基于稀疏图码的离群值去除

本科毕业论文（设计）答辩

庄启源

指导教师：李朋　副教授
兰州大学　数学与统计学院

2024 年 5 月 23 日

- 姓名：庄启源
- 指导教师：李朋　副教授
- 专业：数学（基础理论班）
- 邮箱：zhuangqy20@lzu.edu.cn

- 论文题目：基于稀疏图码的离群值去除
- 关键词：压缩感知；离群值；稀疏信号恢复；稀疏图码

- **主要内容：无噪声情形下，当观测信号存在少量离群值时稀疏信号的精确恢复。**
- **压缩感知问题：** 对一个稀疏信号 $\mathbf{x} \in \mathbb{R}^N$，使用一个测量矩阵 $\mathbf{A}$ 采样后得到一个观测向量 $\mathbf{y} \in \mathbb{R}^M$，再由该观测值估计初始信号。
- **拓展问题：** 考虑观测值中出现离群值的情况，观测过程可表示为

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{f},$$

其中 $\left|\mathrm{supp}(\mathbf{f})\right| = T = \eta M << M$，且 $\mathbf{f}$ 中非零元素的分布方差显著大于 $\mathbf{x}$ 中的非零元素。

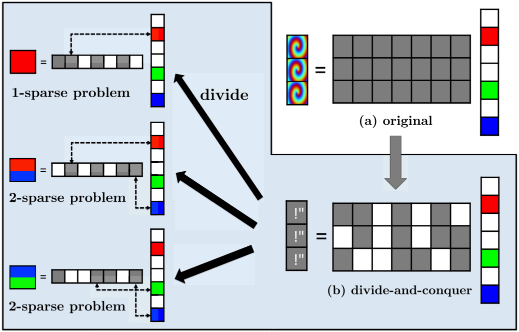
次线性时间的支撑集恢复

- 目的：测量值数目 M 尽可能小；恢复信号所需的时间复杂度尽可能小。
- Li、Yin[1]等人建立了一种基于稀疏图码的压缩感知框架，能够在无噪声情况下，用 $2K$ 次测量精确地恢复任意稀疏度为 K 的信号，解码时间复杂度达到 $O(K)$。
- Bakshi、Jaggi[2]等人提出的 SHO-FA 算法使有噪声情形下的信号恢复仅需 $O(K)$ 次测量且编码解码时间复杂度分别为 $O(N)$ 和 $O(K)$。

[1] Xiao Li et al. “Sub-Linear Time Support Recovery for Compressed Sensing Using Sparse-Graph Codes”. In: *IEEE Transactions on Information Theory* 65.10 (2019), pp. 6580–6619. DOI: 10.1109/TIT.2019.2921757.

[2] Mayank Bakshi et al. “SHO-FA: Robust Compressive Sensing With Order-Optimal Complexity, Measurements, and Bits”. In: *IEEE Transactions on Information Theory* 62.12 (2016), pp. 7419–7444. DOI: 10.1109/TIT.2015.2466601.

稀疏图码



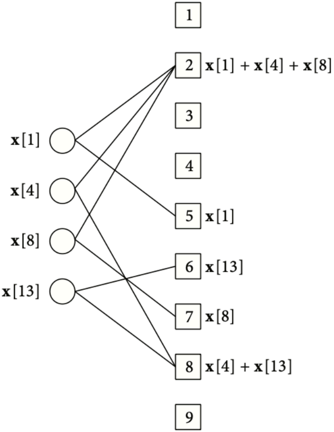
分治法：设计的测量矩阵将稀疏度为 3 的恢复问题分解为多个稀疏度小于 3 的子问题，再逐步剥离解决问题。

存在的问题

- 现有的基于稀疏图码的信号重构策略能快速精确恢复信号，但无法解决存在离群值时的问题。
- 能够解决离群值问题的算法无法达到次线性时间复杂度。

正则度为 2 的左正则二部图邻接矩阵 $\mathbf{H}$
右图表示 $\mathbf{y} = \mathbf{H}\mathbf{x}$

- **零节点**：不包含任何非零元素的右节点，如 y_1。
- **单节点**：仅包含一个非零元素的右节点，如 y_5。
- **多节点**：包含一个以上非零元素的右节点，如 y_2。



- **检测矩阵 S**:

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & W & W^2 & W^3 & W^4 & \dots & W^{15} \end{bmatrix},$$

其中 $W = e^{i\frac{2\pi}{N}}$ 是一个 N 次单位根，$\mathbf{S}$ 是一个 $N \times N$ 的 DFT 矩阵的前两行。

- **行张量算子 $\boxtimes$**:

$$\mathbf{A} = \mathbf{H} \boxtimes \mathbf{S} = [\mathbf{h}_0 \otimes \mathbf{s}_0, \cdots, \mathbf{h}_{N-1} \otimes \mathbf{s}_{N-1}],$$

其中 $\otimes$ 是一个克罗内克积。

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix},$$

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & W & W^2 & W^3 & \dots & W^6 \end{bmatrix}.$$

$\Rightarrow$

$$\mathbf{H} \boxtimes \mathbf{S} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & W & 0 & W^3 & 0 & W^5 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & W & 0 & W^3 & 0 & 0 & W^6 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & W^3 & W^4 & 0 & W^6 \end{bmatrix}.$$

节点类型检测

测量对：二维向量 $\mathbf{y}_r = [y_r[0], y_r[1]]^T$，例如 $\mathbf{y}_5 = x[1] \times \begin{bmatrix} 1 \\ W \end{bmatrix}$.

比值检测：

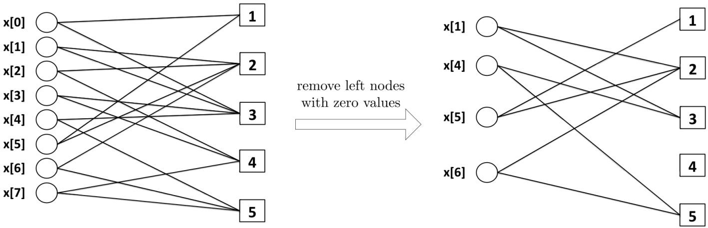
$$\hat{k} = \frac{\angle y_r[1]/y_r[0]}{2\pi/N}.$$

- **零节点对**：测量对是一个全 0 向量，即 $\mathbf{y}_r = \mathbf{0}$.
- **单节点对**：$\hat{k}$ 为整数。
- **多节点对**：$\hat{k}$ 不为非零整数或两个元素不同模 $|y_2[1]| \neq |y_2[0]|$。例如，

$$\hat{k} = \frac{\angle y_2[1]/y_2[0]}{2\pi/16} = 4.85995.$$

剥离解码步骤

- 找到所有零节点（对）并剥离所有相应的左右节点对和边。
- 选取二部图中所有使得右节点度数为 1 的边（找到所有单节点），恢复对应左节点。
- 去除（剥离）所有这些边以及与之对应的左右节点对。
- 去除（剥离）上一步骤中剥离的左节点的其他未被剥离的边。
- 找到上一步骤中去除的左节点连接的所有右节点，将这些左节点的值从右节点的值中减去。



把 $\mathbf{y}$ 的有效非零元素限制在一个指标集

$$S = \{s : |y_s| < \alpha\theta_p(\{|y_j| : j \in \mathrm{supp}(\mathbf{y})\})\}$$

上，其他指标的对应元素全部设为 0。

算法 1 离群值去除算法

输入： 观测值 $\mathbf{y} \in \mathbb{C}^{2R}$（测量对形式），常数 α 和 p。

1: 首先计算得到 $\mathbf{y}$ 的非零元素模的分位数 $q = \theta_p(\{|\mathbf{y}_j[0]| : \mathbf{y}_j[0] \neq 0, \mathbf{y}_j \in \mathbf{y}\})$;

2: **for** $r = 1$ to R **do**

3:     **if** $|\mathbf{y}_r[0]| > \alpha \cdot q$ **then**

4:         将 $\mathbf{y}_r$ 抹去为 $\mathbf{0}$ 向量;

5:     **end if**

6: **end for**

PAMQ 算法

- **两阶段算法**：先去除离群值，再恢复信号。在离群值去除过程中，我们把分位数乘上一定倍率作为分割点；在信号恢复中，我们应用稀疏图码中的剥离解码方法。
  因此我们把这一算法命名为 PAMQ（Peeling after Multiplied Quantile）算法。
- **创新点**：将基于稀疏图码的信号恢复算法应用到存在离群值场景，与自定义的离群值去除技巧结合，形成全新有效算法。

实验设置

评价指标**精确恢复率**：多次重复实验中恢复结果误差 err < thr_err 的次数与总实验次数的比值。其中，

$$err = \frac{||\mathbf{x} - \hat{\mathbf{x}}||_2}{||\mathbf{x}||_2},$$

thr_err = 10^{-2} 或 10^{-7}。

- 初始稀疏信号 $\mathbf{x} \in \mathbb{R}^{200}$ 且其中非零元素 $x_i \sim N(0, 1)$;
- 离群向量 $\mathbf{f}$ 中的非零元素 $f_j \sim N(0, 100)$;
- 编码矩阵 $\mathbf{H}$ 为每列有且仅有三个 1 的矩阵，即正则度为 3 的左正则二部图邻接矩阵;
- 检测矩阵 $\mathbf{S}$ 为 $N \times N$ 的 DFT 矩阵前两行。

表 1: 不同测量次数下的恢复效率

测量值数目 M	10	20	40	80	100	120
精确恢复率	70.0%	84.5%	90.5%	92.5%	95.5%	95.0%
平均用时（毫秒）	48.55	79.25	89.70	95.40	115.25	126.50

表 2: 不同信号稀疏度下的恢复效率

稀疏度 K	1	2	3	4	5
精确恢复率	98.0%	95.4%	92.2%	91.8%	88.8%
平均用时（毫秒）	124.0	129.0	129.6	125.4	130.4

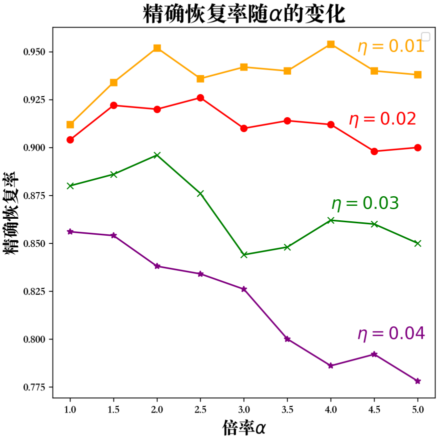


图 1: 不同离群值比例下的恢复效果随算法参数变化展示。图中横坐标表示不同的倍率 α，纵坐标表示精确恢复率，四条曲线分别表示离群值比例 η 取 1%, 2%, 3%, 4% 的情形。

与其他算法的比较

选取对比算法为 ℓ_1-PLAD 算法（Penalized Least Absolute Deviation）。

$$\min_{\mathbf{x}\in\mathbb{R}^N} ||\mathbf{y}-\mathbf{A}\mathbf{x}||_1+\lambda||\mathbf{x}||_1$$

其中 $\mathbf{A}$ 是一个 $M\times N$ 的高斯测量矩阵。

表 3: PAMQ 算法与 PLAD 算法的对比

应用算法	PAMQ	$\text{PLAD}_{\lambda=0.006}$	$\text{PLAD}_{\lambda=0.007}$	$\text{PLAD}_{\lambda=0.008}$	$\text{PLAD}_{\lambda=0.009}$
$\eta=0.02$	92.0%	78.2%	94.6%	91.2%	79.8%
$\eta=0.03$	89.8%	47.8%	66.4%	92.2%	92.0%
平均用时（毫秒）	129.0	1729.9	1268.3	1692.2	2374.5

目录

- 恢复仅需极短的 CPU 时间，速度是 PLAD 算法的十倍。
- 能够适应少量的离群值，恢复比 PLAD 更精确，能够使 $\mathrm{err} < 10^{-7}$。
- 对不同的情况更具一般性和稳定性，不会因算法参数（α, p）变化产生极大波动。

谢　谢！