

基于 GUIG-MCP 的快速低秩矩阵补全算法

李慧勇¹ 马鑫¹ 邱枫¹ 顾强¹ 廖亮^{2*} 赵杰² 陈花竹² 于洪年³ 谭恒⁴

LI Huiyong MA Xin QIU Feng GU Qiang LIAO Liang ZHAO Jie CHEN Huazhu YU Hongnian TAN Heng

摘要

谱正则化是一种广泛使用的通过正则化矩阵奇异值来恢复低秩矩阵 (low-rank matrix recovery, LRMR) 的方法, 主要应用于图像恢复、推荐系统等问题上, 大多数现有的 LRMR 求解器通过在稠密矩阵上应用奇异值分解 (SVD) 来迭代计算奇异值, 但避免不了高昂的计算成本且限制了大规模问题的应用。文章旨在突破现有 LRMR 算法的计算成本高, 计算效率低的问题。提出一种基于 GUIG-MCP (generalized unitarily invariant gauge minimax concave penalty) 的 LRMR 算法。首先, 将问题表示为双线性变分问题, 同时引入了 Nesterov 动量与幂策略加速方法, 设计了基于加速近端交替线性最小化 (Accelerate-PALM) 的加速迭代流程, 避免了显式的 SVD 计算, 降低了计算复杂度并提升了恢复效率。其次, 引入 MCP 嵌入广义酉不变函数 (GUIG) 作为非凸正则化项, 利用 MCP 的闭式近似收缩特性, 为因子矩阵的列范数构造了高效的闭式解, 保证了非凸优化的数值稳定性与近似无偏性。通过对模拟数据以及公开的数据集进行广泛实验, 在不同丢失率的场景下, GUIG-MCP 算法模型在图像恢复的时间以及细节的处理上都具有优势。

关键词

低秩矩阵补全; GUIG 模型; MCP 正则化; 加速近端交替线性最小化; 闭式近端解; 非凸优化

doi: 10.3969/j.issn.1672-9528.2026.06.027

0 引言

低秩矩阵补全问题^[1]广泛出现在图像恢复、推荐系统、视频去噪等领域。其核心目标是仅在观测部分矩阵元素的前提下, 恢复原始的低秩矩阵结构。该问题本质上是一个非凸、NP 难的问题, 直接求解矩阵秩极为困难, 因此许多研究工作者采用对秩函数的替代正则化形式进行近似, 如核范数、Schatten- p 范数等^[2]。从数学上讲矩阵的谱正则化问题可以表示为:

$$\min_{X \in \mathbb{R}^{m \times n}} F(X) = f(X) + g \circ \sigma(X) \quad (1)$$

式中: f 是 $\mathbb{R}^{m \times n}$ 中具有 L_f -Lipschitz 连续梯度的有界光滑函数, $\sigma(X)$ 是在 $\mathbb{R}^{m \times n} \rightarrow \mathbb{R}^d$ 中对矩阵 X 提取的奇异值, g 是在 $\mathbb{R}^d \rightarrow \mathbb{R}$ 中的一个稀疏正则化函数, 符号 $\circ$ 是 g 作用于 $\sigma(X)$, 可表示为 $g \circ \sigma(X) = \sum_{i=1}^r g(\sigma_i(X))$ 。

但这类方法需要在每次迭代中计算奇异值 (SVD) 分解, 对稠密矩阵的复杂度高 $O(mn, \min(m, n))$, 同时存储矩阵 X 为

$O(mn)$ 计算成本较为高昂。

为了解决上述的问题, Jia 等^[3]提出了广义酉不变规范函数 (Generalized Unitarily Invariant Gauge) 的算法框架, 该框架将一类广义非凸谱函数统一建模为奇异值函数的泛化形式, 有效地覆盖了核范数、Schatten- p 范数等多个低秩先验。在此基础上进一步摆脱了对 SVD 的依赖, 通过双线性分解与无 SVD 更新策略降低了计算复杂度。然而, GUIG 框架中的正则函数 $g(x)$ 选择对恢复性能具有重要的影响, 当前缺乏一种在保持稀疏性、非偏性、连续性等统计性质的同时, 又具备高效闭式近端操作的非凸正则。

本文在 GUIG 框架下引入具有良好稀疏性和近似无偏估计能力的最小最大凹惩罚函数 (minimax concave penalty, MCP)^[4], 提出了一种新的 GUIG-MCP 低秩矩阵补全算法。MCP 在向量稀疏建模中已经被证明兼具稀疏性与无偏性质具有明确的近端映射表达式, 具备良好的优化可解性, 利用 MCP 的闭式近端算子, 能对因子矩阵列范数进行高效更新, 并保留数值稳定性。同时, 本文提出一种结合 Nesterov 动量与幂策略加速方法, 基于 PALM 的加速近端交替线性最小化 (Accelerate-PALM) 迭代策略。本文提出的方法避免了显式的 SVD 操作, 极大降低了计算复杂度, 提升了恢复效率。

1 广义酉不变规范 (GUIG) 理论基础

1.1 GUIG 定义与低秩等价性

GUIG (generalized unitarily invariant gauge) 的概念由

1. 中原工学院信息与通信工程学院 河南郑州 450007

2. 中原工学院数学与信息科学学院 河南郑州 450007

3. 爱丁堡龙比亚大学 苏格兰爱丁堡 EH141DJ

4. 中原工学院经济管理学院 河南郑州 450007

[基金项目] 国家自然科学基金 (62202513); 中原工学院“机器智能与高维数据分析”科研团队发展计划 (K2022TD001); 河南省科学技术厅的河南省杰出外籍科学家工作室项目 (GZS2022012)

Jia 等^[3]提出,旨在解决大多数 LRMR 求解器在稠密矩阵上需要应用奇异值分解来进行恢复,从而造成计算成本过于高昂的问题。

GUIG 函数是一种矩阵低秩度量函数。给定矩阵 $X \in \mathbb{R}^{m \times n}$, $G_g(\cdot): \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^d$, 其 GUIG 定义为:

$$G_g(X) = \inf \left\{ \sum_{i=1}^d g(\|\lambda_i\|): X = \sum_{i=1}^d \lambda_i u_i v_i^T, \|u_i\|_2 = \|v_i\|_2 = 1 \right\} \quad (2)$$

式中: $\text{rank}(X) \leq d \leq \min\{m, n\}$, 且 $g(\cdot): \mathbb{R} \rightarrow \mathbb{R}$ 为凹函数。与传统的谱函数不同, $G_g(\cdot)$ 不直接作用于奇异值,但是可以通过低秩等价性定理关联奇异值。

给定矩阵 $X \in \mathbb{R}^{m \times n}$, 若 $g \circ \sigma(X) = \sum_{i=1}^r g(\sigma_i(X))$ 且 $\text{rank}(X) \leq d$, 则表示为:

$$G_g(\cdot) = \sum_{i=1}^r g(\sigma_i(X)) \quad (3)$$

如果函数 g 可以满足以下条件则可以成立:

- (1) g 是一个凹的并且能够在 $(0, +\infty)$ 单调增;
- (2) $\theta(t) = g(e^t)$ 是一个凸函数。

从这一等价性可以看出 GUIG 框架统一了多种谱正则化方法。

1.2 双线性变分等价

虽然谱函数直接作用于奇异值形式清晰,但在优化时不能避免计算, GUIG 可以转化为双线性分解的形式,给定 GUIG 的函数 G_g , 当 g 满足以下定理时就可以很容易构造出双线性分解的形式。

若存在函数 g_1 和 g_2 使得 $g(z) = \min g_1(a) + g_2(b)$, $z = ab$, 则其双线性形式可以表示为:

$$G_g(X) = \min_{X=UV^T} \sum_{i=1}^d g_1(\|U_{:,i}\|_2) + \sum_{i=1}^d g_2(\|V_{:,i}\|_2) \quad (4)$$

式中: $U_{:,i}$ 和 $V_{:,i}$ 是矩阵 U 和 V 的第 i 列。

该模型将原来的问题转化为对因子矩阵 U 和 V 的优化问题,其优势在于:

- (1) 避免了显示的 SVD 计算,只需要矩阵乘法;
- (2) 存储的复杂度从 $O(mn)$ 降为 $O((m+n)d)$, $d \ll \min(m, n)$ 。

2 MCP 正则化的 GUIG 扩展

2.1 MCP 函数阈值特性分析

MCP (minimax concave penalty) 这一正则化方法主要用于解决高维统计中的变量选择问题,旨在平衡 Lasso 方法的偏差问题,其函数形式为分段函数,当变量绝对值小于阈值时呈现凹性,超过阈值后惩罚不再增加,从而避免过度压缩参数,广泛应用于机器学习,图像处理等领域^[5]。

给定参数 $\lambda > 0$, $\theta > 0$, 则 MCP 函数定义为:

$$\phi(x) = \sum_{i=1}^n \phi(\theta, \lambda, x_i) \quad (5)$$

其中:

$$\phi(\theta, \lambda, x_i) = \lambda \int_0^{|x_i|} \left(1 - \frac{\tau}{\theta \lambda}\right) d\tau = \begin{cases} \lambda |x_i| - \frac{\lambda^2}{2\theta}, & |x_i| \leq \theta \lambda \\ \frac{1}{2} \theta \lambda^2, & |x_i| > \theta \lambda \end{cases} \quad (6)$$

根据上述公式,可以得出 MCP 具有以下性质:

- (1) $\phi(\theta, \lambda, t)$ 关于 t 是偶函数;
- (2) MCP 在 $t \in [0, \theta \lambda]$ 上是光滑的,单调递增的。

事实上,在 $[0, \theta \lambda]$ 上 $\phi(\theta, \lambda, t)$ 关于 t 是开口向下的二次函数,其对称轴为 $t = \theta \lambda$,并在 $t = \theta \lambda$ 取得最大值;在 $[\theta \lambda, +\infty)$ 上 $\phi(\theta, \lambda, t)$ 关于 t 是常数函数。这种特点使其方便用来逼近 L_0 范数,此时 $\lambda > 0$ 是惩罚参数, $\theta > 0$ 是调节参数。

2.2 GUIG-MCP 模型构建与理论优势

将第一节中 GUIG 的谱正则 $G_g(X) = \sum_{i=1}^r g(\sigma_i(X))$ 中的泛函数 g 特化为 MCP, 即:

$$G_g(X) = \sum_{i=1}^r \phi_{\theta, \lambda}(\sigma_i(X)) \quad (7)$$

则根据 GUIG 与谱函数等价可知该构造满足 GUIG 的凹性与单调条件。

进一步借助式(4)的双线性表示,引入因子分解,可等价优化问题,即:

$$G_{\text{MCP}}(X) = \min_{X=UV^T} \sum_{i=1}^d \phi_{\theta, \lambda}(\|U_{:,i}\|_2) + \sum_{i=1}^d \phi_{\theta, \lambda}(\|V_{:,i}\|_2) \quad (8)$$

矩阵补全的目标是从其部分观察中完成一个低秩矩阵。

由式(8)将矩阵补全的问题表述为:

$$G_{\text{MCP}}(X) = \min_{U, V} \frac{1}{2} \left\| \rho_{\Omega} (UV^T - M) \right\|_F^2 + \sum_{i=1}^d \phi_{\theta, \lambda}(\|U_{:,i}\|_2) + \sum_{i=1}^d \phi_{\theta, \lambda}(\|V_{:,i}\|_2) \quad (9)$$

式中: $\rho_{\Omega} = \begin{cases} S_{i,j}, & \text{if } (i, j) \in \Omega \\ 0, & \text{otherwise} \end{cases}$ 。

在 GUIG-MCP 中在避免了显示 SVD 计算的同时,具有近似无偏性对小奇异值或列范数有明显压缩作用,促使产生低秩解,在恢复中具有显著的优势,相较于核范数或者 Schatten- p 范数,其设计更加贴合稀疏低秩结构的自然分布。

2.3 因子矩阵列范数子问题闭式解推导

在近端线性交替最小化迭代中,每一步需要交替更新 U 和 V 。以固定 V 更新 U 为例,将第 i 列的问题表述为:

$$\min_{u \in \mathbb{R}^m} \frac{1}{2} \|u - \tilde{u}\|_2^2 + \phi_{\theta, \lambda}(\|u\|_2) \quad (10)$$

式中: $\tilde{u}$ 为梯度线性化之后的中间变量,根据此前计算的阈值特性,其解即:

$$\text{prox}_{\phi}(\tilde{u}) = \begin{cases} \frac{\|x\|_2 - \lambda}{1 - \frac{1}{\theta}} \cdot \frac{x}{\|x\|_2}, & \lambda < \|x\|_2 \leq \theta \lambda \\ x, & \|x\|_2 > \theta \lambda \end{cases} \quad (11)$$

同理可得对 V 的更新闭式计算。由此,近端交替线性最小化全过程只需要矩阵进行上述阈值操作,即可高效推进,避免了显式的 SVD 计算。

3 基于 Accelerated-PALM 的优化算法

为了有效求解 CUIG-MCP 所构建的双线性非凸优化模型, 本文采用基于 PALM (proximal alternating linearized minimization) 框架的交替更新策略^[6], 同时引入两种加速机制: Nesterov 动量 (提升外循环收敛速度) 与幂方法 (压缩奇异空间, 避免高秩冗余), 以实现整体计算效率与数值稳定性的显著提升。

3.1 Nesterov 动量加速外循环收敛

传统的 PALM 方法在每轮更新中以“当前变量”为基础进行近端迭代, 在实际中可能会出现震荡或者缓慢收敛的现象。为此引入 Nesterov 动量加速机制。

对于 Nesterov 动量加速, 引入外推提升收敛速度:

$$\begin{cases} \tilde{U}^k = U^k + \omega_u^k (U^k - U^{k-1}) \\ \tilde{V}^k = V^k + \omega_v^k (V^k - V^{k-1}) \end{cases} \quad (12)$$

$$\text{其中: } \omega_u^k = \frac{t_k - 1}{t_k + 1}, \quad t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}.$$

然后使用 $\tilde{U}^k, \tilde{V}^k$ 来执行每一步的近端更新操作, 该加速策略在凸情形下达到 $O(1/k^2)$ 的收敛率, 在非凸问题中可明显提升收敛速度。

3.2 幂策略压缩奇异值空间维度

为了进一步降低因子维度 d , 避免对高秩成分建模冗余, 引入幂方法 (power method) 对观测矩阵构造了主奇异子空间的近似基, 在更新迭代之前对因子矩阵进行初始化。

幂方法是一种经典的低秩子空间提取算法, 基本思想就是利用对称矩阵的幂次迭代放大主奇异方向, 从而逐步抑制次要方向。算法 1 中给出了使用幂策略加速方法的伪代码。

算法 1: 幂策略的加速方法

输入: 观测矩阵 $M \in \mathbb{R}^{m \times n}$, 估计秩 K , 随机矩阵 $R \in \mathbb{R}^{n \times k}$

输出: 因子矩阵 U_0, V_0

1. $Y = MR$
2. $Q = QR(Y)$ ▷ QR 分解
3. for $i = 1:j$
4. $Y = M(M^T Q)$;
5. $Q = QR(Y)$ ▷ QR 分解
6. end for
7. 输出 Q
8. $[U, S, V] = \text{svd}(Q^T M)$;
9. $U_0 = QU S^{\frac{1}{2}}, V_0 = VS^{\frac{1}{2}}$; ▷ 得到初始化因子矩阵

4 实验

在本节中, 通过具体的实验展示 GUIG-MCP 与其他典型的竞争方法在彩色图像恢复任务中的恢复效率与性能。

4.1 实验参数设置

数据集: 使用 BSDS (berkeley segmentation dataset)

作为数据库。实验采用的图片大小为 $321 \times 481 \times 3$ 或 $481 \times 321 \times 3$ 。对每个颜色通道独立建模与恢复。为了模拟信息缺失, 随机丢弃原图像中 55% 的像素点, 并保持掩膜在所有的对比算法中一致, 使用 PSNR 值的大小来衡量图像恢复的精度。使用的竞争方法包括传统 GUIG_L0, MC_LogDet 以及本文的 GUIG-MCP 方法。

图 1 从左至右为四张图片, 每一列从上到下依次为: 原图、缺失图像、GUIG_L0、MC_LogDet、GUIG-MCP。

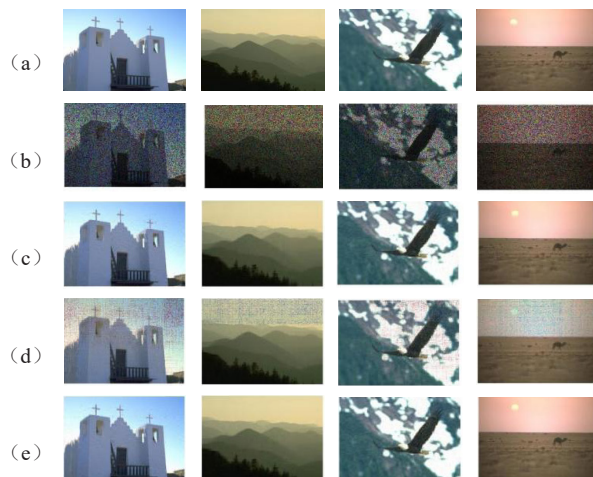


图 1 对比结果

4.2 实验结果及分析

所有竞争方法的恢复结果如图 1 的 (e) 行所示, 各个恢复图像的平均 CPU 时间如表 1 所示。可以看出, 本文方法在大多数情况下恢复的速率更快。

表 1 图 1 中各个恢复的图像 CPU 所用的时间 (单位: s)

图像	1	2	3	4
GUIG_L0	3.93	3.96	5.11	3.96
MC_LogDet	4.73	3.37	4.72	4.30
GUIG-MCP	2.48	3.24	3.23	3.24

在实验过程中, 随机抽取了一幅图展示出本文设计的基于加速近端交替线性最小化的迭代流程在快速收敛上所呈现的效果, 如图 2 所示, 可以明显看出 GUIG-MCP 能够快速地收敛到一个较好的数值。

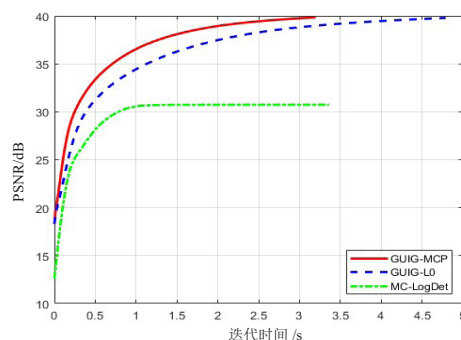


图 2 不同算法 PSNR 随迭代时间的变化曲线

所提出的方法不仅在效率上得到了提升,在最终的图像恢复精度上也有些许的提升,在表2中可以看出图1中图像恢复的PSNR值,GUIG-MCP的快速低秩矩阵补全的方法对缺失的彩色图像能够恢复到一个较好的效果。

表2 图1中各个图像恢复的PSNR(dB)

图像	1	2	3	4
GUIG_L0	34.10	38.66	32.60	41.13
MC_LogDet	18.30	24.50	20.92	25.7
GUIG-MCP	34.84	39.10	32.73	41.89

5 结论

从BSDS数据集随机选择彩色图像作为测试集对传统的GUIG_L0矩阵补全方法与MC_LogDet矩阵补全方法进行实验分析比较。实验数据表明,所提GUIG-MCP方法,同时结合了Nesterov动量和幂方法的一种加速近端交替线性最小化方法提升了整体的计算效率。

参考文献:

[1]Wu Di, Shang Mingsheng, Luo Xin, et al. An L_1 -and- L_2 -norm-oriented latent factor model for recommender systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(10): 5775-5788.
[2]王巍凤,岳亚辉,冯越,等.基于四元数鲁棒主成分分析的彩色视频背景建模[J].信息技术与信息化,2025(4):80-83.
[3]Jia Xixi, Feng Xiangchu, WangWeiwei, et al.Generalized

unitarily invariant gauge regularization for fast low-rank matrix recovery[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021,32(4):1627-1641.
[4]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.
[5]张弦.几类低秩稀疏矩阵优化问题非凸正则化理论与算法研究[D].贵阳:贵州大学,2024.
[6]Zhang Hengmin, Wen Bihan, Zha Zhiyuan, et al. Accelerated PALM for nonconvex low-rank matrix recovery with theoretical analysis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(4): 2304-2317.

【作者简介】

李慧勇(1999—),男,河南商丘人,硕士研究生,研究方向:高维数据与图像处理。

马鑫(1996—),男,河南郑州人,硕士研究生,研究方向:基于广义矩阵的图像分类算法。

邱枫(2001—),男,浙江金华人,硕士研究生,研究方向:轴承寿命预测。

顾强(2000—),男,江苏扬州人,硕士研究生,研究方向:图像处理与模式识别。

廖亮(1975—),通信作者(email:liaoliangis@126.com),男,河南郑州人,博士,教授、研究生导师,研究方向:高维数据分析与图像处理通信。

(收稿日期:2025-11-03 修回日期:2026-06-01)

(上接第120页)

参考文献:

[1] Zhang Yahui, Zhou Taotao, Huang Xufeng, et al. Fault diagnosis of rotating machinery based on recurrent neural networks[J]. Measurement, 2023, 171: 108774.
[2] Pandhare V, Singh J, Lee J. Convolutional neural network based rolling-element bearing fault diagnosis for naturally occurring and progressing defects using time-frequency domain features[C]//2019 Prognostics and System Health Management Conference (PHM-Paris).Piscataway: IEEE, 2023: 320-326.
[3] Shao Siyu, McAleer S, Yan Ruqiang, et al. Highly accurate machine fault diagnosis using deep transfer learning[J]. IEEE Transactions on Industrial Informatics, 2022, 15(4): 2446-2455.
[4] Wang Zhijian, Zheng Likang, Du Wenhua, et al. A novel method for intelligent fault diagnosis of bearing based on capsule neural network[J]. Complexity, 2023: 6943234.
[5] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[6] Chen Tianqi, Guestrin C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.
[7] Mirza M, Osindero S. Conditional generative adversarial nets[PP/OL].(2014-11-06)[2026-05-29].https://doi.org/10.48550/arXiv.1411.1784.
[8] Li Xiang, Zhang Wei, Ding Qian. Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction[J]. Reliability Engineering & System Safety, 2023, 182: 208-218.

【作者简介】

朱美莲(1978—),女,江苏宿迁人,本科,高级工程师、调控值长,研究方向:配网调控运行。

高净净(1996—),女,山西大同人,本科,工程师,研究方向:通信运检及调度自动化。

(收稿日期:2025-10-27 修回日期:2026-06-11)