25]证明了当目标函数局部 Lipschitz 连续且方向可微时，d-稳定点具有如下最优性质：

定理 2.1：设函数 $F:\mathbb{R}^n\to\mathbb{R}$ 在点 $\hat{x}\in\mathbb{R}^n$ 处是局部 Lipschitz 连续且方向可微的，则有如下性质：

i) 若 $\hat{x}$ 是函数 F 的局部最优值点，那么 $\hat{x}$ 是函数 F 的 d-稳定点。

ii) $\hat{x}$ 是函数 F 的严格局部最优值点并满足一阶增长性条件，即存在 $\hat{x}$ 的领域 $\mathcal{W}$ 和 $\delta>0$ 使得：

$$F(x)\geq F(\hat{x})+\delta\|x-\hat{x}\|,\ \forall x\in\mathcal{W},$$

当且仅当 $\hat{x}$ 满足：

$$F'(\hat{x};x-\hat{x})>0,\ \forall x\in\mathbb{R}^n\setminus\{\hat{x}\}.$$

2.2. 问题(5) d-稳定点的下界性质

下述定理表明问题(5)的 d-稳定点的非零分量具有一致的下界。

定理 2.2：设 $\frac{\lambda}{\gamma}>\frac{\delta}{m}\left\|\sum_{i=1}^m A_i\right\|+\alpha\sqrt{q}\|B\|_F$，若 $x^*\in\mathbb{R}^n$ 是问题(5)的 d-稳定点，那么或者 $|x_i^*|\geq\gamma$，或者 $|x_i^*|=0$，$\forall i=1,\cdots,n$。

证明：根据记号，要证明本定理，只需证明 $\Gamma_1(x^*)=\varnothing$。因 x^* 是 d-稳定点，由定义 2.2：

$$f'(x^*;x-x^*)+\Phi'(x^*;x-x^*)+\alpha Q'(x^*;x-x^*)\geq 0,\quad \forall x\in\mathbb{R}^n$$

其中：

i) 因为 Huber 损失函数 $f(x)=\frac{1}{m}\sum_{i=1}^m H(A_i^{\mathrm{T}}x-b_i)$ 是可微的，故

$$f'(x^*;x-x^*)=\nabla f(x^*)^{\top}(x-x^*)=\frac{1}{m}\sum_{i=1}^m H'(A_i^{\top}x^*-b_i)A_i^{\top}(x-x^*),$$

这里

$$H'(A_i^{\top}x^*-b_i)=\begin{cases}A_i^{\top}x^*-b_i, & |A_i^{\top}x^*-b_i|\leq\delta,\\ \delta\cdot\operatorname{sgn}(A_i^{\top}x^*-b_i), & \text{其他.}\end{cases}$$

ii) $\Phi'\left(x^*;x-x^*\right)=\sum_{i=1}^{n}\varphi'\left(x_i^*;x_i-x_i^*\right)$，且由 $\varphi(t)=\lambda\cdot\min\left\{1,\frac{|t|}{\gamma}\right\}$，得：

$$\varphi'\left(x_i^*;x_i-x_i^*\right)=\begin{cases}\dfrac{\lambda|x_i|}{\gamma}, & x_i^*=0,\\ \dfrac{\lambda\left(x_i-x_i^*\right)\operatorname{sgn}\left(x_i^*\right)}{\gamma}, & \left|x_i^*\right|\in(0,\gamma),\\ \min\left\{0,\dfrac{\lambda\left(x_i-x_i^*\right)\operatorname{sgn}\left(x_i^*\right)}{\gamma}\right\}, & \left|x_i^*\right|=\gamma,\\ 0, & \text{其他.}\end{cases}$$

iii) 根据文献[25]，得 $Q'\left(x^*;x-x^*\right)=\Delta:=\sum_{j=1}^{q}\Delta_j$，其中

$$\Delta_j=\begin{cases}0, & \text{如果}\left\langle B_j,x^*\right\rangle<h_j,\\ \max\left\{0,\left\langle B_j,x-x^*\right\rangle\right\}, & \text{如果}\left\langle B_j,x^*\right\rangle=h_j,\\ \left\langle B_j,x-x^*\right\rangle, & \text{其他,}\end{cases}$$

此处，$B_j^{\top}$ 是矩阵 B 的第 j 个行向量。

下面用反证法证明。假设 $\Gamma_1\left(x^*\right)\neq\varnothing$。对每个 $i_0\in\Gamma_1\left(x^*\right)$，定义 $\hat{x}^1,\hat{x}^2\in\mathbb{R}^n$ 如下：

$$\hat{x}_i^1=\begin{cases}2x_{i_0}^*, & \text{如果}i=i_0,\\ x_i^*, & \text{其他,}\end{cases}\quad\hat{x}_i^2=\begin{cases}0, & \text{如果}i=i_0,\\ x_i^*, & \text{其他,}\end{cases}\quad i=1,\cdots,n.$$

则 $F'\left(x^*,\hat{x}^\eta-x^*\right)\geq 0,\eta=1,2$。由上述(i) (ii) (iii)，得：

$$F'\left(x^*,\hat{x}^1-x^*\right)=\left[\nabla f\left(x^*\right)\right]_{i_0}x_{i_0}^*+\frac{\lambda x_{i_0}^*\operatorname{sgn}\left(x_{i_0}^*\right)}{\gamma}+\Delta_{i_0}\geq 0,$$

$$F'\left(x^*,\hat{x}^2-x^*\right)=\left[\nabla f\left(x^*\right)\right]_{i_0}x_{i_0}^*-\frac{\lambda x_{i_0}^*\operatorname{sgn}\left(x_{i_0}^*\right)}{\gamma}-\Delta_{i_0}\geq 0.$$

于是

$$\begin{aligned}\frac{\lambda\left|x_{i_0}^*\right|}{\gamma}&\leq\left|\left[\nabla f\left(x^*\right)\right]_{i_0}\right|+\left|\Delta_{i_0}\right|\leq\left\|\frac{1}{m}\sum_{i=1}^{m}H'\left(A_i^{\top}x^*-b_i\right)A_i\right\|\left|x_{i_0}^*\right|+\left\|B^{i_0}\right\|\left|x_{i_0}^*\right|\\&\leq\left(\frac{\delta}{m}\left\|\sum_{i=1}^{m}A_i\right\|+\alpha\sqrt{q}\|B\|_F\right)\left|x_{i_0}^*\right|.\end{aligned}$$

由 $\left|x_{i_0}^*\right|>0$，得 $\frac{\lambda}{\gamma}\leq\frac{\delta}{m}\left\|\sum_{i=1}^{m}A_i\right\|+\alpha\sqrt{q}\|B\|_F$，这与已知条件 $\frac{\lambda}{\gamma}>\frac{\delta}{m}\left\|\sum_{i=1}^{m}A_i\right\|+\alpha\sqrt{q}\|B\|_F$ 矛盾，所以 $\Gamma_1\left(x^*\right)=\varnothing$。

注：定理 2.2 表明问题(5)的 d-稳定点的非零分量的绝对值具有正下界 γ。这种解的下界性质在理论上反映了解的稀疏性，在数值计算中，如果数值解的某些分量小于下界，则可以直接将这些分量取为 0，这样可以提高解的稀疏度。类似研究可参见文献[11] [19] [22] [24]。

3. 解的等价性

3.1. 问题(3)和问题(4)解的等价性

定理 3.1：若 $\bar{x}\in\Omega$，那么 $\bar{x}$ 是问题(4)的全局最优解当且仅当 $\bar{x}$ 是问题(3)的全局最优解，且问题(3)

和问题(4)具有相同的全局最优值。

证明：

i) 设 $\bar{x}\in\Omega$ 是问题(4)的全局最优解，则

$$f(\bar{x})+\lambda\|\bar{x}\|_0=f(\bar{x})+\Phi(\bar{x})\le f(x)+\Phi(x)\le f(x)+\lambda\|x\|_0,\quad \forall x\in\Omega,$$

其中第一个等式由[19]中引理 2.3 可得，而最后一个不等式由 $\Phi(x)\le\lambda\|x\|_0$ 对任何 $x\in\mathbb{R}^n$ 均成立可得。故 $\bar{x}$ 是问题(3)全局最优解。

ii) 设 $\bar{x}\in\Omega$ 是问题(3)全局最优解，但不是问题(4)的全局最优解，则问题(4)存在一个全局最优解 $\hat{x}$，使得

$$f(\hat{x})+\Phi(\hat{x})<f(\bar{x})+\Phi(\bar{x}).$$

由(i)知 $\hat{x}$ 也是问题(3)全局最优解，因此

$$f(\hat{x})+\lambda\|\hat{x}\|_0<f(\bar{x})+\lambda\|\bar{x}\|_0,$$

这与 $\bar{x}$ 是问题(3)全局最优解相矛盾。所以问题(3)的任何全局最优解都是问题(4)的全局最优解。

3.2. 问题(4)和问题(5)解的等价性

假设 3.1：矩阵 B 和向量 h 满足如下条件：存在一些正数 τ，使得 $\operatorname{dist}(x,\Omega)\le\tau\left\|(Bx-h)_+\right\|_1$ [26]。

定理 3.2：设假设 3。1 成立，则问题(4)的全局最优解都是问题(5)的全局最优解。

证明：因为 $f(x)+\Phi(x)$ 是 Lipschitz 连续的，设其 Lipschitz 常数为 L_f。对所有的 $\alpha>\tau L_f$，有

$$f(x)+\Phi(x)+\frac{\alpha}{\tau}\operatorname{dist}(x,\Omega)\ge f(y)+\Phi(y),\quad \forall x\in\mathbb{R}^n,\forall y\in P_\Omega(x).\tag{6}$$

由假设 3.1 和(6)，可得：

$$\begin{aligned}\inf_{x\in\mathbb{R}^n}f(x)+\Phi(x)+\alpha\left\|(Bx-h)_+\right\|_1&\ge\inf_{x\in\mathbb{R}^n}f(x)+\Phi(x)+\frac{\alpha}{\tau}\operatorname{dist}(x,\Omega)\\&\ge\inf_{x\in\mathbb{R}^n,y\in P_\Omega(x)}f(y)+\Phi(y)\\&=\inf_{x\in\Omega}f(x)+\Phi(x)\\&=\inf_{x\in\Omega}f(x)+\Phi(x)+\alpha\left\|(Bx-h)_+\right\|_1\\&\ge\inf_{x\in\mathbb{R}^n}f(x)+\Phi(x)+\alpha\left\|(Bx-h)_+\right\|_1.\end{aligned}$$

因此，x^* 是问题(5)的全局最优解。

定理 3.3：设假设 3.1 成立，且 $\alpha>\tau L_f$，其中 L_f 是 $f(x)+\Phi(x)$ 的 Lipschitz 常数。若 x^* 是问题(5)的全局最优解，那么 x^* 是问题(4)的全局最优解。

证明：$x^*\in\mathbb{R}^n$ 是问题(5)的全局最优解，则

$$\begin{aligned}f(x^*)+\Phi(x^*)+\alpha\left\|(Bx^*-h)_+\right\|_1&=\inf_{x\in\mathbb{R}^n}f(x)+\Phi(x)+\alpha\left\|(Bx-h)_+\right\|_1\\&\le\inf_{x\in\Omega}f(x)+\Phi(x)\\&\le f(x)+\Phi(x),\quad\forall x\in\Omega.\end{aligned}\tag{7}$$

由假设 3.1，对任意 $x\in P_\Omega(x^*)$，有

$$f(x^*)+\Phi(x^*)+\frac{\alpha}{\tau}\operatorname{dist}(x^*,\Omega)\le f(x)+\Phi(x).$$

考虑到 $f(x)+\Phi(x)$ 是 Lipschitz 连续的，对任意 $x\in P_{\Omega}(x^*)$，由上式得

$$\begin{aligned}\operatorname{dist}(x^*,\Omega)&\le\frac{\tau}{\alpha}\left[f(x)+\Phi(x)-f(x^*)-\Phi(x^*)\right]\\&\le\frac{\tau L_f}{\alpha}\|x-x^*\|=\frac{\tau L_f}{\alpha}\operatorname{dist}(x^*,\Omega).\end{aligned}$$

因为 $\alpha>\tau L_f$，故 $\operatorname{dist}(x^*,\Omega)=0$，因此 $x^*\in\Omega$。再由 $\alpha\left\|(Bx^*-h)_+\right\|_1\ge 0$ 和(7)式，可得 $f(x)+\Phi(x)\ge f(x^*)+\Phi(x^*)$，$\forall x\in\Omega$。故 x^* 是问题(4)的全局最优解。

3.3. 问题(3)和问题(5)解的等价性

由定理 3.1、定理 3.2 和定理 3.3 可得问题(3)与问题(4)之间解的等价性。

定理：设 $\frac{\lambda}{\gamma}>\frac{\delta}{m}\left\|\sum_{i=1}^{m}A_i\right\|+\alpha\sqrt{q}\|B\|_F$，$\alpha>\tau L_f$ 且 $\operatorname{dist}(x,\Omega)\le\tau\left\|(Bx-h)_+\right\|_1$，则 $x^*\in\mathbb{R}^n$ 是问题(5)的全局最优解当且仅当它是问题(3)的全局最优解。

4. 光滑化惩罚算法

由定理 2.1 可知，d-稳定点具有非常好的局部最优性。如何计算 d-稳定点是一个有趣且具有挑战性的问题。光滑逼近方法是一种求解非光滑问题非常有效且被广泛使用的方法，参见[19] [22] [23] [24] [26]。受上述文献启发，本节我们使用光滑化惩罚算法来求解问题(4)。

对于 $t_+=\max\{t,0\}$ 函数，将采用下述光滑化函数

$$h_{\mu}(t):=\begin{cases}t-\dfrac{\mu}{2}, & t\ge\mu,\\ \dfrac{t^2}{2\mu}, & 0<t<\mu,\\ 0, & t\le 0,\end{cases}$$

其中 $\mu>0$ 为光滑化参数。因此，$Q(x)$ 具有如下光滑化函数

$$Q_{\mu}(x):=\sum_{j=1}^{q}h_{\mu}\left(B_j^{\top}x-h_j\right).$$

因 $0\le t_+-h_{\mu}(t)\le\frac{\mu}{2}$，故对任意的 $x\in\mathbb{R}^n$，可得

$$0\le Q_{\mu}(x)\le Q(x)\le Q_{\mu}(x)+\frac{q}{2}\mu.$$

此外，注意到 $h'_{\mu}(t)=\min\left\{\left(\dfrac{t}{\mu}\right)_+,1\right\}$，

$$\nabla Q_{\mu}(x)=\sum_{j=1}^{q}h'_{\mu}\left(B_j^{\top}x-h_j\right)B_j.$$

容易证明，光滑函数 $Q_{\mu}(x)$ 具有下述性质。

i) $\lim_{z\to x,\mu\downarrow 0}Q_{\mu}(z,\mu)=Q(x)$；

ii) 对每个固定的 $\mu>0$，$Q_{\mu}(x)$ 是 x 的凸函数；

iii) 对每个固定的 $\mu>0$，$Q_{\mu}(x)$ 关于 x 是 Lipschitz 连续的，即

$$\left|Q_\mu(x_1)-Q_\mu(x_2)\right|\le\kappa\|x_1-x_2\|;$$

iv) 对每个固定的 $x\in\mathbb{R}^n$， $Q_\mu(x)$ 关于 μ 是 Lipschitz 连续的，即

$$\left|Q_\mu(x_1)-Q_\mu(x_2)\right|\le\kappa'|\mu_1-\mu_2|.$$

下面给出求解问题(4)的光滑化惩罚算法的框架。

算法 4.1. 光滑化惩罚算法

初始步：选取 $x^{\text{feas}}\in\Omega$， $x^0\in\mathbb{R}^n$， $\alpha_0>0$， $\mu_0>0$， $\epsilon_0>0$， $\rho>1$ 和 $\theta\in(0,1)$，令 $x^{0,0}=x^0$， $k=0$。

迭代步：

步 1 如果 $Q_{\mu_k}(x^{k,0})>Q_{\mu_k}(x^{\text{feas}})$，令 $x^{k,0}=x^{\text{feas}}$。

用 $x^{k,0}$ 作为初始点求 $\min\{G_{\lambda_k,\mu_k}(x):=f(x)+\Phi(x)+\alpha_kQ_{\mu_k}(x)\}$ 的近似解 x^k，使得

$$\max\left\{0,-\min_{x\in\mathbb{R}^n}\left\{\left\langle\nabla f(x^k),x-x^k\right\rangle+\Phi'(x^k,x-x^k)+\alpha_k\left\langle\nabla Q_{\mu_k}(x^k),x-x^k\right\rangle\right\}\right\}\le\epsilon_k.$$

步 2 令 $\alpha_{k+1}=\rho\alpha_k$， $\mu_{k+1}=\theta\mu_k$， $\epsilon_{k+1}=\theta\epsilon_k$， $x^{k+1,0}=x^k$。

步 3 令 $k=k+1$，转步 1。

停止

在算法 4.1 中，

$$\nabla f(x^k)=\frac{1}{m}\sum_{i=1}^m H'(A_i^\top x^k-b_i)A_i,$$

$$\Phi'(x^k;x-x^k)=\sum_{i=1}^m\varphi'(x_i^k;x_i-x_i^k),$$

$$\nabla Q_\mu(x^k)=\sum_{j=1}^q h'_\mu(B_j^\top x^k-h_j)B_j,$$

其中，

$$H'(A_i^\top x^k-b_i)=\begin{cases}A_i^\top x^k-b_i, & \left|A_i^\top x^k-b_i\right|\le\delta,\\ \delta\cdot\operatorname{sgn}(A_i^\top x^k-b_i), & \text{其他},\end{cases}$$

$$\varphi'(x_i^k;x_i-x_i^k)=\begin{cases}\dfrac{\lambda|x_i|}{\gamma}, & x_i^k=0,\\ \dfrac{\lambda(x_i-x_i^k)\operatorname{sgn}(x_i^k)}{\gamma}, & |x_i^k|\in(0,\gamma),\\ \min\left\{0,\dfrac{\lambda(x_i-x_i^k)\operatorname{sgn}(x_i^k)}{\gamma}\right\}, & |x_i^k|=\gamma,\\ 0, & \text{其他},\end{cases}$$

$$h'_\mu(B_j^\top x^k-h_j)=\min\left\{\left(\frac{B_j^\top x^k-h_j}{\mu}\right)_+,1\right\}.$$

由上述表达式可知 $\left|H'(A_i^\top x^k-b_i)\right|\le\delta(i=1,\cdots,n)$, $\left\|\nabla f(x^k)\right\|\le\dfrac{\delta}{m}\|A\|_F$ 和

$0 \le h'_\mu\left(B_j^\top x^k - h_j\right) \le 1 (j=1,\cdots,q)$。

注意，这里只是对非光滑项 $Q(x)=\left\|(Bx-h)_+\right\|_1$ 进行了光滑化，并未对非光滑项 $\Phi(x)$ 进行光滑化。因此，如何求解迭代步中步 1 的子问题 $\min_{x\in\mathbb{R}^n} G_{\lambda_k,\mu_k}(x)$ 是非常关键的。该子问题仍是一个非光滑优化，但 $\Phi(x)$ 的邻近函数具有解析表达式[19]，因此，本文建议采用文献[26]中的非单调邻近梯度(NPG)算法对其进行求解。

定理 4.1：设 x_k 是算法 4.1 生成的序列，则 $\{x_k\}$ 的任何聚点 x^* 都是问题(4)的 d-稳定点，即

$$x^* \in \Omega \text{ 且 } f'\left(x^*; x-x^*\right)+\Phi'\left(x^*; x-x^*\right) \ge 0, \quad \forall x \in \Omega.$$

证明：设 $\{x^k\}$ 的收敛子列 $\{x^k\}_{\mathcal{K}}$，使得当 $k\in\mathcal{K}$，$k\to\infty$ 时，$x^k \to x^*$。

1) 首先证明 x^* 是问题(4)的可行点：

$$\begin{aligned}\left\|(Bx^k-h)_+\right\|_1 &\le Q_{\mu_k}\left(x^k\right)+\frac{q}{2}\mu_k \\ &\le \frac{1}{\alpha_k}G_{\lambda_k,\mu_k}\left(x^k\right)+\frac{q}{2}\mu_k \\ &\le \frac{1}{\alpha_k}G_{\lambda_k,\mu_k}\left(x^{\text{feas}}\right)+\frac{q}{2}\mu_k \\ &= \frac{1}{\alpha_k}f\left(x^{\text{feas}}\right)+\frac{1}{\alpha_k}\Phi\left(x^{\text{feas}}\right)+\frac{q}{2}\mu_k,\end{aligned}$$

故当 $k\in\mathcal{K}$，$k\to\infty$ 时，有 $\left\|(Bx^k-h)_+\right\|_1 \le 0$，即 $x^*\in\Omega$。

2) 其次证明 x^* 是问题(4)的 d-稳定点。定义

$$w_j^k := h'_{\mu_k}\left(B_j^\top x^k - h_j\right), \forall j=1,\cdots,p$$

和

$$I^* := \left\{j\in\{1,\cdots,p\} : B_j^\top x^* - h_j = 0\right\},$$

则 $0\le w_j^k \le 1, \forall j=1,\cdots,p$；当 $j\notin I^*$ 时，$B_j^\top x^* - h_j < 0$，且当 k 充分大时，有 $B_j^\top x^k - h_j < 0$，此时也有 $w_j^k = 0$；当 $j\in I^*$ 时，$B_j^\top x^k - h_j \to B_j^\top x^* - h_j = 0$，$w_j^k = h'_{\mu_k}\left(B_j^\top x^k - h_j\right)\to 0$。因

$$\min_{x\in\mathbb{R}^n}\left\{\left\langle \nabla f\left(x^k\right), x-x^k\right\rangle + \Phi'\left(x^k, x-x^k\right)+\alpha_k\left\langle \nabla Q_{\mu_k}\left(x^k\right), x-x^k\right\rangle\right\} \ge -\epsilon_k, \quad \forall x\in\mathbb{R}^n.$$

再由方向导数的表示，存在 $\zeta^k := \left(\zeta_1^k,\cdots,\zeta_n^k\right)^\top$ 且 $\zeta_i^k \in \partial\phi\left(x_i^k\right), i=1,\cdots,n$，使得

$$\left\langle \nabla f\left(x^k\right)+\zeta^k+\alpha_k\sum_{j\in I^*} w_j^k B_j, x-x^k\right\rangle \ge -\epsilon_k, \quad \forall x\in\mathbb{R}^n. \tag{8}$$

因 $\Phi(x)$ 是全局 lipschitz 的，故 $\{\zeta^k\}$ 都是有界的。由上式及 $\left\{\nabla f\left(x^k\right)\right\}$ 和 $\{\zeta^k\}$ 的有界性，对每个 $j\in I^*$，$\{\alpha_k w_j^k\}$ 都是有界的，否则可取 $\hat{x} = x^k - \left[\nabla f\left(x^k\right)+\zeta^k+\alpha_k\sum_{j\in I^*} w_j^k B_j\right]$，使得

$$\left\langle \nabla f\left(x^k\right)+\zeta^k+\alpha_k\sum_{j\in I^*} w_j^k B_j, \hat{x}-x^k\right\rangle = -\left\|\nabla f\left(x^k\right)+\zeta^k+\alpha_k\sum_{j\in I^*} w_j^k B_j\right\|^2 \to -\infty,$$

与(8)矛盾。因此，不妨设

$$\left\{\zeta^k\right\}_{\mathcal{K}} \to \zeta^* = \left(\zeta_1^*, \cdots, \zeta_n^*\right)^\top \in \partial\Phi\left(x^*\right), \quad \left\{\alpha_k w_j^k\right\}_{\mathcal{K}} \to y_j \in [0, C], \quad j \in I^*,$$

其中 $C > 0$ 为某一常数。在(8)中，由 $\epsilon_k \to 0$，得

$$\left\langle \nabla f\left(x^*\right) + \zeta^* + \sum_{i \in I^*} y_i B_j, x - x^* \right\rangle \geq 0, \quad \forall x \in \mathbb{R}^n.$$

取 $x = x^* - \left[\nabla f\left(x^*\right) + \zeta^* + \sum_{i \in I^*} y_i B_j\right]$，由上式得 $-\left\|\nabla f\left(x^*\right) + \zeta^* + \sum_{i \in I^*} y_i B_j\right\|^2 \geq 0$，故

$$\nabla f\left(x^*\right) + \zeta^* + \sum_{i \in I^*} y_i B_j = 0.$$

由 $x^* \in \Omega = \{x : Bx \leq h\}$，知 $\sum_{i \in I^*} y_i B_j \in N_\Omega\left(x^*\right)$，故 $-\left[\nabla f\left(x^*\right) + \zeta^*\right] \in N_\Omega\left(x^*\right)$。注意到，$\forall x \in \Omega$，有 $x - x^* \in T_\Omega\left(x^*\right)$。因此，

$$\left\langle \nabla f\left(x^*\right) + \zeta^*, x - x^* \right\rangle \geq 0, \quad \forall x \in \Omega.$$

进而，$\forall x \in \Omega$，有

$$\begin{aligned} 0 &\leq \left\langle \nabla f\left(x^*\right) + \zeta^*, x - x^* \right\rangle \\ &\leq \left\langle \nabla f\left(x^*\right), x - x^* \right\rangle + \max_{\zeta \in \partial\Phi\left(x^*\right)} \left\langle \zeta, x - x^* \right\rangle \\ &= f'\left(x^*; x - x^*\right) + \Phi'\left(x^*; x - x^*\right). \end{aligned}$$

上式表明，x^* 是问题(4)的 d-稳定点。

5. 总结

本文研究了基于 Huber 损失的线性不等式约束稀疏优化问题。我们给出了稀疏优化的原问题、松弛问题和惩罚问题等三种模型，在一定条件下分析了三种模型全局最优解的等价性，提出了求解该问题的光滑化惩罚算法，并证明了该算法的收敛性。本文为求解线性不等式约束稀疏优化问题提供了理论和方法基础。下一步将通过数值实验和算例进一步检验算法的实际效果。

基金项目

国家自然科学基金项目(11861020, 12261020)、贵州省高层次留学人才创新创业择优资助重点项目([2018] 03)、贵州省科技计划项目(ZK[2021] 009, [2018] 5781)、贵州省青年科技人才成长项目([2018] 121)。

参考文献

[1] Natarajan, B. (1995) Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, **24**, 227-234. https://doi.org/10.1137/S0097539792240406

[2] Donoho, D. (2006) Compressed Sensing. *IEEE Transactions on Information Theory*, **52**, 1289-1306. https://doi.org/10.1109/TIT.2006.871582

[3] Candès, E., Romberg, J. and Tao, T. (2006) Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory*, **52**, 489-509. https://doi.org/10.1109/TIT.2005.862083

[4] Tibshirani, R. (1996) Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[5] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[6] Zhang, C. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *Annals of Statistics*, **38**, 894-942. https://doi.org/10.1214/09-AOS729

[7] Ong, C. and An, L. (2013) Learning Sparse Classifiers with Difference of Convex Functions Algorithms. *Optimization Methods and Software*, **28**, 830-854. https://doi.org/10.1080/10556788.2011.652630

[8] Peleg, D. and Meir, R. (2008) A Bilinear Formulation for Vector Sparsity Optimization. *Signal Processing*, **88**, 375-389. https://doi.org/10.1016/j.sigpro.2007.08.015

[9] Thi, H., Dinh, T., Le, H. and Vo, X. (2015) DC Approximation Approaches for Sparse Optimization. *European Journal of Operational Research*, **244**, 26-46. https://doi.org/10.1016/j.ejor.2014.11.031

[10] Zhang, T. (2013) Multi-Stage Convex Relaxation for Feature Selection. *Bernoulli*, **19**, 2277-2293. https://doi.org/10.3150/12-BEJ452

[11] Bian, W. and Chen, X. (2017) Optimality and Complexity for Constrained Optimization Problems with Nonconvex Regularization. *Mathematics of Operations Research*, **42**, 1063-1084. https://doi.org/10.1287/moor.2016.0837

[12] Chartrand, R. and Staneva, V. (2008) Restricted Isometry Properties and Nonconvex Compressive Sensing. *Inverse Problems*, **24**, 1-14. https://doi.org/10.1088/0266-5611/24/3/035020

[13] Huang, J., Horowitz, J. and Ma. S. (2008) Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models. *Annals of Statistics*, **36**, 587-613. https://doi.org/10.1214/009053607000000875

[14] Ahn, M., Pang, J. and Xin, J. (2017) Difference-of-Convex Learning: Directional Stationarity, Optimality, and Sparsity. *SIAM Journal on Optimization*, **27**, 1637-1655. https://doi.org/10.1137/16M1084754

[15] An, L. and Tao, P. (2005) The DC (Difference of Convex Functions) Programming and DCA Revisited with DC Models of Real World Nonconvex Optimization Problems. *Annals of Operations Research*, **133**, 23-46. https://doi.org/10.1007/s10479-004-5022-1

[16] Pang, J., Razaviyayn, M. and Alvarado, A. (2017) Computing B-Stationary Points of Nonsmooth DC Programs. *Mathematics of Operations Research*, **42**, 95-118. https://doi.org/10.1287/moor.2016.0795

[17] Chen, X., Niu, L. and Yuan, Y. (2013) Optimality Conditions and Smoothing Trust Region Newton Method for Non-Lipschitz Optimization. *SIAM Journal on Optimization*, **23**, 1528-1552. https://doi.org/10.1137/120871390

[18] Candès, E., Walkin, M. and Boyd, S. (2008) Enhancing Sparsity by Reweighted ℓ_1 Minimization. *Journal of Fourier Analysis and Applications*, **14**, 877-905. https://doi.org/10.1007/s00041-008-9045-x

[19] Bian, W. and Chen, X. (2020) A Smoothing Proximal Gradient Algorithm for Non-Smooth Convex Regression with Cardinality Penalty. *SIAM Journal on Numerical Analysis*, **58**, 858-883. https://doi.org/10.1137/18M1186009

[20] 罗孝敏, 彭定涛, 张弦. 基于 MCP 正则的最小一乘回归问题研究[J]. 系统科学与数学, 2021, 41(8): 2327-2337.

[21] 彭定涛, 唐琦, 张弦. 组稀疏优化问题精确连续 Capped-L1 松弛[J]. 数学学报, 2022, 65(2): 243-262.

[22] Pan, L. and Chen, X. (2021) Group Sparse Optimization for Images Recovery Using Capped Folded Concave Functions. *SIAM Journal on Imaging Sciences*, **14**, 1-25. https://doi.org/10.1137/19M1304799

[23] Zhang, X. and Peng, D. (2022) Solving Constrained Nonsmooth Group Sparse Optimization via Group Capped-ℓ_1 Relaxation and Group Smoothing Proximal Gradient Algorithm. *Computational Optimization and Applications*, **83**, 801-804. https://doi.org/10.1007/s10589-022-00419-2

[24] Peng, D. and Chen, X. (2020) Computation of Second-Order Directional Stationary Points for Group Sparse Optimization. *Optimization Methods and Software*, **35**, 348-376. https://doi.org/10.1080/10556788.2019.1684492

[25] Rockafellar, R. and Wets, R. (2009) Variational Analysis. 3rd Edition, Springer-Verlag, Berlin.

[26] Chen, X., Lu, Z. and Pong, T. (2016) Penalty Methods for a Class of Non-Lipschitz Optimization Problems. *SIAM Journal on Optimization*, **26**, 1465-1492. https://doi.org/10.1137/15M1028054