



ℓ^0 正则化下衰减信号稀疏恢复的 PDASC 算法

张虎¹, 曹永秀^{1*}, 焦雨领¹, 石跃勇^{2,3}

1. 中南财经政法大学统计与数学学院, 武汉 430073

2. 中国地质大学(武汉) 经济管理学院, 武汉 430074

3. 中国地质大学(武汉) 资源环境经济研究中心, 武汉 430074

* 通信作者. E-mail: yxcao@zuel.edu.cn

收稿日期: 2018-04-16; 修回日期: 2018-06-25; 接受日期: 2018-07-23; 网络出版日期: 2019-04-09

国家自然科学基金(批准号: 11501578, 11501579, 11701571, 11801531, 11871474, 41572315) 和国家社科基金(批准号: 17BTJ017) 资助项目

摘要 稀疏恢复的主要任务之一是发展和分析在计算上易于处理的算法以得到欠定线性系统的稀疏解. 焦雨领等(2015)提出了一种 PDASC 算法求解 ℓ^0 正则化最小二乘问题, 并在一定的 RIP 条件下分析了 PDASC 的误差估计, 进而建立了有限步全局收敛性. 对一类具有强衰减特性的信号, 本文将 RIP 常数条件减弱到与稀疏度 T 无关. 此外, 数据驱动的选择准则被提出用来选择适合的正则参数. 数值实验验证了 PDASC 算法和数据驱动参数选择准则的效率和精度.

关键词 连续化, 衰减信号, 全局收敛, 原始对偶积极集, 正则参数选择

1 引言

考虑压缩感知的非凸 ℓ^0 正则极小化问题如下:

$$\min_{x \in \mathbb{R}^p} J_\lambda(x) = \frac{1}{2} \|\Psi x - y\|^2 + \lambda \|x\|_0, \quad (1)$$

其中感知矩阵 $\Psi \in \mathbb{R}^{n \times p}$ ($p \gg n$) 是列单位化的, 即 $\|\psi_i\| = 1$, $\|\cdot\|$ 为通常的欧氏范数或 ℓ^2 范数; $\|x\|_0$ 表示 p 维列向量 x 的非零元个数; y 是一组可能带有噪声的观测值. 在实际中, 由于很多的自然信号/图像都具有某种近似稀疏表示, 故压缩感知被广泛应用于信号处理、图像处理和机器学习等领域^[1~5]. 压缩感知的主要任务之一是设计高效的算法以恢复欠定线性系统 $\Psi x = y$ 的稀疏解. 一种常用的方法是 ℓ^1 凸松弛法, 亦即基追踪去噪 (basis pursuit denoising)^[6] 或 LASSO^[7]. 凸松弛模型的普及主要归因于一方面压缩感知、统计学、反问题等领域的学者们深入地研究了 ℓ^1 正则化理论并在一定条件下证明了 ℓ^1 和 ℓ^0 模型的等价性^[1, 2, 8, 9]; 另一方面, 由于 ℓ^1 模型在数值求解时是凸优化问题, 在计算上发展了很多的快速算法^[10, 11]. 近些年, 非凸模型在理论^[12~14] 和应用^[15~18] 中都显示了区

引用格式: 张虎, 曹永秀, 焦雨领, 等. ℓ^0 正则化下衰减信号稀疏恢复的 PDASC 算法. 中国科学: 信息科学, 2019, 49: 900–910, doi: 10.1360/N112018-00085
Zhang H, Cao Y X, Jiao Y L, et al. Sparse recovery decaying signals based on ℓ^0 regularization via PDASC (in Chinese). Sci Sin Inform, 2019, 49: 900–910, doi: 10.1360/N112018-00085

别于凸模型的优良性. 由于 ℓ^0 模型等价于最优子集回归^[19], 其研究在非凸模型中处于中心位置, 所以需要人们设计高效的算法去求解问题 (1), 同时需要发展有效的准则去选择适合的正则参数 λ . 除了在压缩感知领域, 问题 (1) 的研究在变量选择^[13,19,20]、图像复原^[21~24] 等方面也非常重要.

由于式 (1) 中的目标函数 $J_\lambda(x)$ 是不连续且非凸的, 所以其求解过程是 NP 困难的, 并且非常不容易找到其全局极小值. 现有的大多数求解问题 (1) 的算法, 包括迭代硬阈值 (iterative hard thresholding)^[25] 和惩罚分解 (penalty decomposition)^[26] 等, 至多能够收敛到一个局部极小值或者一个稳定点. 基于全局极小值的一种新的必要性最优条件 (同时也是坐标极小值^[27] 的一种充分性条件), 文献^[28,29] 发展了一种原始对偶积极集 (primal dual active set, PDAS) 算法求解 (1). 结合正则参数 λ 的一种连续化 (continuation) 技术^[30,31], 在 RIP 满足 $\delta_{T+1} \leq \frac{1}{2\sqrt{T+1}}$ 的条件下, T 为稀疏度, 文献^[32] 证明了 PDASC (PDAS with continuation) 算法的全局收敛性.

本文对 PDASC 算法进行研究和分析, 主要的工作和贡献有两方面: (1) 在真实信号满足衰减性质的假设下, 例如从正弦调制的指数信号采样获得的数字信号^[33], 将文献^[32] 中的 RIP 条件 δ_{T+1} 减弱为与稀疏度 T 无关; (2) 3 种后验准则被提出用来选择合适的正则参数. 特别地, 当噪声水平已知时, 采用偏差原则 (the discrepancy principle, DP) 选择正则化参数; 当噪声水平未知时, 采用另外两种分别基于 BIC (Bayesian information criterion) 和积极集大小 AS (the active set size) 的数据驱动准则选择正则化参数.

本文的组织结构如下: 第 2 节介绍 PDASC 算法的问题设置、算法步骤、收敛性分析和正则参数选择准则, 第 3 节通过数值算例验证算法和参数选择准则的有效性, 第 4 节进行总结.

2 PDASC 算法

2.1 问题设置

设真实信号向量 x^* 有 T 个非零分量, $y = \sum_{i \in \mathcal{A}^*} x_i^* \psi_i + \eta$, 其中 \mathcal{A}^* 是真实积极集 (即非零分量指标集), $\mathcal{A}^* = \{i : x_i^* \neq 0\}$, 从而有 $T = |\mathcal{A}^*|$. 不失一般性, 假定 $\mathcal{A}^* = \{1, 2, \dots, T\}$, 且 $|x_1^*| \geq |x_2^*| \geq \dots \geq |x_T^*| > 0$. 设噪声向量 η 满足 $\|\eta\| \leq \epsilon$, 其中 ϵ 为噪声水平. 记 $\mathcal{S} = \{1, 2, \dots, p\}$, $\mathcal{I}^* = \mathcal{S} \setminus \mathcal{A}^*$. 对任意指标集 $\mathcal{A} \subset \mathcal{S}$, 令 $x_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ ($\Psi_{\mathcal{A}} \in \mathbb{R}^{n \times |\mathcal{A}|}$) 表示将向量 x (矩阵 Ψ 的列) 限制在 \mathcal{A} 上得到的子向量 (子矩阵). $\Psi_{\mathcal{A}}^\dagger$ 表示子矩阵 $\Psi_{\mathcal{A}}$ 的广义逆. RIP 常数^[34] $\delta_s \in (0, 1)$ 为满足下式的最小正常数:

$$(1 - \delta_s)\|x\|^2 \leq \|\Psi x\|^2 \leq (1 + \delta_s)\|x\|^2, \quad \forall x \text{ 满足 } \|x\|_0 \leq s.$$

另记 $\|x\|_\infty = \max_{1 \leq i \leq p} |x_i|$ 表示向量 x 的 ℓ^∞ 范数, Ψ^t 表示矩阵 Ψ 的转置.

2.2 算法步骤

文献^[32] 提出了求解 ℓ^0 问题的 PDASC 算法. PDASC 结合了 PDAS 算法^[28,29] 和连续化技术^[30,31] 的长处, 相应地具有了快速局部收敛性和可以控制积极集大小的优点. 为了方便本文的叙述, 回顾 PDASC 算法求解 (1) 的详细步骤如算法 1 所示.

2.3 收敛性分析

为了对算法进行分析, 需要对真实信号 x^* 、感知矩阵 Ψ 和噪声水平 ϵ 作如下假设:

算法 1 PDASC

输入: $\lambda_0 = \frac{1}{2} \|\Psi^t y\|_\infty^2$, $\mathcal{A}(\lambda_0) = \emptyset$, $x(\lambda_0) = 0$, $d(\lambda_0) = \Psi^t y$, 下降因子 $\mu \in (0, 1)$.

1: **for** $k = 1, 2, \dots$ **do**

2: 令 $\lambda_k = \mu \lambda_{k-1}$, $\mathcal{A}_0 = \mathcal{A}(\lambda_{k-1})$, $(x^0, d^0) = (x(\lambda_{k-1}), d(\lambda_{k-1}))$;

3: **for** $j = 1, 2, \dots$ **do**

4: 定义积极集 \mathcal{A}_j 和非积极集 \mathcal{I}_j 分别为

$$\mathcal{A}_j = \left\{ i : |x_i^{j-1} + d_i^{j-1}| > \sqrt{2\lambda_k} \right\}, \quad \mathcal{I}_j = \mathcal{A}_j^c;$$

5: 检查停止准则 $\mathcal{A}_j = \mathcal{A}_{j-1}$. 如果满足停止准则, 令 $\mathcal{A}(\lambda_k) = \mathcal{A}_j$, $(x(\lambda_k), d(\lambda_k)) = (x^{j-1}, d^{j-1})$;

6: 更新原始变量 x^j 和对偶变量 d^j 如下:

$$\begin{cases} x_{\mathcal{I}_j}^j = 0, \\ \Psi_{\mathcal{A}_j}^t \Psi_{\mathcal{A}_j} x_{\mathcal{A}_j}^j = \Psi_{\mathcal{A}_j}^t y, \\ d^j = \Psi^t (\Psi x^j - y); \end{cases}$$

7: **end for**

8: 检查停止准则;

9: **end for**

输出: 解的路径 $\{x(\lambda_k)\}_{k=1,2,\dots}$.

(H1) 真实信号 x^* 满足衰减性

$$\max_{1 \leq i \leq T-1} \frac{|x_{i+1}^*|}{|x_i^*|} = \rho < 1;$$

(H2) 感知矩阵 Ψ 满足 RIP 条件 $\delta \triangleq \delta_T \leq \frac{(1-\rho)\sqrt{1-\rho^2}}{6}$;

(H3) 噪声水平 ϵ 满足 $\epsilon \leq \frac{|x_T^*|(1-\rho)}{30}$.

注释1 条件 (H1) 要求真解信号的强度有指数衰减的趋势. 在实际中, 从由正弦调制的指数信号采样获得的数字信号满足 (H1), 参见文献 [33]. 条件 (H2) 要求采样矩阵 Ψ 满足 RIP 条件, 其中阶数为 T 的 RIP 常数充分小. 根据随机矩阵的理论, 当 Ψ 中的元素独立同分布的来自 $\mathcal{N}(0, \frac{1}{n})$ 时, 只要采样数 $n > \frac{36}{(1-\rho)^3(1+\rho)} T \log \frac{n}{T}$, 条件 (H2) 以高概率成立, 参见文献 [35]. 条件 (H3) 要求噪声水平较小, 即信噪比较大. 这里假设 (H1) 成立, 即本文里面的真解是具有强衰减性的, 由此, 可以得到附录中 (A4)~(A7), 从而把文献 [32] 中的 RIP 条件减弱为与稀疏度无关, 详见附录中的证明.

若噪声水平 ϵ 已知, 设定算法 1 中第 8 行的停止准则为偏差原则, 即

$$\|\Psi x(\lambda_k) - y\| \leq \epsilon. \quad (2)$$

然后有定理 1 关于算法 1 的全局收敛性结果, 证明在附录中给出. 注意到, 对于一类具有快速衰减的信号, 假设 (H2) 中的 RIP 条件与真实信号 x^* 的稀疏度 T 无关, 这减弱了文献 [32] 对采样矩阵 RIP 常数的要求.

定理1 令假设 (H1)~(H3) 成立, 下降因子 $\mu = 0.25$, 则算法 1 在有限 k 步后停止, 且有 $\mathcal{A}(\lambda_k) = \mathcal{A}^*$, $x(\lambda_k) = \Psi_{\mathcal{A}^*}^\dagger y$.

注释2 在满足条件 (H1)~(H3) 的条件下, PDASC 算法可以在有限步内停机, 且此时 PDASC 的输出就是 Oracle 解. 文献 [32] 中定理 3.2 在加速采样矩阵 RIP 常数 $\delta = \mathcal{O}(\frac{1}{2\sqrt{T}+1})$ 且噪声水平不大的条件下得到了类似的结论. 相比较, 本文的 RIP 常数减弱到了与稀疏度 T 无关. 由随机矩阵的理

论^[35]可知, 采样个数 $n \geq \mathcal{O}(T \log \frac{p}{T})$ 时 (H2) 以高概率成立, 而文献 [32] 中对 RIP 常数的条件需要 $n \geq \mathcal{O}(T^2 \log \frac{p}{T})$ 时才能成立. 正交匹配追踪 (orthogonal matching pursuit, OMP) 在一定的条件下也可以得到 Oracle 解^[36,37]. 不过 OMP 对采样矩阵 RIP 常数的要求与文献 [32] 类似. 此外, OMP 需要知道稀疏度 T , 而 PDASC 不需要.

2.4 正则参数选择

本小节考虑正则参数 λ 的选择, 其对于获得良好的信号重建 (reconstruction) 非常重要. 若噪声水平 ϵ 已知, 可以采用偏差原则^[38] 选择参数, 即, 当 $\frac{1}{2} \|\Psi x(\lambda_k) - y\|^2 \leq \epsilon$ 成立时停止 PDASC 并把此时的 $x(\lambda_k)$ 作为输出解. 即选择使得式 (2) 成立的第 1 个 λ_k 作为最优的 λ . 然而, 有时候精确地估计噪声水平并不是一件容易的事情, 此种情形下 (即噪声水平未知), 需要利用数据驱动 (data-driven) 的准则去选择正则参数, 比如可以根据 BIC 准则或 AS 准则去选择适合的 λ . BIC 准则在统计学中应用较为普遍, 因为文献 [39~44] 等在适当条件下证明了其模型选择相合性. 而 AS 准则的提出是源自对 $\{|\mathcal{A}_k|\}_k$ 的启发式观察, 其中 $\mathcal{A}_k = \{i : x_i(\lambda_k) \neq 0\}$. 特别地, 在算法 1 运行过程中, 当其满足 $\|x(\lambda_k)\|_0 \geq \theta n$ 时, 算法被停止, 其中 $\theta \in [0.5, 1)$, 在实际中可固定 $\theta = 0.5$. 然后, 通过以下准则确定合适的 λ 及相应的 $x(\lambda)$.

- BIC 准则. 定义为

$$\text{BIC}(\lambda_k) := \frac{1}{2} \|\Psi x(\lambda_k) - y\|^2 + \frac{\ln p}{n} df_{\lambda_k},$$

其中 $df_{\lambda_k} = \|x(\lambda_k)\|_0$ 表示 $x(\lambda_k)$ 的近似自由度^[45]. 然后选择 $\lambda = \arg \min_{\{\lambda_k\}} \text{BIC}(\lambda_k)$.

- AS 准则. 对任意的 $d = 1, \dots, p$, 记 $S_d = \{\lambda_k : |\mathcal{A}_k| = d\}$, $\bar{d} = \arg \max_d \{|S_d|\}$. 然后选择

$$\lambda = \max_k \lambda_k, \quad \text{s.t. } \lambda_k \in S_{\bar{d}}.$$

3 数值算例

本节通过数值算例来说明算法 1 和参数选择准则的表现. 实验程序用 Matlab 语言编写; 实验环境为双核台式机, 3.16 GHz 处理器, 4 GB 内存.

首先, 给出数据的生成方式和算法 1 的参数设置. 按照文献 [46] 里的方式生成随机 T -稀疏的幅度 (dynamic range, DR) 为

$$\text{DR} := \frac{\max\{|x_i^*| : x_i^* \neq 0\}}{\min\{|x_i^*| : x_i^* \neq 0\}}$$

的真实信号 x^* . 由于随机性, 一个大的 DR 可能意味着 x^* 的衰减更快. 感知矩阵 $\Psi \in \mathbb{R}^{n \times p}$ 为随机 Gauss 矩阵. 观测向量 y 由 $y = \Psi x^* + \eta$ 生成, 其中 η 为 Gauss 噪声, 分量 $\eta_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, σ 为标准差. 记噪声水平为 $\epsilon = \|\eta\|$. 用 $(T, n, p, \text{DR}, \sigma)$ 表示生成的数据集. 下降因子 μ 固定为 $\mu = 0.7$.

其次, 考察算法 1 与信号 x^* 的幅度 DR 之间的相关性. 为此, 对每个 $\text{DR} \in \{1, 10, \dots, 10^5\}$, 按照 $(T, n, p, \text{DR}, \sigma) = (10 : 20 : 390, 200, 1000, \text{DR}, 0.01)$ 生成数据集, 其中 $10 : 20 : 390$ 表示 $10, 30, 50, \dots, 390$. 我们进行 $N = 100$ 次独立实验, 用 DP 准则 (使得式 (2) 成立的第 1 个 λ_k) 选择适合的 λ , 记每次实验得到的估计积极集为 $\hat{\mathcal{A}}^{(m)}$ ($m = 1, 2, \dots, N$), 然后计算 \mathcal{A}^* 的精确恢复概率 (exact recovery probability) 为 $N^{-1} \sum_{m=1}^N I(\hat{\mathcal{A}}^{(m)} = \mathcal{A}^*)$. 算法 1 关于稀疏度 T 对不同 DR 的精确恢复概率的影响如图 1 所示. 由图 1 可以看出, 算法 1 的精确恢复概率会随着信号 DR 的增大而增加. 特别地, 若 $\text{DR} > 10^3$, 即使 $|\mathcal{A}^*| > n/2$, 算法 1 也能完全恢复 \mathcal{A}^* .

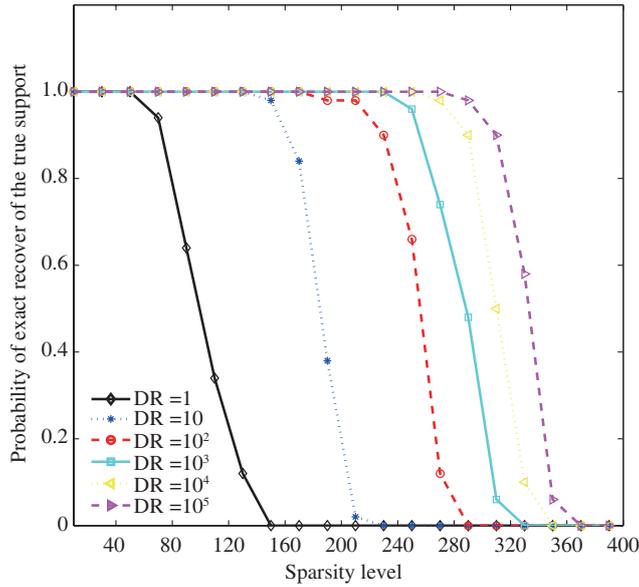


图 1 (网络版彩图) 不同 DR 的精确恢复概率

Figure 1 (Color online) The exact support recovery probability for different dynamical ranges

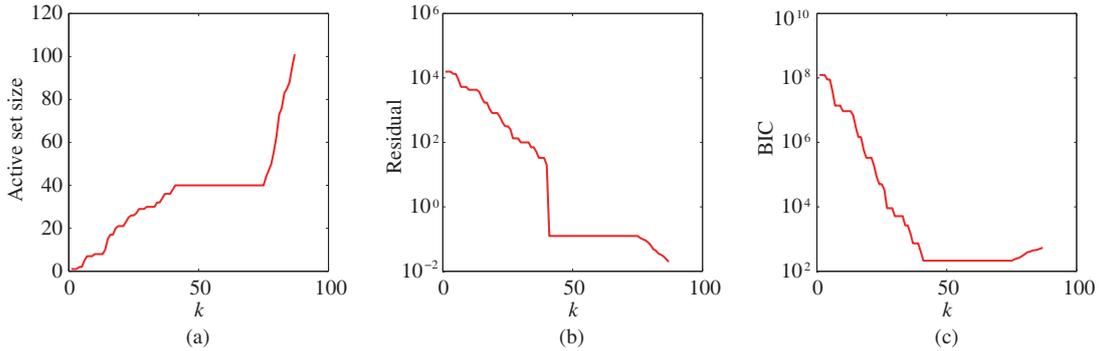
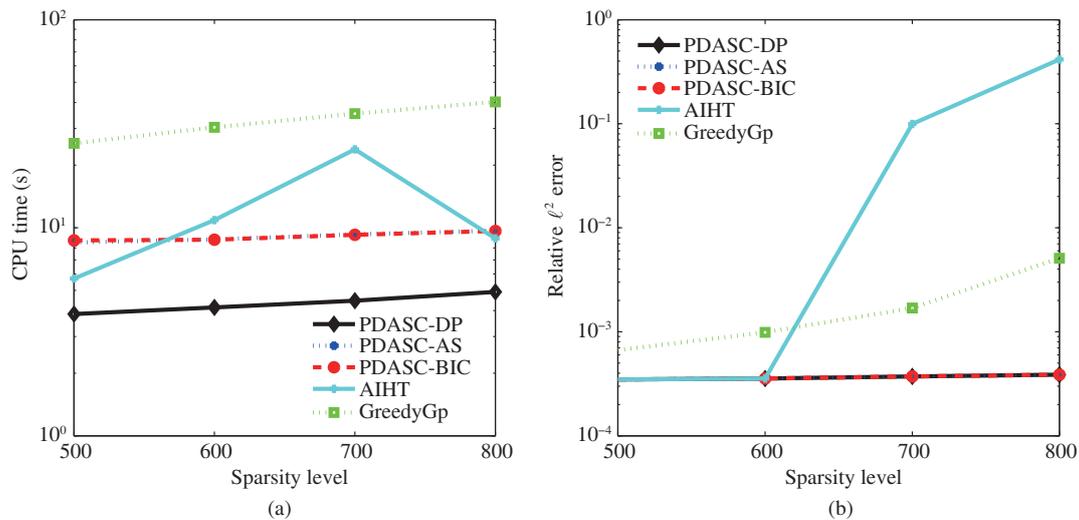


图 2 (网络版彩图) 解路径的 (a) 积极集大小曲线 $|\mathcal{A}_k|$, (b) 残差曲线 $\|\Psi x(\lambda_k) - y\|$, (c) BIC 曲线

Figure 2 (Color online) The size of the active set $|\mathcal{A}_k|$ (a), the residual $\|\Psi x(\lambda_k) - y\|$ (b) and BIC values (c) on the solution path

然后, 考察 2.4 小节的参数选择准则. 为此, 先按照 $(T, n, p, DR, \sigma) = (40, 200, 1000, 10^3, 0.01)$ 生成数据集, 再运行算法 1 直到满足条件 $\|x(\lambda_k)\|_0 \geq n/2$ 为止. 取 $\lambda_{\max} = \frac{1}{2} \|\Psi^t y\|_\infty^2$, $\lambda_{\min} = 1e - 15 \lambda_{\max}$, 格子数 $G = 100$, 从而得到 $\{\lambda_k\}$ 序列为 $\lambda_0 = \lambda_{\max} > \lambda_1 > \dots > \lambda_G = \lambda_{\min}$. 图 2(a)~(c) 分别给出 $|\mathcal{A}_k|$, $\|\Psi x(\lambda_k) - y\|$ 和 BIC 随 λ_k 变化 (因为 $\{\lambda_k\}$ 为有序序列, 亦即随 k 变化) 的关系图. 由图 2(a) 可知, $|\mathcal{A}_k|$ 会随着 λ_k 的减小而逐渐增大, 在若干个 λ_k 上保持值不变 (图形上呈平坦状态), 且相应的 $|\mathcal{A}_k|$ 恰好等于真实积极集 \mathcal{A}^* 的大小, 即 $|\mathcal{A}_k| = T = |\mathcal{A}^*|$. 由图 2(b) 可知, 残差 $\|\Psi x(\lambda_k) - y\|$ 会随着 λ_k 的减小而减小, 在若干个 λ_k 上保持平坦状态之前在曲线上有一个很大的跳跃, 且这些 λ_k 对应的残差要比噪声水平小得多. 由图 2(c) 可知, BIC 曲线同样具有和曲线 (a) 和 (b) 相似的一段平坦状态. 有趣的是, 图 2(a)~(c) 中的 3 段平坦线段是相同的, 均给出了真实信号 x^* 的积极集 \mathcal{A}^* . 以上观察是在 2.4 小节中提出 3 种参数选择准则选取合适 λ 的基础和动机.

图 3 (网络版彩图) (a) CPU 时间; (b) ℓ^2 相对误差Figure 3 (Color online) (a) CPU time; (b) relative ℓ^2 error

最后, 基于 DP, BIC 和 AS 3 种参数选择准则, 考察算法 1 的效率和精度. 为方便计, 分别记为 PDASC-DP, PDASC-BIC 和 PDASC-AS, 并与两种公认的贪婪算法 GreedyGp^[47] 和 AIHT^[48] 进行比较. 数据集按照 $(T, n, p, DR, \sigma) = (500 : 100 : 800, 2 \times 10^3, 10^5, 10^2, 0.01)$ 生成. 实验重复 100 次, 平均 CPU 时间 (以秒计时) 和平均 ℓ^2 相对误差 (relative ℓ^2 error) 如图 3 所示. 由图 3 可知, PDASC-DP, PDASC-BIC 和 PDASC-AS 3 种方法在 CPU 时间和误差方面表现彼此相当, 均要优于另外两种算法.

4 总结

本文研究了文献 [32] 提出的连续化原始对偶积极集 (PDASC) 算法. 对一类具有强衰减特性的信号, 在与积极集大小无关的 RIP 条件下, 建立了算法的全局收敛性. 几种参数选择准则被建议用来选择合适的正则化参数. 数值实验结果充分验证了算法的效率和精度.

致谢 感谢审稿人提出的建设性修改意见. 感谢伦敦大学学院金邦梯教授和武汉大学吕锡亮教授的热心帮助和有益讨论.

参考文献

- 1 Candes E J, Romberg J, Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans Inform Theory*, 2006, 52: 489–509
- 2 Donoho D L. Compressed sensing. *IEEE Trans Inform Theory*, 2006, 52: 1289–1306
- 3 Tropp J A. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans Inform Theory*, 2006, 52: 1030–1051
- 4 Zhang H, Wang Y, Chang X Y, et al. $L_{1/2}$ regularization. *Sci China Inf Sci*, 2010, 40: 412–422 [张海, 王尧, 常象宇, 等. $L_{1/2}$ 正则化. *中国科学: 信息科学*, 2010, 40: 412–422]
- 5 Zhang H, Zhang H. Approximate message passing algorithm for $L_{1/2}$ regularization. *Sci Sin Inform*, 2017, 47: 58–72 [张会, 张海. 基于 AMP 的 $L_{1/2}$ 正则化方法. *中国科学: 信息科学*, 2017, 47: 58–72]
- 6 Chen S S, Donoho D L, Saunders M A. Atomic decomposition by basis pursuit. *SIAM J Sci Comput*, 1998, 20: 33–61

- 7 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Method*, 1996, 58: 267–288
- 8 Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann Stat*, 2006, 34: 1436–1462
- 9 Grasmair M, Scherzer O, Haltmeier M. Necessary and sufficient conditions for linear convergence of ℓ^1 -regularization. *Comm Pure Appl Math*, 2011, 64: 161–182
- 10 Tropp J A, Wright S J. Computational methods for sparse solution of linear inverse problems. *Proc IEEE*, 2010, 98: 948–958
- 11 Rish I, Grabarnik G Y. *Sparse Modeling: Theory, Algorithms, and Applications*. Boca Raton: CRC Press, 2014
- 12 Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*, 2001, 96: 1348–1360
- 13 Zhang C H, Zhang T. A general theory of concave regularization for high-dimensional sparse estimation problems. *Stat Sci*, 2012, 27: 576–593
- 14 Elsener A, Sara V. Sharp oracle inequalities for stationary points of nonconvex penalized M-estimators. 2018. ArXiv:1802.09733v1
- 15 Zeng J S, Fang J, Xu Z B. Sparse SAR imaging based on $L_{1/2}$ regularization. *Sci China Inf Sci*, 2012, 55: 1755–1775
- 16 Zeng J S, Xu Z B, Zhang B C, et al. Accelerated regularization based SAR imaging via BCR and reduced Newton skills. *Signal Process*, 2013, 93: 1831–1844
- 17 Zeng J S, Lin S B, Wang Y, et al. $L_{1/2}$ regularization: convergence of iterative half thresholding algorithm. *IEEE Trans Signal Process*, 2014, 62: 2317–2329
- 18 Huang J, Jiao Y L, Jin B T, et al. A unified primal dual active set algorithm for nonconvex sparse recovery. 2018. ArXiv:1310.1147v4
- 19 Fan J Q, Lv J C. A selective overview of variable selection in high dimensional feature space. *Stat Sin*, 2010, 20: 101–148
- 20 Chang X Y, Xu Z B, Zhang H, et al. Robust regularization theory based on L_q ($0 < q < 1$) regularization: the asymptotic distribution and variable selection consistence of solutions. *Sci Sin Math*, 2010, 40: 985–998 [常象宇, 徐宗本, 张海, 等. 稳健 L_q ($0 < q < 1$) 正则化理论: 解的渐近分布与变量选择一致性. *中国科学: 数学*, 2010, 40: 985–998]
- 21 Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intel*, 1984, 6: 721–741
- 22 Leclerc Y G. Constructing simple stable descriptions for image partitioning. *Int J Comput Vision*, 1989, 3: 73–102
- 23 Yan M, Yang Y, Osher S. Robust 1-bit compressive sensing using adaptive outlier pursuit. *IEEE Trans Signal Process*, 2012, 60: 3868–3875
- 24 Zhang Y, Dong B, Lu Z S. ℓ_0 minimization for wavelet frame based image restoration. *Math Comput*, 2013, 82: 995–1015
- 25 Blumensath T, Davies M E. Iterative thresholding for sparse approximations. *J Fourier Anal Appl*, 2008, 14: 629–654
- 26 Lu Z S, Zhang Y. Sparse approximation via penalty decomposition methods. *SIAM J Optim*, 2013, 23: 2448–2478
- 27 Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J Optim Theory Appl*, 2001, 109: 475–494
- 28 Ito K, Kunisch K. A variational approach to sparsity optimization based on Lagrange multiplier theory. *Inverse Probl*, 2014, 30: 015001
- 29 Jiao Y L, Jin B T, Lu X L. A primal dual active set algorithm for a class of nonconvex sparsity optimization. 2013. ArXiv:1310.1147v1
- 30 Hale E T, Yin W T, Zhang Y. Fixed-point continuation for ℓ_1 -minimization: methodology and convergence. *SIAM J Optim*, 2008, 19: 1107–1130
- 31 Fan Q B, Jiao Y L, Lu X L. A primal dual active set algorithm with continuation for compressed sensing. *IEEE Trans Signal Process*, 2014, 62: 6276–6285
- 32 Jiao Y L, Jin B T, Lu X L. A primal dual active set with continuation algorithm for the ℓ^0 -regularized optimization problem. *Appl Comput Harmonic Anal*, 2015, 39: 400–426
- 33 Proakis J G, Manolakis D G. *Digital Signal Processing: Principles Algorithms and Applications*. New Jersey: Prentice Hall, 1996

34 Candes E J, Tao T. Decoding by linear programming. *IEEE Trans Inform Theory*, 2005, 51: 4203–4215

35 Vershynin R. Introduction to the non-asymptotic analysis of random matrices. 2010. ArXiv:1011.3027

36 Huang S S, Zhu J B. Recovery of sparse signals using OMP and its variants: convergence analysis based on RIP. *Inverse Probl*, 2011, 27: 035003

37 Mo Q, Shen Y. A remark on the restricted isometry property in orthogonal matching pursuit. *IEEE Trans Inform Theory*, 2012, 58: 3654–3656

38 Ito K, Jin B. *Inverse Problems: Tikhonov Theory and Algorithms*. Singapore: World Scientific, 2014

39 Schwarz G. Estimating the dimension of a model. *Ann Stat*, 1978, 6: 461–464

40 Wang H, Li R, Tsai C L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 2007, 94: 553–568

41 Chen J H, Chen Z H. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 2008, 95: 759–771

42 Wang H S, Li B, Leng C L. Shrinkage tuning parameter selection with a diverging number of parameters. *J R Stat Soc-Ser B*, 2009, 71: 671–683

43 Wang L, Kim Y D, Li R Z. Calibrating non-convex penalized regression in ultra-high dimension. *Ann Stat*, 2013, 41: 2505–2536

44 Shi Y Y, Jiao Y L, Yan L, et al. A modified BIC tuning parameter selector for SICA-penalized Cox regression models with diverging dimensionality. *J Math*, 2017, 37: 723–730 [石跃勇, 焦雨领, 严良, 等. 发散维数 SICA 惩罚 Cox 回归模型的一种修正 BIC 调节参数选择器. *数学杂志*, 2017, 37: 723–730]

45 Zou H, Hastie T, Tibshirani R. On the “degrees of freedom” of the lasso. *Ann Stat*, 2007, 35: 2173–2192

46 Becker S, Bobin J, Candés E J. NESTA: a fast and accurate first-order method for sparse recovery. *SIAM J Imag Sci*, 2011, 4: 1–39

47 Blumensath T, Davies M E. Stagewise weak gradient pursuits. *IEEE Trans Signal Process*, 2009, 57: 4333–4346

48 Blumensath T. Accelerated iterative hard thresholding. *Signal Process*, 2012, 92: 752–756

附录 A 基本估计式

首先, 回顾 RIP 常数的一些基本性质¹⁾. 设 A, B 为 S 的不相交子集, 那么

$$\begin{aligned} \|\Psi_A^t \Psi_A x_A\| &\geq (1 \mp \delta_{|A|}) \|x_A\|, \\ \|(\Psi_A^t \Psi_A)^{-1} x_A\| &\leq \frac{1}{1 \mp \delta_{|A|}} \|x_A\|, \\ \|\Psi_A^t \Psi_B\| &\leq \delta_{|A|+|B|}, \\ \|\Psi_A^\dagger y\| &\leq \frac{1}{\sqrt{1 - \delta_{|A|}}} \|y\|, \end{aligned}$$

δ_s 关于 s 是单调递增的.

令假设 (H1)~(H3) 成立. 给定任意 $\mathcal{A} \subset \mathcal{A}^*$, 记 $\mathcal{B} = \mathcal{A}^* \setminus \mathcal{A}$, $\mathcal{I} = S \setminus \mathcal{A}$. 考虑一步迭代

$$x_{\mathcal{A}} = \Psi_{\mathcal{A}}^\dagger y, \quad x_{\mathcal{I}} = 0, \quad d = \Psi^t(y - \Psi x).$$

令 $\bar{x}_{\mathcal{A}} = x_{\mathcal{A}} - x_{\mathcal{A}}^*$, 有

$$\begin{aligned} \bar{x}_{\mathcal{A}} &= (\Psi_{\mathcal{A}}^t \Psi_{\mathcal{A}})^{-1} \Psi_{\mathcal{A}}^t (\Psi_{\mathcal{A}^*} x_{\mathcal{A}^*}^* + \eta - \Psi_{\mathcal{A}} x_{\mathcal{A}}) \\ &= (\Psi_{\mathcal{A}}^t \Psi_{\mathcal{A}})^{-1} \Psi_{\mathcal{A}}^t (\Psi_{\mathcal{B}} x_{\mathcal{B}}^* + \eta), \\ d_j &= \Psi_j^t (y - \Psi_{\mathcal{A}} x_{\mathcal{A}}) = \Psi_j^t (\Psi_{\mathcal{B}} x_{\mathcal{B}}^* + \eta - \Psi_{\mathcal{A}} \bar{x}_{\mathcal{A}}). \end{aligned}$$

因此

$$\begin{aligned} \|\bar{x}_{\mathcal{A}}\| &\leq \frac{\delta_{|A|+|B|}}{1 - \delta_{|A|}} \|x_{\mathcal{B}}^*\| + \frac{1}{\sqrt{1 - \delta_{|A|}}} \epsilon \\ &\leq \frac{\delta_T}{1 - \delta_T} \|x_{\mathcal{B}}^*\| + \frac{1}{\sqrt{1 - \delta_T}} \epsilon. \end{aligned} \tag{A1}$$

1) Needell D, Tropp J A. CoSaMP: iterative signal recovery from incomplete and inaccurate samples. *Appl Comput Harmon Anal*, 2009, 26: 301–321.

现在分别考虑下面两种情形.

情形A1 若 $\mathcal{B} = \emptyset$, 即 $\mathcal{A} = \mathcal{A}^*$, 则 $\|\bar{x}_{\mathcal{A}}\| \leq \frac{1}{\sqrt{1-\delta_T}}\epsilon$, 且

$$\begin{aligned} |d_j| &\leq \epsilon + \delta_{T+1}\|\bar{x}_{\mathcal{A}}\| \leq \epsilon + \frac{\delta_{T+1}}{\sqrt{1-\delta_T}}\epsilon, \quad \forall j \in \mathcal{I}, \\ |x_i| &\geq |x_i^*| - \|\bar{x}_{\mathcal{A}}\| \geq |x_i^*| - \frac{1}{\sqrt{1-\delta_T}}\epsilon, \quad \forall i \in \mathcal{A}. \end{aligned}$$

情形A2 若 $\mathcal{B} \neq \emptyset$, 令 $i_{\mathcal{A}}$ 为 \mathcal{B} 中的最小指标, $c = |x_{i_{\mathcal{A}}}^*|$, 则

$$\begin{aligned} ||d_{i_{\mathcal{A}}}| - c| &= \left| |\Psi_{i_{\mathcal{A}}}^t (\Psi_{i_{\mathcal{A}}} x_{i_{\mathcal{A}}}^* + \Psi_{\mathcal{B} \setminus i_{\mathcal{A}}} x_{\mathcal{B} \setminus \{i_{\mathcal{A}}\}}^* + \eta - \Psi_{\mathcal{A}} \bar{x}_{\mathcal{A}})| - c \right| \\ &\leq \delta_{|\mathcal{B}|} \|x_{\mathcal{B} \setminus \{i_{\mathcal{A}}\}}^*\| + \epsilon + \delta_{|\mathcal{A}|+1} \|\bar{x}_{\mathcal{A}}\|, \end{aligned} \quad (\text{A2})$$

且

$$\begin{aligned} |d_j| &= |\Psi_j^T (\Psi_{\mathcal{B}} x_{\mathcal{B}}^* + \eta - \Psi_{\mathcal{A}} \bar{x}_{\mathcal{A}})| \\ &\leq \begin{cases} \delta_{|\mathcal{B}|+1} \|x_{\mathcal{B}}^*\| + \epsilon + \delta_{|\mathcal{A}|+1} \|\bar{x}_{\mathcal{A}}\|, & \forall j \in \mathcal{I}^* \cap \mathcal{I}, \\ \delta_{|\mathcal{B}|} \|x_{\mathcal{B}}^*\| + |x_j^*| + \epsilon + \delta_{|\mathcal{A}|+1} \|\bar{x}_{\mathcal{A}}\|, & \forall j \in \mathcal{B}, j \neq i_{\mathcal{A}}. \end{cases} \end{aligned} \quad (\text{A3})$$

假设 (H1) 中信号的衰减性蕴含

$$\begin{aligned} \|x_{\mathcal{B}}^*\| &\leq \sqrt{c^2 + c^2 \rho^2 + \dots} \leq \frac{c}{\sqrt{1-\rho^2}}, \\ \|x_{\mathcal{B} \setminus \{i_{\mathcal{A}}\}}^*\| &\leq \frac{c\rho}{\sqrt{1-\rho^2}}. \end{aligned} \quad (\text{A4})$$

根据假设 (H1)~(H3) 和估计式 (A1)~(A3), 有

$$\|\bar{x}_{\mathcal{A}}\| \leq \frac{\delta}{1-\delta} \frac{c}{\sqrt{1-\rho^2}} + (1+\delta)\epsilon \leq \frac{1}{4}(1-\rho)c, \quad (\text{A5})$$

$$||d_{i_{\mathcal{A}}}| - c| \leq \frac{\delta c \rho}{\sqrt{1-\rho^2}} + \epsilon + \delta \|\bar{x}_{\mathcal{A}}\| \leq \frac{1}{4}(1-\rho)c, \quad (\text{A6})$$

$$|d_j| \leq \begin{cases} \frac{1}{4}(1-\rho)c, & \forall j \in \mathcal{I}^* \cap \mathcal{I}, \\ c - \frac{3(1-\rho)}{4}c, & \forall j \in \mathcal{B}, j \neq i_{\mathcal{A}}. \end{cases} \quad (\text{A7})$$

作为上述估计式的直接结果, 有推论 A1.

推论A1 令假设 (H1)~(H3) 成立, 且 $l \leq T$, $\mathcal{A} = \{1, \dots, l\}$, $x_{\mathcal{A}} = \Psi_{\mathcal{A}}^\dagger y$, $d = \Psi^t(y - \Psi_{\mathcal{A}} x_{\mathcal{A}})$, 则 $\min_{i \in \mathcal{A}} |x_i| \geq \|d\|_{\infty}$.

附录 B 定理 1 的证明

定理 1 的证明依赖于引理 A1 和 A2.

引理A1 令假设 (H2) 成立, 且 $l \leq T$, $\mathcal{A}_1 = \{1, \dots, l-1\}$, $\mathcal{A}_2 = \{1, \dots, l\}$, $x_{\mathcal{A}_2}^2 = \Psi_{\mathcal{A}_2}^\dagger y$, $d^1 = \Psi^t(y - \Psi_{\mathcal{A}_1} \Psi_{\mathcal{A}_1}^\dagger y)$, 则 $|x_i^2| \geq |d_i^1|$.

证明 记 $D = \Psi_{\mathcal{A}_1}^t \Psi_{\mathcal{A}_1}$, $b = (b_1, \dots, b_{l-1})^t = \Psi_{\mathcal{A}_1}^t \Psi_l$. 对任意的 $z \in \mathbb{R}^{l-1}$, 有

$$\begin{aligned} z^t(D - bb^t)z &= \|\Psi_{\mathcal{A}_1} z\|^2 - \|\Psi_l \Psi_{\mathcal{A}_1}^t z\|^2 \\ &\geq (1-\delta)\|z\|^2 - \delta^2\|z\|^2 \\ &= (1-\delta-\delta^2)\|z\|^2, \\ &\geq 1 - \delta^t D^{-1}b \geq 1 - \|b\| \|D^{-1}b\| \\ &\geq 1 - \frac{1}{1-\delta} \|\Psi_l^t \Psi_{\mathcal{A}_1}\|^2 \\ &\geq 1 - \frac{\delta^2}{1-\delta} > 0. \end{aligned}$$

故矩阵 $D - bb^t$ 是可逆的, 从而由分块矩阵求逆公式有

$$(\Psi_{\mathcal{A}_2}^t \Psi_{\mathcal{A}_2})^{-1} = \begin{bmatrix} D & b \\ b^t & 1 \end{bmatrix}^{-1} = \begin{bmatrix} (D - bb^t)^{-1} & -\frac{D^{-1}b}{1 - b^t D^{-1}b} \\ -b^t (D - bb^t)^{-1} & \frac{1}{1 - b^t D^{-1}b} \end{bmatrix}.$$

令 $h = \Psi_{\mathcal{A}_1}^t(\sum_{l=1}^T \Psi_l x_l^* + \eta) \in \mathbb{R}^{l-1}$, $c = \Psi_l^t(\sum_{l+1}^T \Psi_l x_l^* + \eta)$, 得

$$\begin{aligned} d_l^1 &= \Psi_l^t \left(\sum_{i=l}^T \Psi_i x_i^* + \eta - \Psi_{\mathcal{A}_1} D^{-1} \Psi_{\mathcal{A}_1}^t \left(\sum_{i=l}^T \Psi_i x_i^* + \eta \right) \right) \\ &= x_l^* + c - b^t D^{-1} h - b^t D^{-1} b x_l^*, \\ x_l^2 &= x_l^* + \left((\Psi_{\mathcal{A}_2}^t \Psi_{\mathcal{A}_2})^{-1} \Psi_{\mathcal{A}_2}^t \left(\sum_{l+1}^T \Psi_l x_l^* + \eta \right) \right)_l \\ &= x_l^* + \frac{c - b^t D^{-1} h}{1 - b^t D^{-1} b} = \frac{d_l^1}{1 - b^t D^{-1} b}. \end{aligned}$$

因此 $|x_l^2| \geq |d_l^1|$. 证毕.

引理A2 令假设 (H1)~(H3) 成立, 且 $\mathcal{A}_0 = \{1, \dots, s\}$, $s < T$, $\mathcal{I}_0 = \mathcal{A}_0^c$, $x_{\mathcal{A}_0}^0 = \Psi_{\mathcal{A}_0}^\dagger y$, $x_{\mathcal{I}_0}^0 = 0$, $d^0 = \Psi^t(y - \Psi)$, 则

$$\|d^0\|_\infty \leq \min_{i \in \mathcal{A}_0} |x_i^0|.$$

再令 $\sqrt{2\lambda} \in [\|d^0\|_\infty, \min_{i \in \mathcal{A}_0} |x_i^0|]$, $\lambda = 0.25\hat{\lambda}$, 则算法 1 中第 2~7 行在有限步后终止, 且满足停止准则 $\mathcal{A}_j = \mathcal{A}_{j-1}$.

证明 第 1 个结论直接由推论 A 可得, 我们只需证明第 2 个结论. 令 $\sqrt{2\lambda} \in [|x_{q+1}^*|, |x_q^*|]$, $i_k = \min\{i : i \in \mathcal{I}_k\}$, 其中 q 为正整数且 $1 \leq q < T$. 为了证明的方便, 记

$$a = |x_{q+1}^*|, \quad d = |x_q^*|, \quad b = a + \frac{d-a}{4}, \quad c = d - \frac{d-a}{4}.$$

接下来, 分别讨论以下 3 种情形:

(i) $\sqrt{2\lambda} \in (b, c)$. 可以验证: $i_k \leq q \Rightarrow i_{k+1} > i_k$, $i_k \geq q+1 \Rightarrow \mathcal{A}_{k+1} = \{1, 2, \dots, q\}$, $\mathcal{A}_k \subset \mathcal{A}^*$, 且 $\{1, 2, \dots, q\}$ 是稳定的.

(ii) $\sqrt{2\lambda} \in [a, b]$. 可以验证: $i_k \leq q \Rightarrow i_{k+1} > i_k$, $i_k \geq q+1 \Rightarrow \mathcal{A}_{k+1} = \{1, 2, \dots, q\}$ 或 $\mathcal{A}_{k+1} = \{1, 2, \dots, q, q+1\}$. 有限步后, \mathcal{A}_k 为 $\{1, 2, \dots, q\}$ 或 $\{1, 2, \dots, q, q+1\}$. 由引理 B, 若 $\sqrt{2\lambda} < x_{q+1}$, 则 $\{1, 2, \dots, q+1\}$ 是稳定的; 若 $\sqrt{2\lambda} \geq d_{q+1}$, 则 $\{1, 2, \dots, q\}$ 是稳定的. 收敛性得证.

(iii) $\sqrt{2\lambda} \in [c, d]$. 类似的讨论可知 $\{1, 2, \dots, q\}$ 或 $\{1, 2, \dots, q-1\}$ 是稳定的. 此外, 可以证明 $q > s$.

若 $q = T$, 只需证明 $i_k \leq T-1 \Rightarrow i_{k+1} > i_k$ 和 $\mathcal{A}_k \subset \mathcal{A}^*$, 然后 \mathcal{A}_k 可达 $\{1, \dots, T-1\}$ 或 $\{1, 2, \dots, T\}$. 类似的讨论可得算法停止.

下面证明定理 1.

证明 根据引理 A2, 积极集 \mathcal{A}_k 是单调递增的, 但其总是包含于集合 \mathcal{A}^* 中, 则当 $\sqrt{2\lambda} \leq \frac{3}{4}|x_T^*|$ 时, 积极集 \mathcal{A} 必然与 \mathcal{A}^* 相等, 此时算法停止.

Sparse recovery decaying signals based on ℓ^0 regularization via PDASC

Hu ZHANG¹, Yongxiu CAO^{1*}, Yuling JIAO¹ & Yueyong SHI^{2,3}

1. School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China;
2. School of Economics and Management, China University of Geosciences, Wuhan 430074, China;
3. Research Center of Resource and Environmental Economics, China University of Geosciences, Wuhan 430074, China

* Corresponding author. E-mail: yxcao@zuel.edu.cn

Abstract One of the main tasks for sparse recovery is to develop and analyze computationally tractable algorithms for obtaining sparse solutions of underdetermined linear systems. Jiao et al. (2015) proposed a primal-dual active set with a continuation (PDASC) algorithm for the ℓ^0 -regularized least-squares problem and established finite step global convergence and obtained a sharp estimation error under a certain restricted isometry property (RIP) condition. In this paper, we relax the condition on the RIP constant to be independent of the sparsity level T for a class of signals with a strong decay property. Furthermore, we propose a novel data-driven regularization parameter selection rule. Several numerical examples are presented to verify the efficiency and accuracy of the PDASC algorithm and the data-driven parameter choice rule.

Keywords continuation, decaying signals, global convergence, primal-dual active set, regularization parameter selection



Hu ZHANG was born in 1963. He received his Ph.D. degree from the Zhongnan University of Economics and Law, Wuhan in 1997. He is currently a professor at the School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China. His main research interests include finance and econometrics, risk management, statistical modeling and computing, and applications of statistics in business and economics.



Yongxiu CAO was born in 1985. She received her Ph.D. degree from Wuhan University, Wuhan in 2015. She is currently a lecturer at the School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China. Her main research interests include survival analysis, high-dimensional data analysis, and statistical computing.



Yuling JIAO was born in 1986. He received his Ph.D. degree from Wuhan University, Wuhan, in 2014. He is currently an associate professor at the School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China. His main research interests include signal/image processing, inverse problem, sparse optimization, machine learning, statistical computing, and fast stochastic, parallel, and distributed algorithms.



Yueyong SHI was born in 1984. He received his Ph.D. degree from Wuhan University, Wuhan, in 2013. He is currently a lecturer at the School of Economics and Management, China University of Geosciences, Wuhan, China. His main research interests include semiparametric models, high-dimensional data analysis, machine learning, and statistical computing.