低秩稀疏矩阵优化问题的模型与算法

潘少华* 文再文†

2020 年 3 月 21 日

摘要

低秩稀疏矩阵优化问题是一类带有组合性质的非凸非光滑优化问题. 由于零模与秩函数的重要性和特殊性, 这类 NP-难矩阵优化问题的模型与算法研究在过去十几年里取得了长足发展. 本文从稀疏矩阵优化问题、低秩矩阵优化问题、低秩加稀疏矩阵优化问题、以及低秩张量优化问题四个方面来综述其研究现状; 其中, 对稀疏矩阵优化问题, 主要以稀疏逆协方差矩阵估计和列稀疏矩阵优化问题为典例进行概述, 而对低秩矩阵优化问题, 主要从凸松弛和因子分解法两个角度来概述秩约束优化和秩 (正则) 极小化问题的模型与算法研究. 最后，总结了低秩稀疏矩阵优化研究中的一些关键与挑战问题, 并提出了一些可以探讨的问题.

关键词: 低秩稀疏矩阵优化; 凸松弛模型; 因子分解模型; 精确恢复条件; 收敛性

1 引言

低秩稀疏矩阵优化问题是应大数据时代需求而产生的一类新型优化问题, 其主旨是寻求满足一定约束条件且使某损失函数值尽可能小的低秩和/或稀疏矩阵. 矩阵的秩与稀疏度不仅是数据线性关系及可解释性的基本度量, 还是某些代数与几何关系的表达工具 [27,73]. 不仅如此, 在许多科学工程领域中, 矩阵的秩还是某些模型或设计的阶数、复杂度、以及维数的度量 [31]; 例如, 低秩矩阵可以对应某系统的低阶控制器、某拟合随机过程的低阶统计模型、某种可嵌入到低维空间的形状、或者某种含少量元件的设计. 鉴于矩阵秩函数与零模的重要性以及大数据中普遍呈现的低秩和/或稀疏性, 这类优化问题在统计 [75]、控制与系统辨识 [32]、信号与图像处理 [46]、机器学习 [92]、量子计算 [44]、金融 [84]、计算化学 [96] 等领域中具有十分广泛的应用也就不足为奇了.

从拟寻求矩阵 (简称目标矩阵) 的特点来看, 低秩稀疏矩阵优化本质是双目标规划, 即目标矩阵不仅要有尽可能小的秩和/或很高的稀疏度, 还要使损失函数值尽可能的小; 从零模与秩函数的变分刻画来看, 低秩稀疏矩阵优化问题本质是带有矩阵平衡约束的数学规划 (简称矩阵 MPEC); 而从矩阵的分解角度看, 低秩矩阵优化问题又是非凸双线性因子矩阵规划. 由于实际应用中的低

*华南理工大学数学学院 (shhpan@scut.edu.cn), 该工作得到国家自然科学基金 (项目号 11971177) 资助.

†北京大学北京国际数学研究中心 (wenzw@pku.edu.cn), 该工作得到国家自然科学基金 (项目号 11831002) 和北京智源人工智能研究院资助.

秩稀疏矩阵优化问题通常具有较大规模, 采用全局优化的策略来寻求其全局最优解的思路是不可行的. 这样, 如何根据这类优化问题的特点设计既有理论保证又能快速求解的算法成为优化领域面临的新挑战.

在过去十几年里, 基于零模与秩函数的特殊性质, 提出了许多处理低秩稀疏矩阵优化问题的模型与算法. 本文第 2 节为预备知识, 主要介绍了零模与秩函数的一些非光滑性质和在低秩稀疏矩阵优化理论分析中常用的正则性质; 第 3 节以稀疏逆协方差矩阵估计和列稀疏矩阵优化为典例来概述稀疏矩阵优化的模型与算法研究; 第 4 节从凸松弛和因子分解两个角度概述了秩约束优化和秩 (正则) 极小化问题的模型与算法研究; 第 5 节概述了低秩加稀疏矩阵优化问题的模型与算法的研究; 第 6 节主要介绍与低秩矩阵优化相关的低秩张量问题的模型与算法研究; 第 7 节总结了低秩稀疏矩阵优化研究中的关键问题与挑战; 最后在第 8 节中提出低秩稀疏矩阵优化中可以探讨的一些问题.

2 符号说明与预备知识

符号说明: 本文用 $\mathbb{R}^{n\times m}$ 表示所有 $n\times m$ 且 $n\le m$ 实矩阵构成且赋予迹内积 $\langle\cdot,\cdot\rangle$ 及其诱导的 Frobenius 范数 $\|\cdot\|_F$ 的向量空间, $\mathbb{S}^n$ 表示所有 $n\times n$ 实对称矩阵构成的集合, $\mathbb{S}^n_{++}$ 表示 $\mathbb{S}^n$ 中所有正定矩阵构成的开凸锥, 而 $\mathbb{O}^n$ 表示所有 $n\times n$ 正交矩阵构成的集合. 符号 I 和 e 分别表示单位矩阵和全 1 向量, 其维数可从文中推知. 对任意 $X\in\mathbb{R}^{n\times m}$, $\|X\|_*$ 和 $\|X\|$ 分别表示矩阵 X 的核范数与谱范数, $\|X\|_{2,0}=\sum_{j=1}^m \mathrm{sign}(\|X_{\cdot j}\|)$ 表示 X 的列 $\ell_{2,0}$-范数, 其中 $X_{\cdot j}$ 表示矩阵 X 的第 j 列, 而 $\sigma(X)=(\sigma_1(X),\ldots,\sigma_n(X))^{\mathbb{T}}$ 表示元素按从大到小排序的奇异值向量, 并记 $\mathbb{O}^{n,m}(X):=\big\{(U,V)\in\mathbb{O}^n\times\mathbb{O}^m \mid X=U[\mathrm{Diag}(\sigma(X))\ \ 0]V^{\mathbb{T}}\big\}$. 对给定的闭集 Λ, δ_Λ 表示集合 Λ 的指示函数, 即 $\delta_\Lambda(x)=0$ 若 $x\in\Lambda$, 否则 $\delta_\Lambda(x)=+\infty$; $\mathcal{P}_\Lambda(\cdot)$ 表示集合 Λ 上的投影算子; 而 $\mathcal{N}_\Lambda(x)$ 表示集合 Λ 在 $x\in\Lambda$ 点处的 (极限) 法锥. 对任意给定的向量, $[\![x]\!]$ 与 $[\![x]\!]^\perp$ 分别表示由 x 生成的子空间及其正交补空间. 对给定函数 f, 用 $\mathcal{P}_{\tau,f}$ 表示 f 的参数 $\tau>0$ 的邻近映射, 即

$$\mathcal{P}_{\tau,f}(x):=\mathop{\arg\min}_{z\in\mathbb{X}}\left\{\frac{1}{2\tau}\|z-x\|^2+f(z)\right\}.$$

2.1 零模与秩函数的非光滑性质

零模与秩函数不仅是具有一定组合性的下半连续函数, 还是带平衡约束数学规划的值函数. 设 $\mathscr{L}$ 是由满足如下条件的闭正常凸函数 $\phi\colon\mathbb{R}\to(-\infty,+\infty]$ 构成的函数簇:

$$\mathrm{int}(\mathrm{dom}\,\phi)\supseteq[0,1],\ 1>t^*:=\mathop{\arg\min}_{0\le t\le 1}\phi(t),\ \phi(t^*)=0\ \text{ 且 }\ \phi(1)=1. \tag{1}$$

由于 $\phi(1)=1$ 且 t^* 是函数 ϕ 在闭区间 $[0,1]$ 上的唯一最小值点, 故对任意 $x\in\mathbb{R}^n$ 有

$$\|x\|_0=\min_{w\in\mathbb{R}^n}\left\{\sum_{i=1}^n\phi(w_i)\quad \text{s.t.}\ \ \langle e-w,|x|\rangle=0,\,0\le w\le e\right\}. \tag{2}$$

这表明零模是带有互补约束 $e-w\geq 0,|x|\geq 0,\langle e-w,|x|\rangle=0$ 数学规划的值函数. 据我们所知, Mangasarian [72] 是最早针对非负向量利用 $\phi(t)=t$ 给出这种变分刻画的. 任取 $\phi\in\mathscr{L}$, 应用 Von Neumann 迹不等式 [49, 式 (3.3.25)] 可以验证, 对任意 $X\in\mathbb{R}^{n\times m}$,

$$\mathrm{rank}(X)=\min_{W\in\mathbb{R}^{n\times m}}\Big\{\sum_{i=1}^{n}\phi(\sigma_i(W))\ \ \text{s.t.}\ \ \|X\|_*-\langle X,W\rangle=0,W\in\mathbb{B}\Big\}\tag{3}$$

其中 $\mathbb{B}:=\{Z\in\mathbb{R}^{n\times m}\mid \|Z\|\leq 1\}$ 表示谱范数单位球. 由于 $\|X\|_*-\langle X,W\rangle=0,W\in\mathbb{B}$ 成立当且仅当 $W\in\arg\max_{Z\in\mathbb{B}}\langle Z,X\rangle$ 或 $X\in\mathcal{N}_{\mathbb{B}}(W)$, 故 (3) 式表明秩函数是带有平衡约束 $X\in\mathcal{N}_{\mathbb{B}}(W)$ 数学规划的值函数. 此外, 从矩阵因子分解角度看, 秩函数还是参变量双线性等式约束规划的值函数. 的确, 对任意秩不超过 κ 的矩阵 $X\in\mathbb{R}^{n\times m}$, 不难验证

$$\mathrm{rank}(X)=\min_{U\in\mathbb{R}^{n\times\kappa},V\in\mathbb{R}^{m\times\kappa}}\Big\{\frac{1}{2}\big(\|U\|_{2,0}+\|V\|_{2,0}\big)\ \ \text{s.t.}\ \ UV^{\mathbb{T}}=X\Big\}.\tag{4}$$

下面来回顾零模与秩函数的广义次微分, 它们是刻画低秩稀疏矩阵优化各类稳定点的关键. 为此, 设 $\mathbb{X}$ 为有限维向量空间 $\mathbb{R}^n$ 或 $\mathbb{R}^{n\times m}$, 对广义实值函数 $f\colon\mathbb{X}\to[-\infty,+\infty]$, 分别用 $\widetilde{\partial}f$, $\widehat{\partial}f,\partial f$ 及 $\overline{\partial}f$ 表示 f 的邻近、正则、极限和 Clarke 次微分映射; 当 f 为某闭集 $\Lambda\subseteq\mathbb{X}$ 上的指示函数时, 它们分别对应 Λ 的邻近、正则、极限和 Clarke 法锥映射.

2.1.1 零模与零模约束集的非光滑性质

文献 [61, 定理 1] 阐述零模在任意点处的邻近、正则、极限和 Clarke 次微分相同. 此结果可直接由零模的邻近正则性 (prox-regularity) 得到, 下面来证明零模的此性质.

命题 2.1 *零模是邻近正则函数, 从而对任意 $x\in\mathbb{R}^n$ 下述几个等式成立:*

$$\widetilde{\partial}\|\cdot\|_0(x)=\widehat{\partial}\|\cdot\|_0(x)=\partial\|\cdot\|_0(x)=\overline{\partial}\|\cdot\|_0(x)=[\![x]\!]^{\perp}.$$

证明: 对任意 $z\in\mathbb{R}^n$, 记 $\mathrm{supp}(z)=\{i\in\{1,\ldots,n\}\mid z_i\neq 0\}$. 固定任意 $\overline{x}\in\mathbb{R}^n$ 以及任意 $\overline{v}\in\partial\|\cdot\|_0(\overline{x})$. 根据连续性, 必存在 $\varepsilon_1>0$ 使得对所有 $z\in\mathbb{B}(\overline{x},\varepsilon_1)$, $\mathrm{supp}(z)\supseteq\mathrm{supp}(\overline{x})$. 令 $\varepsilon=\frac{1}{4}\min(\varepsilon_1,\frac{1}{\|\overline{v}\|+1})$. 任取 $x'\in\mathbb{B}(\overline{x},\varepsilon)$. 考虑 $\|x-\overline{x}\|<\varepsilon$ 且 $\|x\|_0<\|\overline{x}\|_0+\varepsilon$ 以及 $v\in\partial\|\cdot\|_0(x)$ 且 $\|v-\overline{v}\|<\varepsilon$. 注意到 $\mathrm{supp}(x)\supseteq\mathrm{supp}(\overline{x})$. 结合 $\|x\|_0<\|\overline{x}\|_0+\varepsilon$, 容易推知 $\mathrm{supp}(x)=\mathrm{supp}(\overline{x})$. 注意到 $\mathrm{supp}(x')\supseteq\mathrm{supp}(\overline{x})$, 下面分两种情况进行讨论:

情形 1: $\mathrm{supp}(x')\neq\mathrm{supp}(\overline{x})$. 此时, $\|x'\|_0-\|x\|_0\geq 1$, 从而对任意 $\rho>0$ 都成立着

$$\|x'\|_0\geq\|x\|_0+\langle v,x'-x\rangle-(\rho/2)\|x'-x\|^2.\tag{5}$$

情形 2: $\mathrm{supp}(x')=\mathrm{supp}(\overline{x})$. 此时, 结合 $\mathrm{supp}(x)=\mathrm{supp}(\overline{x})$ 以及 $\partial\|\cdot\|_0(x)=[\![x]\!]^{\perp}$, 容易得到 $\|x'\|_0=\|\overline{x}\|_0=\|x\|_0$ 和 $\langle v,x'-x\rangle=0$, 从而不等式 (5) 对任意的 $\rho>0$ 也成立.

由 [91, 定义 13.27] 及 $\overline{x}\in\mathbb{R}^n$ 和 $\overline{v}\in\partial\|\cdot\|_0(\overline{x})$ 的任意性, 零模是邻近正则函数. □

下述引理提供了零模约束集的邻近、正则和极限法锥的刻画, 其证明详见文献 [7].

引理 2.1 *设 r 是给定的正整数, 定义集合 $\Lambda := \{z \in \mathbb{R}^n \mid \|z\|_0 \le r\}$. 考虑任意 $x \in \Lambda$.*

(i) 若 $\|x\|_0 = r$, 则 $\widetilde{\mathcal{N}}_\Lambda(x) = \widehat{\mathcal{N}}_\Lambda(x) = [\![x]\!]^\perp = \mathcal{N}_\Lambda(x)$.

(ii) 若 $\|x\|_0 < r$, 则 $\widetilde{\mathcal{N}}_\Lambda(x) = \{0\} = \widehat{\mathcal{N}}_\Lambda(x) \subseteq \mathcal{N}_\Lambda(x) = \big\{v \in \mathbb{R}^n \mid \|v\|_0 \le n-r\big\} \cap [\![x]\!]^\perp$.

2.1.2 秩函数与秩约束集的非光滑性质

由于秩函数是零模与奇异值向量的复合, 而零模是绝对对称函数 (即对任意 $x \in \mathbb{R}^n$ 以及任意 $n \times n$ 符号置换矩阵 P 都有 $\|Px\|_0 = \|x\|_0$), 将命题 2.1 与文献 [28, 定理 4.2] 结合即可得知秩函数是邻近正则的. 再结合文献 [63, 64] 中的结论, 便有下述命题成立.

命题 2.2 *秩函数是邻近正则函数, 从而对任意 $X \in \mathbb{R}^{n\times m}$ 成立着*

$$\begin{aligned}\widetilde{\partial}\mathrm{rank}(X) &= \widehat{\partial}\mathrm{rank}(X) = \partial\mathrm{rank}(X) = \overline{\partial}\mathrm{rank}(X) = [\![X]\!]^\perp \\ &= \big\{U[\mathrm{Diag}(w)\ \ 0]V^{\mathbb{T}} \mid w \in \partial\|\cdot\|_0(\sigma(X)), (U,V) \in \mathbb{O}^{n,m}(X)\big\}.\end{aligned}$$

注意到, 秩约束集的指示函数是零模约束集的指示函数与奇异值向量的复合, 而零模约束集的指示函数是绝对对称的. 将引理 2.1 与文献 [63, 64] 结合, 即可得到下述结论.

命题 2.3 *设 r 是给定的正整数, 定义集合 $\Xi := \{Z \in \mathbb{R}^{n\times m} \mid \mathrm{rank}(Z) \le r\}$. 考虑任意 $X \in \Xi$. 若 $\mathrm{rank}(X) = r$, 则有 $\widetilde{\mathcal{N}}_\Xi(X) = \widehat{\mathcal{N}}_\Xi(X) = \mathcal{N}_\Xi(X) = [\![X]\!]^\perp$; 若 $\mathrm{rank}(X) < r$, 则有 $\widetilde{\mathcal{N}}_\Xi(X) = \{0\} = \widehat{\mathcal{N}}_\Xi(X) \subseteq \mathcal{N}_\Xi(X) = \big\{W \in \mathbb{R}^{n\times m} \mid \mathrm{rank}(W) \le n-r\big\} \cap [\![X]\!]^\perp$.*

由命题 2.3 知, 秩约束集在边界点 M 处的法锥恰好是法空间 $[\![M]\!]^\perp$, 从而它在点 M 处的切锥是子空间 $[\![M]\!]$. 这样, 任取 $(U,V) \in \mathbb{O}^{n,m}(M)$ 并记 $U = [U_1\ \ U_2], V = [V_1\ \ V_2]$ 其中 $U_1 \in \mathbb{R}^{n\times r}$ 与 $V_1 \in \mathbb{R}^{m\times r}$ 是分别表示由 U 和 V 的前 r 列构成的子矩阵, 便有

$$[\![M]\!] = \big\{H \mid H = U_1U_1^{\mathbb{T}}HV_2V_2^{\mathbb{T}} + HV_1V_1^{\mathbb{T}}\big\},\ [\![M]\!]^\perp = \big\{H \mid H = U_2U_2^{\mathbb{T}}HV_2V_2^{\mathbb{T}}\big\}.$$

根据零模与秩函数的广义次微分, 容易验证零模与秩函数还是指数 1/2 的 KL 函数 (关于 KL 的定义见定义 2.5). 另外, 零模与秩函数的邻近映射 (尽管多值) 还有闭式表达, 是低秩稀疏矩阵优化问题的硬阈值算法的核心. 的确, 对任意 $x \in \mathbb{R}^n$ 及 $\tau > 0$,

$$\mathcal{P}_{\tau,\|\cdot\|_0}(x) = \left\{x^* \in \mathbb{R}^n \mid x_i^* = \begin{cases} x_i & \text{if } x_i \ge \sqrt{2\tau}; \\ 0 & \text{otherwise,}\end{cases}\quad i = 1,\ldots,n\right\},$$

而对任意矩阵 $X \in \mathbb{R}^{n\times m}$, 根据 Von Neumann 迹不等式以及秩函数与零模的关系,

$$\mathcal{P}_{\tau,\mathrm{rank}(\cdot)}(X) = \Big\{U[\mathrm{Diag}(x^*)\ \ 0]V^{\mathbb{T}} \mid x^* \in \mathcal{P}_{\tau,\|\cdot\|_0}(\sigma(X)), (U,V) \in \mathbb{O}^{n,m}(X)\Big\}.$$

值得指出的是, 在低秩稀疏矩阵优化问题的算法分析中, 更多需要的是零模 (或秩) 正则复合函数的广义次微分及相应指数 1/2 的 KL 性质刻画; 譬如, 根据零模与秩函数的邻近映射具有闭式表达, 对某些特殊低秩稀疏矩阵优化问题可直接藉助零模与秩函数的邻近映射设计算法, 但要分析这些算法的线性收敛速率, 则需刻画零模或秩函数正则复合函数的指数 1/2 的 KL 性质. 另外, 虽然单独的零模约束集或秩约束集的法锥映射比较容易刻画, 但零模 (或秩) 约束与其他约束条件构成的复合闭集的法锥刻画并不容易, 需要一定的约束规范, 而且, 因零模与秩函数的组合性, Robinson 约束规范条件通常是不成立的. 这样, 识别出满足更弱约束规范条件的零模 (或秩) 约束复合集类尤为重要.

2.2 几类常用的正则性质

本小节回顾低秩稀疏矩阵优化的理论分析中常用的几类正则性质, 包括损失函数的限制强凸与限制光滑性、线性算子的限制同构性、以及非凸非光滑函数的 KL 性质.

2.2.1 损失函数的限制强凸与光滑性

首先, 我们来介绍连续可微函数的限制秩型的强凸 (RSC) 与光滑性 (RSS) 定义.

定义 2.1 *(见 [55])* 设 $\Phi\colon \mathbb{R}^{n\times m}\to\mathbb{R}$ 是给定的连续可微函数. 若存在常数 $\alpha>0$ 使得对任意秩不超过 r 的矩阵 $X,Z\in\mathbb{R}^{n\times m}$ 都有下述不等式成立

$$\Phi(X)-\Phi(Z)-\langle\nabla\Phi(Z),X-Z\rangle\geq(\alpha/2)\|X-Z\|_F^2,$$

则称 Φ 满足 (r,α)-*RSC*; 若存在常数 $\beta>0$ 使得对任意秩不超过 r 的 $X,Z\in\mathbb{R}^{n\times m}$ 都有

$$\Phi(X)-\Phi(Z)-\langle\nabla\Phi(Z),X-Z\rangle\leq(\beta/2)\|X-Z\|_F^2,$$

则称 Φ 满足 (r,β)-*RSS*; 并称这样的最大 α 和最小 β 分别为 r-*RSC* 模和 r-*RSS* 模.

当将定义 2.1 中的 X,Z 限制为对角矩阵时, 即可得到稀疏条件下的限制强凸与限制光滑性 (见 [93]). 上述限制强凸与光滑性是通过限制变量的秩 (或稀疏度) 给出的, 主要用于矩阵零模约束和秩约束优化问题的理论分析 (如硬阈值迭代法和因子分解法), 而在低秩稀疏矩阵优化的凸松弛法分析中, 通常采用限制方向集型的强凸与光滑性.

定义 2.2 *(见 [1])* 设 $\Phi\colon\mathbb{R}^{n\times m}\to\mathbb{R}$ 与 $\vartheta\colon\mathbb{R}^{n\times m}\to\mathbb{R}_+$ 分别是给定的连续可微函数与正则函数. 若存在常数 $\gamma_l>0,\tau_l\geq0$ 使得对任意 $X,Z\in\mathcal{C}\subseteq\mathbb{R}^{n\times m}$ 都有

$$\Phi(X)-\Phi(Z)-\langle\nabla\Phi(Z),X-Z\rangle\geq(\gamma_l/2)\|X-Z\|_F^2-\tau_l\,\vartheta^2(X-Z),$$

则称 Φ 在 $\mathcal{C}$ 上满足关于正则函数 ϑ 的 (γ_l,τ_l)-*RSC*; 若存在常数 $\gamma_u>0,\tau_u\geq0$ 使得

$$\Phi(X)-\Phi(Z)-\langle\nabla\Phi(Z),X-Z\rangle\leq(\gamma_u/2)\|X-Z\|_F^2+\tau_u\,\vartheta^2(X-Z)\quad\forall X,Z\in\mathcal{C},$$

则称函数 Φ 在集合 $\mathcal{C}$ 上满足关于正则函数 ϑ 的 (γ_u,τ_u)-RSS.

不难看到, 当 Φ 是对应采样算子 $\mathcal{A}\colon\mathbb{R}^{n\times m}\to\mathbb{R}^N$ 的最小二乘损失函数 $\frac{1}{2}\|\mathcal{A}(X)-b\|^2$ 时, 其模 α 的 r-RSC 与模 β 的 r-RSS 分别退化为 $\mathcal{A}^*\mathcal{A}$ 的 $2r$-限制最小与最大特征值, 即

$$\alpha=\min_{\mathrm{rank}(Y)\le 2r,\|Y\|_F=1}\|\mathcal{A}(Y)\|^2,\quad \beta=\max_{\mathrm{rank}(Y)\le 2r,\|Y\|_F=1}\|\mathcal{A}(Y)\|^2;$$

而其在集合 $\mathcal{C}$ 上关于任意正则函数 ϑ 的 $(\gamma_l,0)$-RSC 等价于要求对所有 $H\in\mathcal{C}-\mathcal{C}$, $\|\mathcal{A}(H)\|^2\ge\gamma_l\|H\|_F^2$, 此时也称线性算子 $\mathcal{A}$ 满足集合 $\mathcal{C}-\mathcal{C}$ 上的限制强凸性.

2.2.2 线性算子的限制同构性

定义 2.3 *(见 [87]) 对给定 $r\in\{1,\dots,n\}$, 线性算子 $\mathcal{A}\colon\mathbb{R}^{n\times m}\to\mathbb{R}^N$ 的 r-限制同构常数 $\delta_r(\mathcal{A})$ 是使下述不等式对所有秩不超过 r 的矩阵 $X\in\mathbb{R}^{n\times m}$ 都成立的最小正常数 δ:*

$$(1-\delta)\|X\|_F\le\|\mathcal{A}X\|\le(1+\delta)\|X\|_F;$$

若存在某个常数 $\overline{\delta}\in[0,1)$ 使得 $\delta_r(\mathcal{A})\le\overline{\delta}$, 则称线性算子 $\mathcal{A}$ 满足秩 r-限制同构性.

对照定义 2.1, 若线性算子 $\mathcal{A}$ 满足秩 $2r$-限制同构性, 则相应最小二乘损失函数满足模 $[1-\delta_{2r}(\mathcal{A})]^2$ 的 r-RSC 与模 $[1+\delta_{2r}(\mathcal{A})]^2$ 的 r-RSS. 显然, $\mathcal{A}$ 的 $2r$-限制同构常数 $\delta_{2r}(\mathcal{A})$ 越接近 0 等价于最小二乘损失函数的 r-限制条件数 $\frac{[1+\delta_{2r}(\mathcal{A})]^2}{[1-\delta_{2r}(\mathcal{A})]^2}$ 越接近 1. 文献 [87] 中证明 "对任意给定的 $\delta\in(0,1)$, 若 $\mathcal{A}$ 是几乎同构随机线性映射, 则当 $N\ge c_0r(n+m)\log(nm)$, $\delta_r(\mathcal{A})\le\delta$ 会以概率 $1-\exp(-c_1N)$ 成立, 其中 c_0 和 c_1 是仅依赖于 δ 的正常数." 虽然文献 [87] 提供了一些几乎同构随机线性映射的例子 (如 $\mathcal{A}$ 的相应矩阵具有独立同分布元素且每个元素服从高斯分布或对称伯努利分布), 但限制同构性还是比较强的. 文献 [75] 还通过例子说明即使映射 $\mathcal{A}$ 不满足限制同构性, 但会以高概率满足闭凸锥 $\mathcal{C}:=\{H\mid\|\mathcal{P}_{[\![M]\!]^\perp}(H)\|_*\le 3\|\mathcal{P}_{[\![M]\!]}(H)\|_*\}$ 上的限制强凸性, 其中 $M\in\mathbb{R}^{n\times m}$ 是真实矩阵.

正如文献 [17] 中提到的, 矩阵填充问题的采样算子一般不满足限制同构性. 对这类低秩矩阵优化问题的理论分析, 通常需假设真实矩阵满足不相关 (incoherence) 条件, 其本质是要求真实矩阵的奇异值向量元素尽可能均匀, 见文献 [23] 给出的下述定义:

定义 2.4 *设秩 r 矩阵 $M\in\mathbb{R}^{n\times m}$ 有奇异值分解 $U[\mathrm{Diag}(\sigma(M))\ \ 0]V^{\mathbb{T}}$. 若对某个 $\mu_0>0$,*

$$\max_{1\le i\le n}\|U^{\mathbb{T}}e_i\|\le\sqrt{\mu_0r/n}\ \ 且\ \ \max_{1\le j\le m}\|V^{\mathbb{T}}e_j\|\le\sqrt{\mu_0r/m},$$

则称矩阵 M 满足参数 μ_0 的不相关条件, 其中 e_i 是第 i 个标准基向量.

2.2.3 非凸非光滑函数的 KL 性质

对低秩稀疏矩阵优化问题, 为了分析基于近似或等价非凸代理模型设计的算法的全局收敛性及收敛速率, 通常需要代理模型目标函数的 KL 性质, 这里给出其严格定义.

定义 2.5 *设 $h:\mathbb{X}\to[-\infty,+\infty]$ 是闭正常广义实值函数, 考虑任意点 $\overline{x}\in\operatorname{dom}\partial h$. 如果存在点 $\overline{x}$ 的领域 $\mathcal{U}$, 常数 $\eta\in(0,+\infty]$ 以及满足如下性质的连续凹函数 $\varphi:[0,\eta)\to\mathbb{R}_+$:*

(i) $\varphi(0)=0$ 且 φ 在开区间 $(0,\eta)$ 上连续可微;

(ii) 对所有 $s\in(0,\eta)$ 都有 $\varphi'(s)>0$

使得对所有 $x\in\mathcal{U}\cap\big[h(\overline{x})<h(x)<h(\overline{x})+\eta\big]$, 都有 $\varphi'(h(x)-h(\overline{x}))\mathrm{dist}(0,\partial h(x))\geq 1$, 则称函数 h 在点 $\overline{x}$ 处具有 KL 性质; 特别地, 当 $\varphi(s)\equiv c\sqrt{s}$ (其中 $c>0$ 为常数) 时, 此不等式仍成立, 则称函数 h 在点 $\overline{x}$ 处具有指数 1/2 的 KL 性质. 如果 h 在集合 $\operatorname{dom}\partial h$ 中的每点处都满足 (指数 1/2 的) KL 性质, 则称 h 是 (指数 1/2 的) KL 函数.

正如文献 [3] 中第 4 节所讨论的, 许多类函数都具有 KL 性质 (如半代数函数), 但要识别它们 (尤其是复合非凸非光滑函数) 是否具有指数 1/2 的 KL 性质并不是容易的事, 感兴趣的读者可参阅文献 [66] 中提供的一些准则.

3 稀疏矩阵优化问题的模型与算法

稀疏矩阵优化问题旨在寻求使某损失函数值尽可能小且有很多零元或零列 (行) 的矩阵, 其典例是稀疏逆协方差矩阵估计问题和列 (行) 稀疏矩阵优化问题.

3.1 稀疏逆协方差矩阵估计的模型与算法

稀疏逆协方差矩阵估计问题就是基于随机向量 $\mathbf{X}\in\mathbb{R}^p$ 的独立同分布观测 $\{\mathbf{X}^{(k)}\}_{k=1}^p$ 来估计其稀疏逆协方差矩阵. 由于当随机向量服从高斯分布时, 其协方差矩阵逆的稀疏结构与高斯马尔可夫随机域图的边结构对应, 该问题在机器学习、信号处理、计算生物等领域中有着十分重要的应用 [34,106]. 设随机向量 $\mathbf{X}\in\mathbb{R}^p$ 服从高斯分布 $\mathcal{N}(0,\Sigma^*)$, 其稀疏逆协方差矩阵 $\Theta^*=(\Sigma^*)^{-1}$ 估计的理想选择是零模正则的最大似然估计:

$$\min_{\Theta\in\mathbb{S}^p_{++}}\Big\{-\log(\det\Theta)+\langle\Sigma^*,\Theta\rangle+\lambda\|\Theta\|_0\Big\}. \tag{6}$$

由于真实协方差矩阵的未知性和矩阵零模的组合性, 问题 (6) 是不可解的. 对于 Σ^* 的未知性, 自然采用样本协方差矩阵 S 代替 Σ^* 来克服; 而对矩阵零模 $\|\Theta\|_0$ 的组合性, 通常用其在无穷范数单位球 $\{Z\in\mathbb{S}^p\mid\max_{i,j}|Z_{ij}|\leq 1\}$ 上的凸包 $\|\Theta\|_1$ (见文献 [31]) 来代替零模克服. Yuan 等人 [6,106] 正是率先通过解这样的凸规划来得到 Θ^* 的稀疏估计:

$$\widehat{\Theta}:=\mathop{\arg\min}_{\Theta\in\mathbb{S}^p_{++}}\Big\{-\log(\det\Theta)+\langle S,\Theta\rangle+\lambda\|\Theta\|_1\Big\}. \tag{7}$$

由于最优解 $\widehat{\Theta}$ 的非零元表征了图模型变量的相关性, 模型 (7) 也常被称为高斯图模型. 另一类常见的模型是根据关系式 $\Sigma^*\Theta^*=I$ 以及 S 是 Σ^* 的无偏估计的事实, 通过在矩阵集 $\{\Theta\in\mathbb{R}^{p\times p}:$

$\|S\Theta - I\|_\infty \le \lambda\}$ 中寻求 ℓ_1-范数最小的矩阵来得到 Θ^* 的稀疏估计 [15], 即将如下凸优化问题最优解的适当对称化 $\widetilde{\Theta}$ 作为 Θ^* 的稀疏估计:

$$\widetilde{\Theta}' \in \mathop{\arg\min}_{\Theta \in \mathbb{R}^{p\times p}} \Big\{ \|\Theta\|_1 \ \ \text{s.t.}\ \|S\Theta - I\|_\infty \le \lambda \Big\}. \tag{8}$$

对于凸松弛问题 (7), Ravikumar 等人 [86] 在真实逆协方差矩阵 Θ^* 的不可表示条件和样本协方差矩阵的适当尾条件下, 不仅导出了唯一最优解 $\widehat{\Theta}$ 到真实解 Θ^* 的元素最大无穷范数统计界, 还建立了最优解 $\widehat{\Theta}$ 的符号相容性; 而对凸松弛问题 (8), Cai 等人 [15] 通过限制真实逆协方差矩阵 Θ^* 的 ℓ_1-范数及近似稀疏度, 在适当的矩条件下建立了 $\widetilde{\Theta}$ 到真实解 Θ^* 的谱范数、Frobenius 范数以及元素最大无穷范数的统计界. 这些结果为矩阵零模正则化问题 (6) 的凸松弛模型 (7) 和 (8) 提供了统计意义下的理论保证.

鉴于稀疏逆协方差矩阵估计的重要性, 文献中提出了许多求解凸规划 (7) 的算法, 如对偶一阶算法和牛顿法、原一阶算法和牛顿法, 这里主要介绍 Banerjee 等人 [6] 提出的对偶块坐标下降法和 Hsieh 等人 [53] 提出的原块坐标牛顿法.

3.1.1 对偶块坐标下降法

根据凸规划的强对偶定理, 凸规划 (7) 的最优解可以通过解其如下对偶问题来得到:

$$\min_{\Gamma \in \mathbb{S}^p_{++}} \Big\{ -\log(\det \Gamma) \quad \text{s.t.}\ \ \|\Gamma - S\|_\infty \le \lambda \Big\}.$$

Banerjee 等人 [6] 提出的对偶块坐标下降法是每步迭代通过求解 p 个凸二次规划对变量 Γ 的每行/列进行优化. 给定初始点 $\Gamma^0 = S + \lambda I$, 该方法的基本迭代公式如下:

$$\left\{\begin{array}{l}
\textbf{For } k = 0,1,2,\ldots \\
\qquad \textbf{For } j = 1,2,\ldots,p \\
\qquad\quad (a)\ \ \text{设 } \Gamma^{j-1} \text{是当前迭代, 计算如下二次规划} \\
\qquad\qquad\quad \widehat{y} := \mathop{\arg\min}_{y \in \mathbb{R}^{p-1}} \Big\{ \langle y, (\Gamma^{j-1}_{\backslash j \backslash j})^{-1} y \rangle \quad \text{s.t.}\ \ \|y - S_j\|_\infty \le \lambda \Big\}; \qquad (9a) \\
\qquad\quad (b)\ \ \Gamma^j \longleftarrow \text{用 } \widehat{y} \text{ 代替除去对角元的矩阵 } \Gamma^{j-1} \text{的第 } j \text{ 列.} \qquad (9b) \\
\qquad \textbf{End} \\
\qquad \text{置 } \Gamma^{k+1} \longleftarrow \Gamma^p. \text{ 若 } \Gamma^{k+1} \text{不满足终止条件, 则 } \Gamma^0 \longleftarrow \Gamma^{k+1}, k \longleftarrow k+1. \\
\textbf{End}
\end{array}\right.$$

其中, $\Gamma^{j-1}_{\backslash j \backslash j}$ 是将 Γ^{j-1} 去掉第 j 行和第 j 列而得到的矩阵, 而 S_j 表示矩阵 S 除去对角元的第 j 列. 注意到式 (9a) 中的凸二次规划恰好是标准 Lasso 问题的对偶, 故该方法实质是将复杂的凸规划 (7) 分解为一系列简单 Lasso 问题来求解, 因此也常被称为图 Lasso 算法. 值得强调的是, 上述块坐标下降法是通过估计协方差矩阵来得到其逆的稀疏估计.

3.1.2 原块坐标牛顿法

求解凸规划 (7) 的牛顿法是每步通过解目标函数 f 在当前迭代点处的二次近似来产生牛顿方向, 其中 f 的二次近似是藉助其光滑部分的二次近似和其非光滑部分形成. 记 $g(\Theta) := -\log(\det\Theta) + \langle S,\Theta\rangle$, 则 f 在迭代点 Θ^k 处的牛顿方向由下述优化问题定义

$$D^k := \mathop{\arg\min}_{\Delta\in\mathbb{S}^p}\Big\{\mathrm{vec}(\nabla g(\Theta^k))^{\mathbb{T}}\mathrm{vec}(\Delta) + \frac{1}{2}\mathrm{vec}(\Delta)^{\mathbb{T}}\nabla^2 g(\Theta^k)\mathrm{vec}(\Delta) + \lambda\|\Theta^k+\Delta\|_1\Big\}.$$

藉助 $H^k := \nabla^2 g(\Theta^k)$ 和 $b^k := \mathrm{vec}(\nabla g(\Theta^k))$, 此优化问题可重新写成 Lasso 问题

$$D^k := \mathop{\arg\min}_{\Delta\in\mathbb{S}^p}\Big\{\frac{1}{2}\big\|(H^k)^{1/2}\mathrm{vec}(\Delta) + (H^k)^{-1/2}b^k\big\|^2 + \lambda\|\Theta^k+\Delta\|_1\Big\}. \tag{10}$$

文献 [53] 的块坐标牛顿法是每步使用块坐标下降法求解子问题 (10) 而提出的二阶方法, 由于它将块坐标下降法的优点与问题 (10) 的结构巧妙结合起来, 适于求解大规模问题.

3.2 列 (行) 稀疏矩阵优化的模型与算法

列稀疏矩阵优化问题主要源于高维多元线性回归中的变量选择 [78] 和多任务学习中的联合变量或子空间选择问题 [79, 110], 其目的是恢复一组同时与多个任务相关的协变量. 给定数据样本集 $\big\{(x_i^k, y_i^k)\in\mathbb{R}^p\times\mathbb{R},\, i=1,\dots,N_k, k=1,\dots,K\big\}$, 其中指标 k 和 i 分别表征任务及任务的独立同分布观测. 若 $w^k\in\mathbb{R}^p$ 表示任务 k 的线性评判函数的参数, $J_k(\langle w^k, x_i^k\rangle, y_i^k)$ 表示任务 k 关于样本 (x_i^k, y_i^k) 的损失, 其中 $J_k\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$ 是任务 k 的损失函数, 则选择一组与这 K 个任务同时相关的变量就模型化为列 $\ell_{2,0}$-正则化问题

$$\min_{W\in\mathbb{R}^{K\times p}}\left\{\sum_{k=1}^{K}\sum_{i=1}^{N_k} J_k(\langle w^k, x_i^k\rangle, y_i^k) + \lambda\|W\|_{2,0}\right\} \tag{11}$$

其中 W 是以 $(w^k)^{\mathbb{T}}$ 为行向量的矩阵, 其第 j 列表示协变量 j 对应各任务的系数. 根据向量 ℓ_1-范数是向量零模在 ℓ_∞-范数单位球上的凸包 [31], 自然用 $\|W\|_{2,0}$ 在紧凸集 $\{Z\in\mathbb{R}^{K\times p}\mid \|Z_{\cdot j}\|\le 1,\, j=1,\dots,p\}$ 上的凸包 $\|W\|_{2,1}$ 代替 $\|W\|_{2,0}$, 通过解如下凸规划

$$\min_{W\in\mathbb{R}^{K\times p}}\left\{\sum_{k=1}^{K}\sum_{i=1}^{N_k} J_k(\langle w^k, x_i^k\rangle, y_i^k) + \lambda\|W\|_{2,1}\right\} \tag{12}$$

来得到非凸问题 (11) 的满意解. 针对高维多元线性回归中的变量选择问题, 文献 [78] 在随机高斯设计矩阵的协方差满足适当不可表示条件、自不相关性和有界特征谱条件下研究了相应凸松弛模型 (12) 的组变量选择相容性, 为这类列 $\ell_{2,0}$-正则最小二乘损失问题的凸松弛法提供理论支撑. 虽然凸规划 (12) 隶属非多面体二阶锥优化, 但仍适于采用加速邻近梯度法 [5]、块坐标下降法 [103]、交替方向乘子法 [41, 43] 等一阶算法进行求解.

对前两节中的稀疏矩阵优化问题, 我们主要介绍了基于矩阵零模在特殊紧凸集上的凸包而发

展的凸松弛算法. 值得指出的是, 这些凸包诱导稀疏解的能力是弱的, 尤其是当矩阵变量限制在某些复杂约束集, 如非负矩阵集 $\left\{Z \in \mathbb{R}_{+}^{n\times p} \mid Ze = e, Z^{\mathbb{T}}e = e\right\}$ 时, 采用相应的凸松弛方法甚至不能产生稀疏解. 在过去十几年里, 对向量零模提出了许多非凸代理, 如 ℓ_p $(0 < p < 1)$ 代理 [21,22,52]、光滑凹近似 [13,88,102]、SCAD 和 MCP 等折叠凹函数 [33,109]、Soubies 等人 [94] 针对零模正则最小二乘问题提出的精确连续松弛、以及文献 [69] 针对约束组零模正则化问题, 通过研究其等价 MPEC 问题的全局精确罚导出的等价 DC 代理. 与零模的凸包相比, 这些非凸代理可以更好地逼近甚至等价零模, 从而藉助它们可以期望设计出更有效的稀疏矩阵优化问题的凸松弛算法.

4 低秩矩阵优化问题的模型与算法

根据对目标矩阵的秩是否有较紧的上估计, 低秩优化问题一般模型化为秩约束优化和秩 (正则) 极小化问题. 设 $f\colon \mathbb{R}^{n\times m} \to \mathbb{R}_+$ 是损失函数, 秩约束优化问题通常描述为

$$\min_{X\in\mathbb{R}^{n\times m}} \Big\{ f(X) \ \ \text{s.t. } \operatorname{rank}(X) \le r, X \in \Omega \Big\} \tag{13}$$

其中正整数 r 是目标矩阵秩的上紧估计, $\Omega \subseteq \mathbb{R}^{n\times m}$ 为闭凸集; 而秩极小化问题一般为

$$\min_{X\in\mathbb{R}^{n\times m}} \Big\{ \operatorname{rank}(X) \ \ \text{s.t. } f(X) \le \delta,\ X \in \Omega \Big\} \tag{14}$$

其中常数 $\delta \ge 0$ 表示噪声水平. 与秩极小化问题不同, 秩正则极小化问题是通过控制正则参数 $\nu > 0$ 在损失函数极小与秩函数极小之间寻求一个权衡目标矩阵, 其模型如下

$$\min_{X\in\mathbb{R}^{n\times m}} \Big\{ \nu f(X) + \operatorname{rank}(X) \ \ \text{s.t. } X \in \Omega \Big\}. \tag{15}$$

由于秩函数的组合性, 一种处理低秩矩阵优化问题 (13)-(15) 的方法是寻求秩函数的有效凸或非凸代理、然后通过解相应的代理模型或其凸松弛模型来得到秩优化问题的满意解, 这种方法是直接对矩阵变量 X 进行优化, 每步迭代会涉及矩阵的奇异值分解; 另一种处理低秩矩阵优化问题 (13)-(15) 的方法是利用目标矩阵是低秩的信息将变量 X 用低秩双因子形式 $UV^{\mathbb{T}}$ 代替，然后通过解双因子分解模型来得到秩优化问题的满意解.

4.1 代理模型与凸松弛法

本节主要概述低秩优化问题 (13)-(15) 的几类常见代理模型 (包括核范数代理模型、Schatten-p 拟范数代理模型和等价 DC 代理模型) 和基于这些模型而发展的凸松弛法. 如无特殊说明, 下文出现的 $\mathcal{A}$ 是 $\mathbb{R}^{n\times m}$ 到 $\mathbb{R}^N$ 的线性映射, $M \in \mathbb{R}^{n\times m}$ 是未知真实矩阵.

4.1.1 核范数凸代理模型

核范数凸松弛法, 作为控制领域中迹范数启发式法的推广, 是由 Fazel 在其博士论文中首次提出的 [31]. 受向量稀疏优化 ℓ_1-范数凸代理法的良好性能启发, 她证实了核范数是秩函数在算子范数单位球上的最紧凸代理, 并提出用核范数代替秩函数、通过解单个核范数凸优化问题来得到低秩解的凸松弛方法. 特别地, 就仿射约束 $\mathcal{A}(X)=\mathcal{A}(M)$ 的秩极小化问题, Recht 等人 [87] 率先在 $\mathcal{A}$ 的限制同构条件下证明了如下核范数优化问题

$$\min_{X\in\mathbb{R}^{n\times m}}\Big\{\|X\|_* \quad \text{s.t.}\ \ \mathcal{A}(X)=\mathcal{A}(M)\Big\} \tag{16}$$

有唯一最优解 $X^*=M$, 并证实当 $\mathcal{A}$ 是服从某些分布的随机采样算子时, 所需的限制同构条件会以高概率成立, 从而建立核范数凸松弛法对仿射约束秩极小化问题的精确恢复保证. 几乎同时, Candès 和 Recht [17] 注意到源于矩阵填充的仿射约束秩极小化问题的 $\mathcal{A}$ 并不满足限制同构性, 他们在 M 的行和列空间满足一定不相关条件下, 证实当无噪声均匀采样的采样数达到一定下界时, 相应的问题 (16) 会以高概率有唯一最优解 $X^*=M$. 之后, 就噪声矩阵感知和填充问题, Candès [19] 和 Negahban 等人 [75,76] 分别在采样算子的限制同构和限制强凸条件下, 建立了如下核范数正则化问题

$$\min_{X\in\mathbb{R}^{n\times m}}\Big\{\frac{1}{2}\|\mathcal{A}(X)-b\|^2+\lambda\,\|X\|_*\Big\} \tag{17}$$

的最优解到真实解的误差界. 这些理论结果为核范数凸代理模型提供了相应的统计保证.

受核范数凸代理模型的理论保证激发, 许多求解核范数优化问题的有效算法被相继提出, 如求解核范数极小化问题 (16) 的奇异值阈值算法 [16] 以及 (线性化) 增广拉格朗日函数法 [68,105]; 求解核范数正则化问题 (17) 的不动点延拓算法 [71]、加速邻近梯度法 [98] 和线性化交替方向法. 这些算法基本都依赖于核范数的邻近算子, 如奇异值阈值算法 [16] 是采用对偶法解 (16) 的 Tiknohov 强凸正则化形式, 其基本迭代公式为

$$\begin{cases} X^k=\mathcal{P}_{\tau,\|\cdot\|_*}(\mathcal{A}^*(y^{k-1})) \quad (\tau>0\ \text{为常数});\\ y^k=y^{k-1}+\alpha_k(\mathcal{A}(M)-\mathcal{A}(X^k)) \end{cases}$$

其中 $\alpha_k>0$ 为步长; 而不动点延拓算法 [71] 和加速邻近梯度法 [98] 的基本迭代公式为

$$\begin{cases} Y^k=X^k-\tau\mathcal{A}^*(\mathcal{A}(X^k)-b);\\ X^{k+1}=\mathcal{P}_{\tau\lambda,\|\cdot\|_*}(Y^k) \quad (\tau>0\ \text{为常数}). \end{cases}$$

4.1.2 Schatten-p 拟范数代理模型

虽然核范数是秩函数在谱范数单位球上的最紧凸包, 但两者之间存在很大差别, 因而对一般的尤其是某些结构的低秩矩阵恢复问题, 该方法将面临挑战甚至失效; 例如, 当用此方法求解金融领域中的低秩相关矩阵填充问题时, 由于核范数在可行域上恒为常数, 解核范数凸松弛问题并不能产生低秩解. 鉴于此, 一些研究者通过寻求秩函数的有效非凸代理来设计启发式法, 如 Schatten-p 拟范数代理法. 该方法是用 Schatten-p 拟范数或其光滑形式代替秩函数、通过解

代理问题的一阶线性近似来得到秩 (正则) 极小化问题的满意解. 文献 [58] 和 [108] 就仿射约束 $\mathcal{A}(X)=\mathcal{A}(M)$ 秩极小化问题, 分别在线性映射 $\mathcal{A}$ 的适当限制同构条件或 Schatten-p 拟范数诱导的零空间条件下, 证实如下非凸代理问题

$$\min_{X\in\mathbb{R}^{n\times m}}\Big\{\|X\|_p^p\quad\text{s.t.}\ \ \mathcal{A}(X)=\mathcal{A}(M)\Big\}\tag{18}$$

具有唯一全局最优解 $X^*=M$, 由此建立了此类非凸代理法对仿射约束秩极小化问题的精确恢复保证, 其中 $\|X\|_p^p$ $(0<p<1)$ 表示矩阵 X 的 Schatten-p 拟范数. 此外, Rohde 等人 [90] 还在线性映射 $\mathcal{A}$ 的适当限制同构条件下导出如下 Schatten-p 拟范数正则问题

$$\min_{X\in\mathbb{R}^{n\times m}}\Big\{\|\mathcal{A}(X)-b\|^2+\lambda\|X\|_p^p\Big\}\tag{19}$$

的预测误差界以及全局最优解到真实解的 Schatten-q $(p\le q\le 2)$ 误差界. 这些结果在一定程度上证实了非凸代理模型 (18)-(19) 是好的, 但因问题 (18)-(19) 的非凸性, 求解它们仍需采用凸松弛方法, 故这些结果与计算满意解的凸松弛法的理论之间还存在着差距.

目前, 求解 Schatten-p 拟范数代理问题的凸松弛法主要是迭代重加权最小二乘法, 其基本思想是寻求光滑化的 Schatten-p 拟范数代理问题一阶近似的最优解. 例如, 针对光滑化的 Schatten-p 拟范数正则最小二乘问题, Lai 等人 [59] 提出如下迭代重加权法

$$\begin{cases} W^k=\big((X^k)^{\mathbb{T}}X^k+\epsilon_k^2 I\big)^{p/2-1}; & \text{(20a)}\\ X^{k+1}\in\mathop{\arg\min}\limits_{X\in\mathbb{R}^{n\times m}}\Big\{\dfrac{1}{2}\|X(W^k)^{1/2}\|_F^2+\dfrac{1}{2\lambda q}\|\mathcal{A}(X)-b\|^2\Big\}; & \text{(20b)}\\ \epsilon_{k+1}=\min\big(\epsilon_k,\alpha\sigma_{\kappa+1}(X^{k+1})\big) & \text{(20c)}\end{cases}$$

并在线性映射 $\mathcal{A}$ 的限制同构条件下建立了迭代序列 $\{X^k\}$ 的任意极限点到真实解的误差界, 其中迭代步 (20c) 中的 $\alpha\in(0,1)$ 为常数, 正整数 κ 是预估的秩.

4.1.3 等价 DC 代理模型

前两节讨论的凸代理模型或非凸代理模型都只是低秩矩阵优化问题的近似, 它们与低秩优化问题一般具有不同的全局最优解集, 因而基于这些代理模型的凸松弛法的有效性在很大程度上取决于它们与低秩优化模型的逼近程度. 根据 (3) 中秩函数的变分刻画, 秩 (正则) 极小化问题本质是带平衡约束的矩阵规划, 如问题 (15) 等价如下矩阵 MPEC

$$\min_{X\in\Omega,W\in\mathbb{B}}\Big\{\nu f(X)+\sum_{i=1}^n\phi(\sigma_i(W))\ \ \text{s.t.}\ \ \|X\|_*-\langle X,W\rangle=0\Big\}.$$

文献 [12, 69] 通过研究平衡约束 $\|X\|_* - \langle X, W\rangle = 0$ 诱导的罚问题的全局精确惩罚性, 即证实存在 $\overline{\rho} > 0$ 使得对应每个 $\rho > \overline{\rho}$ 的如下罚问题与矩阵 MPEC 有相同全局最优解集:

$$\min_{X\in\Omega, W\in\mathbb{B}} \Big\{\nu f(X) + \sum_{i=1}^{n} \phi(\sigma_i(W)) + \rho\big(\|X\|_* - \langle X, W\rangle\big)\Big\},$$

提供了一种构造低秩矩阵优化问题等价 DC 代理的机制, 并藉助等价 DC 代理或 MPEC 的全局精确罚提出了多阶段凸松弛法. 以秩正则最小二乘问题为例, 其基本迭代步为

$$\begin{cases} X^{k+1} \in \mathop{\arg\min}\limits_{X\in\Omega} \Big\{\dfrac{1}{2}\|\mathcal{A}(X) - b\|^2 + \rho(\|X\|_* - \langle X, W^k\rangle)\Big\}; & \text{(21a)} \\ W^{k+1} \in \mathop{\arg\min}\limits_{W\in\mathbb{B}} \Big\{\displaystyle\sum_{i=1}^{n} \phi(\sigma_i(W)) - \rho\langle X^k, W\rangle\Big\}. & \text{(21b)} \end{cases}$$

由于子问题 (21b) 具有闭式全局最优解, 该方法每步迭代只需求解一个核半范数正则化问题. 文献 [12] 在映射 $\mathcal{A}$ 的适当限制特征值条件下, 导出了每阶段最优解 X^k 到真实解的误差界、量化了第一阶段最优解 X^1 的误差界在之后各阶段的下降量、并建立了误差界序列 $\{\|X^k - M\|\}$ 在统计意义下的几何收敛速率. 对于此方法, 目前尚不能得到整个序列 $\{X^k\}$ 的收敛性, 但将 (21a) 中的损失函数用其在 X^k 点处的二次近似代替来产生点列 $\{X^k\}$, 则藉助适当辅助函数及其 KL 性质可以证得 $\{X^k\}$ 的收敛性, 详见 [70] 的分析.

对于秩约束优化问题 (13), 注意到 $\mathrm{rank}(X) \le r$ 成立当且仅当 $\|X\|_* - \|X\|_{(r)} = 0$, 其中 $\|\cdot\|_{(r)}$ 表示矩阵的 Ky-Fan r-范数, 因此可将其等价写成如下 DC 约束矩阵规划

$$\min_{X\in\mathbb{R}^{n\times m}} \Big\{f(X) \ \ \text{s.t.}\ \|X\|_* - \|X\|_{(r)} = 0, X \in \Omega\Big\}.$$

文献 [38] 藉助 DC 约束 $\|X\|_* - \|X\|_{(r)} = 0$ 诱导的罚问题提出了求解秩约束优化问题 majorized 罚极小化法. 对于某些简单闭凸集 Ω, 文献 [11] 证实这种 DC 约束诱导的罚实际是全局精确罚, 从而相应的秩约束优化问题在全局意义下等价于矩阵 DC 规划.

低秩矩阵优化问题的凸松弛法历经十几年的研究, 虽然在理论上已取得很大的进展, 但因每步迭代都涉及一个 (通常是满秩矩阵的) 奇异值分解, 在数值计算上仍面临着严峻的挑战. 特别地, 此计算瓶颈制约了凸松弛方法在一些大规模低秩矩阵优化中的应用.

4.2 因子分解模型与算法

因子分解法是由 Burer 和 Monteiro [14] 最早提出并用于求解低秩半正定规划. 与凸松弛法不同, 它不是直接优化矩阵变量 $X \in \mathbb{R}^{n\times m}$, 而是根据目标矩阵是低秩的信息将变量 X 用其双因子形式 $UV^{\mathbb{T}}$ 代替, 然后直接求解所得到的非凸因子模型. 当对目标矩阵的秩有较紧的上估计 (如正整数 r) 时, 通常采用秩约束优化问题 (13) 的因子模型

$$\min_{U\in\mathbb{R}^{n\times r}, V\in\mathbb{R}^{m\times r}} \Big\{f(UV^{\mathbb{T}}) \quad \text{s.t.}\ \ UV^{\mathbb{T}} \in \Omega\Big\}; \tag{22}$$

否则考虑秩 (正则) 极小化模型或其代理模型的因子形式更为合理. 设正整数 κ 是目标矩阵秩的粗略上估计, 根据 (4) 中秩函数的因子刻画, 秩正则问题 (15) 的因子模型为

$$\min_{U\in\mathbb{R}^{n\times\kappa},V\in\mathbb{R}^{m\times\kappa}}\left\{\nu f(UV^{\mathbb{T}})+\frac{1}{2}\big(\|U\|_{2,0}+\|V\|_{2,0}\big)\quad\text{s.t.}\ \ UV^{\mathbb{T}}\in\Omega\right\}\tag{23}$$

其中列 $\ell_{2,0}$-范数正则项可藉助参数 $\nu>0$ 的调整极大缩减所求因子的非零列数, 从而得到接近最优秩的双因子. 由 [89] 中核范数的双因子刻画, 核范数代理模型的因子形式为

$$\min_{U\in\mathbb{R}^{n\times\kappa},V\in\mathbb{R}^{m\times\kappa}}\left\{\nu f(UV^{\mathbb{T}})+\frac{1}{2}\big(\|U\|_F^2+\|V\|_F^2\big)\quad\text{s.t.}\ \ UV^{\mathbb{T}}\in\Omega\right\}.\tag{24}$$

与因子模型 (23) 相比, 模型 (24) 中的正则项并不是使所求因子矩阵的某些列为零来接近最优秩, 而是藉助核范数的降秩功能来得到秩比 κ 更小的双因子. 正如前面所提到的, 核范数的降秩能力是弱的, 因此当真实矩阵 (或目标矩阵) 的秩与其粗略上估计 κ 相差很大时, 求解核范数因子模型 (24) 依然会得不到满意的低秩解.

由于因子分解模型比原低秩矩阵优化模型涉及更少的变量个数、尤其是因子分解法每步迭代免去了凸松弛方法所需的奇异值分解的工作量, 它在一些大规模低秩矩阵优化问题 (如传感器定位 [10]、用户偏好预测 [29,57]、量子层析 [44]) 的处理中得到广泛应用, 引发近几年对因子分解模型理论进行研究的热潮. 虽然因子模型极大缩减了变量数, 但因子分解的双线性结构使得因子模型具有很高的非凸性, 引入许多额外的非全局最优的稳定点. 因此, 对因子分解模型的一个研究方向是致力于刻画其目标函数的几何蓝图或证实目标函数是否具有严格鞍点性质, 即验证每个稳定点要么是局部最优解、要么其海瑟矩阵至少存在一个严格负的特征值, 从而沿着负曲率方向可以继续缩减目标值. 目前, 此方向的大部分研究工作 (见文献 [9,39,40,82,116,117]) 是围绕着秩约束优化问题的精确参数化因子模型 (即 $r=\mathrm{rank}(M)$) 或其带平衡项 $\|U^{\mathbb{T}}U-V^{\mathbb{T}}V\|_F^2$ 的正则化形式开展的; 只有文献 [65] 考虑了核范数正则化因子模型 (24) 的精确和过参数形式 (即 $\kappa\geq\mathrm{rank}(M)$), 并在损失函数的适当限制强凸与限制光滑条件下证实了每个稳定点要么是全局最优解、要么其海瑟矩阵有严格负的特征值. 此方向的研究成果意味着像梯度下降这类局部搜索法求解相应的因子模型可以收敛到全局最优解.

另一研究方向是从局部视角来研究因子模型 (22)-(24) 或其带平衡项的正则化因子模型, 其目的是刻画目标函数在全局最优解集附近的增长性, 此方向的研究工作大多与具体的算法联系在一起 (见文献 [54,81,95,100,113,114]). 下面来介绍求解因子分解模型的常见算法. **(1)** 当损失函数 f 连续可微时, 可直接应用某些经典非线性规划方法来进行求解; 例如, 文献 [87] 中就采用了增广拉格朗日函数法解核范数极小化问题 (16) 的因子模型. **(2)** 对因子分解模型 (22)-(24), 当因子矩阵 U 和 V 之一固定时, 若其变成可解的矩阵优化问题, 则可以采用块坐标下降法求解, 以因子模型 (22) 为例, 其基本迭代步为

$$\left\{\begin{aligned}&U^{k+1}\in\mathop{\arg\min}_{U\in\mathbb{R}^{n\times r}}\Big\{f(U(V^k)^{\mathbb{T}})\quad\text{s.t.}\ \ U(V^k)^{\mathbb{T}}\in\Omega\Big\},&&\text{(25a)}\\&V^{k+1}\in\mathop{\arg\min}_{V\in\mathbb{R}^{m\times r}}\Big\{f(U^{k+1}V^{\mathbb{T}})\quad\text{s.t.}\ \ U^{k+1}V^{\mathbb{T}}\in\Omega\Big\}.&&\text{(25b)}\end{aligned}\right.$$

例如, 针对仿射约束 $\mathcal{A}(X)=\mathcal{A}(M)$ 秩极小化问题, Jain 等人 [54] 应用上述块坐标下降法求解 f 为最小二乘损失的因子模型 (22), 并通过限制初始点的选取, 在映射 $\mathcal{A}$ 的限制同构条下证实了块坐标下降法 (25a)-(25b) 几何收敛到矩阵感知问题的真实解 M, 在真实矩阵 M 的行和列空间满足一定不相关条件下, 证实该方法在一定采样数下会以高概率几何收敛到矩阵填充问题的真实解 M. 文献 [101] 还将块坐标下降法应用于求解矩阵填充问题的如下因子分解形式, 其中 $\Upsilon\subseteq\{1,\ldots,n\}\times\{1,\ldots,m\}$ 为采样指标集:

$$\min_{X\in\mathbb{R}^{n\times r},Y\in\mathbb{R}^{m\times r},Z\in\mathbb{R}^{n\times m}}\Big\{\frac{1}{2}\|XY-Z\|_F^2\quad\text{s.t.}\ \ Z_{ij}=M_{ij}\ \ \forall(i,j)\in\Upsilon\Big\}.$$

通过缩减最小二乘子问题中的矩阵乘积工作量, 该文提出了如下非线性过松弛法:

$$\begin{cases} Z_\omega\leftarrow\omega Z+(1-\omega)XY, & \text{(26a)}\\ X_+(\omega)\leftarrow Z_\omega Y^{\mathbb{T}}\ \text{或}\ Z_\omega Y^{\mathbb{T}}(YY^{\mathbb{T}})^{\dagger}, & \text{(26b)}\\ Y_+(\omega)\leftarrow\big(X_+(\omega)^{\mathbb{T}}X_+(\omega)\big)^{\dagger}(X_+(\omega)^{\mathbb{T}}Z_\omega), & \text{(26c)}\\ \mathcal{P}_{\Upsilon^c}(Z_+(\omega))\leftarrow\mathcal{P}_{\Upsilon^c}(X_+(\omega)Y_+(\omega)),\ \mathcal{P}_{\Upsilon}(Z_+(\omega))\leftarrow\mathcal{P}_{\Upsilon}(M), & \text{(26d)}\end{cases}$$

其中 $\omega\geq 1$ 是松弛参数, $A^{\dagger}$ 表示矩阵 A 的 Moore-Penrose 伪逆, 而 Υ^c 代表指标集 Υ 的补集. 虽然该算法 (简称 LMaFit) 目前尚缺乏恢复性理论保证, 但在实际计算中展示出良好性能. **(3)** 对涉及光滑损失函数的因子模型, 局部搜索梯度下降法也是常见的解法. 以仿射约束 $\mathcal{A}(X)=\mathcal{A}(M)$ 的秩极小化问题为例, 该方法先运行有限个低秩投影梯度步

$$\begin{cases}\textbf{For}\ k=0,1,2,\ldots,T_0-1\\ \qquad\widetilde{M}^{k+1}:=\mathcal{P}_{\Xi}\big(\widetilde{M}^k-\alpha_{k+1}\mathcal{A}^*(\mathcal{A}(\widetilde{M}^k-b))\big)\\ \textbf{End}\qquad(\text{其中}\ \widetilde{M}^0=0,\Xi\ \text{表示秩}\ r\ \text{约束集})\end{cases}$$

产生满意初始解 $(U^0,V^0)=(P\Sigma,Q\Sigma)$, 其中 $(P,Q)\in\mathbb{O}^{n,m}(\widetilde{M}^{T_0}),\Sigma=\mathrm{Diag}(\sigma(\widetilde{M}^{T_0})^{\frac{1}{2}})$; 然后以 (U^0,V^0) 为初始点、采用梯度下降算法求解带平衡正则项的光滑非凸优化问题

$$\min_{U\mathbb{R}^{n\times r},V\in\mathbb{R}^{m\times r}}\Big\{\frac{1}{2}\|\mathcal{A}(UV^{\mathbb{T}})-b\|^2+\frac{1}{16}\|U^{\mathbb{T}}U-V^{\mathbb{T}}V\|_F^2\Big\}.$$

Tu 等人 [100] 在线性映射 $\mathcal{A}$ 的适当限制同构条件下, 证实该算法所产生的迭代点列会几何收敛到真实矩阵 M; Park 等人 [81] 对一般的秩约束优化问题, 在损失函数的适当限制强凸与限制光滑条件下也证实适当初始点启动的梯度下降法会具有局部线性收敛速率. **(4)** 针对矩阵填充问题, Sun 和 Luo [95] 通过研究适当约束最小二乘损失的罚因子模型的几何性质, 在标准的不相关条件下证实许多常用一阶方法 (如梯度下降、随机梯度下降、块坐标下降等) 求解该罚因子模型都会以高概率线性收敛到全局最优解.

除上述研究方向外, 还有一些工作研究因子模型的局部最优解到真实矩阵的误差界; 例如, 文献 [9] 就噪声低秩半定恢复问题, 在采样算子的限制同构条件下证实精确参数化的因子分解模型的所有局部最优解都靠近真实解, 而文献 [112] 在损失函数的限制强凸与光滑条件下证实正则化

因子模型的所有局部最优解也靠近真实解. 此外, 文献 [24] 还对噪声矩阵填充问题, 藉助因子分解模型证实核范数凸松弛法产生的误差界几乎接近最优.

4.3 非负矩阵分解问题的模型与算法

非负矩阵分解问题源于上世纪八十年代线性代数领域的研究, 后来因 Lee 和 Seung 的工作 [62] 而得到广泛关注. 作为数据降维的一种重要方法, 它在文本挖掘、计算机视觉、光谱数据分析、盲源分离等领域中具有广泛的应用 [36,37]. 非负矩阵分解问题的主旨是对已知高维数据矩阵 $A\in\mathbb{R}^{n\times m}$, 在给定维数为 $\kappa<n$ 的空间中寻求非负低秩矩阵 $W^*\in\mathbb{R}^{n\times\kappa}$ 和 $H^*\in\mathbb{R}^{m\times\kappa}$ 使得它们的积 $W^*(H^*)^{\mathbb{T}}$ 等于矩阵 A, 其标准优化模型为

$$\min_{W\in\mathbb{R}^{n\times\kappa},H\in\mathbb{R}^{m\times\kappa}}\Big\{\|WH^{\mathbb{T}}-A\|_F^2\quad\text{s.t.}\ \ W\geq 0,H\geq 0\Big\}. \tag{28}$$

对照第 4.2 节中的模型 (22), 显然, 这类问题可视为特殊限制的低秩矩阵因子分解问题.

由于非负矩阵分解问题带了一类硬型约束, 不仅寻求其全局最优解证实是 NP-难的, 而且一般也不存在能成功求解问题 (28) 的凸松弛方法. 目前所期望的计算问题 (28) 的好方法是能够产生其局部最优解的算法. 在求解 (28) 的诸多算法中, 比较有效的是交替非负最小二乘法和分层交替最小二乘法, 其中前者是两矩阵块的坐标下降法, 迭代公式为

$$\begin{cases} W^{k+1}\leftarrow\operatorname*{arg\,min}_{W\geq 0}\Big\{\|H^kW^{\mathbb{T}}-A^{\mathbb{T}}\|_F^2\Big\}, & \text{(29a)}\\ H^{k+1}\leftarrow\operatorname*{arg\,min}_{H\geq 0}\Big\{\|W^{k+1}H^{\mathbb{T}}-A\|_F^2\Big\}; & \text{(29b)}\end{cases}$$

后者是基于问题 (28) 向量形式的 2κ 个向量块的坐标下降法, 基本迭代公式如下

$$\begin{cases}\textbf{For } j=1,2,\ldots,\kappa\\ \qquad W_j^{k+1}\leftarrow\operatorname*{arg\,min}_{W_j\geq 0}\Big\{\Big\|\sum\limits_{i<j}^{\kappa}W_i^{k+1}(H_i^k)^{\mathbb{T}}+W_j(H_j^k)^{\mathbb{T}}+\sum\limits_{i>j}^{\kappa}W_i^k(H_i^k)^{\mathbb{T}}-A\Big\|_F^2\Big\}\\ \textbf{End}\\ \textbf{For } j=1,2,\ldots,\kappa\\ \qquad H_j^{k+1}\leftarrow\operatorname*{arg\,min}_{H_j\geq 0}\Big\{\Big\|\sum\limits_{i<j}^{\kappa}W_i^{k+1}(H_i^{k+1})^{\mathbb{T}}+W_j^{k+1}H_j^{\mathbb{T}}+\sum\limits_{i>j}^{\kappa}W_i^{k+1}(H_j^k)^{\mathbb{T}}-A\Big\|_F^2\Big\}\\ \textbf{End}\end{cases}$$

其中 W_i 和 H_i 分别是矩阵 W 和 H 的第 i 列. 在适当的条件下，上述两个算法所产生点列的极限点是非凸优化问题 (28) 的稳定点, 详细讨论见文献 [56].

对于非凸优化模型 (28), 在假设数据矩阵 A 是由公式 $A=\overline{W}\,\overline{H}^{\mathbb{T}}$ 产生且生成因子 $\overline{W}$ 与 $\overline{H}$ 之一满足适当的可分离与零元模式条件下, Donoho [30] 与 Laurberg 等人 [60] 证实了其可识别性, 即模型 (28) 的全局最优解以某种意义恢复真实因子矩阵对 $(\overline{W},\overline{H})$. 注意到真实因子矩阵的可分离条件在实际中通常并不满足, Huang 等人 [50] 在生成因子满足充分分散条件及零元模式

下证实了优化模型 (28) 的可识别性. 近来, 为了除去前面工作中对真实因子的零元模式限制, 文献 [36,37] 提出了新的非负矩阵分解的优化模型.

5 低秩加稀疏矩阵优化的模型和算法

低秩加稀疏矩阵优化问题的典例是矩阵分离问题, 即已知真实矩阵 M 是低秩矩阵 $\overline{L}$ 与稀疏型矩阵 $\overline{S}$ (如元素稀疏或列稀疏) 的叠加, 能否在适当数量的观测下将 M 的低秩部分与稀疏部分恢复出来? 从主成分分析角度看, 就是如何从 M 的高噪声观测中恢复出其主成分, 从而提供一种鲁棒主成分分析. 矩阵分离或鲁棒主成分分析在计算机视觉、多任务学习、潜在语义索引、视频监控等领域中有着广泛应用 [2,74]. 用 $\psi(S)$ 表示矩阵 S 的元素零模 $\|S\|_0$ 或列 $\ell_{2,0}$-范数 $\|S\|_{2,0}$, 矩阵分离问题可模型化为秩加零模极小化问题

$$\min_{L,S\in\mathbb{R}^{n\times m}}\Big\{\mathrm{rank}(L)+\mu\psi(S)\quad \text{s.t.}\quad \|\mathcal{A}(L+S)-Y\|_F\le\delta,\, L\in\Omega\Big\},\tag{31}$$

其中 $\mathcal{A}\colon\mathbb{R}^{n\times m}\to\mathbb{R}^{n\times m}$ 是采样算子, $Y\in\mathbb{R}^{n\times m}$ 是噪声观测, $\Omega\subseteq\mathbb{R}^{n\times m}$ 是简单闭凸集. 当 $\delta>0$ (即带噪声) 时, 矩阵分离问题也经常被模型化为如下秩加零模正则化问题

$$\min_{L,S\in\mathbb{R}^{n\times m}}\Big\{\frac{1}{2}\|\mathcal{A}(L+S)-Y\|_F^2+\lambda\,\mathrm{rank}(X)+\mu\psi(S)\quad \text{s.t.}\quad L\in\Omega\Big\}.\tag{32}$$

类似于低秩矩阵优化问题, 一种处理秩加零模优化问题 (31)-(32) 的方法是用秩函数与零模的有效凸与非凸代理代替它们，然后求解相应的代理模型或其凸松弛模型; 另一方法是根据目标矩阵 $\overline{L}$ 是低秩的信息, 将变量 L 用其低秩因子分解代替，然后求解所得到的因子分解模型. 下面从这两个方面来概述低秩加稀疏矩阵优化的模型与算法发展.

5.1 代理模型与凸松弛法

受零模优化问题的 ℓ_1-凸代理和秩优化问题的核范数凸代理的影响, 核范数加矩阵 ℓ_1-范数 (或列 $\ell_{2,1}$-范数) 模型自然成为秩加零模优化问题的首选代理模型, 求解该凸代理模型的方法通常称为主成分追踪法. 用 $\mathcal{R}(S)$ 表示矩阵 S 的元素 ℓ_1-范数或列 $\ell_{2,1}$-范数. 秩加零模极小化问题 (31) 的主成分追踪法是通过解如下凸优化问题来得到其满意解:

$$\min_{L,S\in\mathbb{R}^{n\times m}}\Big\{\|X\|_*+\lambda\mathcal{R}(S)\quad \text{s.t.}\quad \|\mathcal{A}(L+S)-Y\|_F\le\delta,\, L\in\Omega\Big\};\tag{33}$$

而秩加零模正则化问题 (32) 的压缩主成分追踪法是通过解如下凸规划得到满意解的:

$$\min_{L,S\in\mathbb{R}^{n\times m}}\Big\{\frac{1}{2}\|\mathcal{A}(L+S)-Y\|_F^2+\lambda\|X\|_*+\mu\mathcal{R}(S)\quad \text{s.t.}\quad L\in\Omega\Big\}.\tag{34}$$

针对 $\Omega=\mathbb{R}^{n\times m}$, Chandrasekaran 等人 [20] 就确定性的稀疏部分, 在秩-稀疏不相关条件下建立了无噪声全采环境下核范数加 ℓ_1-范数极小化问题 (33) 的精确恢复保证, 而 Candès 等人 [18] 就

随机稀疏部分, 在真实低秩矩阵的不相关条件下建立其精确恢复保证; Zhou 等人 [51,115] 在真实低秩矩阵的不相关条件下建立了噪声全采环境下凸优化模型 (33) 的最优解到真实解 $(\overline{L},\overline{S})$ 的误差界; 而 Agarwal 等人 [2] 在较弱的 spikiness 条件下建立了核范数加矩阵 ℓ_1-范数正则化问题 (34) 的最优解到真实解 $(\overline{L},\overline{S})$ 的误差界. 这些结果为秩加零模 (正则) 极小化问题的主成分追踪法提供了理论保证.

考虑到核范数与矩阵 ℓ_1-范数 (或列 $\ell_{2,1}$-范数) 具有较弱的诱导低秩解与稀疏解能力, 文献 [47] 还藉助秩函数与零模的局部 Lipschitz 非凸代理提出了求解秩加零模极小化问题 (31) 的两阶段凸松弛法, 并通过建立每阶段最优解到真实解 $(\overline{L},\overline{S})$ 的误差界、以及证实两阶段凸松弛会缩减第一阶段凸松弛的误差界为其提供理论保证.

上面提到的秩加零模 (正则) 极小化问题的凸松弛模型一般都是涉及两块或三块非光滑的线性等式约束凸规划, 因此, 可直接应用乘子交替方向法 (ADMM) [41–43] 及近来发展的对称 Gauss-Seidel 型的 ADMM [25,67] 对其进行求解.

5.2 因子分解模型与块坐标下降法

考虑到低秩矩阵优化问题的凸松弛方法在计算上所面临的瓶颈, Gu 等人 [45] 利用目标矩阵是低秩的信息, 将变量 L 用其低秩双因子形式 $UV^{\mathbb{T}}$ 代替给出了如下分解模型

$$\min_{U\in\mathbb{R}^{n\times\kappa},V\in\mathbb{R}^{m\times\kappa},S\in\Lambda}\left\{\frac{1}{2}\big\|\mathcal{A}(UV^{\mathbb{T}}+S)-Y\big\|_F^2\right\},\tag{35}$$

其中 $\Lambda:=\{S\in\mathbb{R}^{n\times m}\mid\|S\|_0\le s\}$ 是零模约束集. 观察到变量对 (U,S) 和 (V,S) 之一固定时, 模型 (35) 就变成可解凸规划, 而零模约束集 Λ 上的投影算子具有闭式表达, 因此, 可以采用如下块坐标下降法进行求解, 其中 $\mathrm{QR}(Z)$ 表示矩阵 Z 的 QR 分解:

$$\begin{cases}U^{k+1}\in\mathop{\arg\min}\limits_{U\in\mathbb{R}^{n\times\kappa}}\left\{\frac{1}{2}\big\|\mathcal{A}\big(U(\overline{V}^k)^{\mathbb{T}}+\overline{S}^k\big)-Y\big\|_F^2\right\}, & \text{(36a)}\\ \{\overline{U}^{k+1},R^{k+1}\}=\mathrm{QR}(U^{k+1}), & \text{(36b)}\\ V^{k+1}\in\mathop{\arg\min}\limits_{V\in\mathbb{R}^{m\times\kappa}}\left\{\frac{1}{2}\big\|\mathcal{A}\big(\overline{U}^{k+1}V^{\mathbb{T}}+\overline{S}^k\big)-Y\big\|_F^2\right\}, & \text{(36c)}\\ \{\overline{V}^{k+1},R^{k+1}\}=\mathrm{QR}(V^{k+1}), & \text{(36d)}\\ S^{k+1}\in\mathop{\arg\min}\limits_{S\in\mathbb{R}^{n\times m}}\left\{\frac{1}{2}\big\|\mathcal{A}\big(\overline{U}^{k+1}(V^{k+1})^{\mathbb{T}}+S\big)-Y\big\|_F^2\right\}, & \text{(36e)}\\ \overline{S}^{k+1}\in\mathcal{P}_\Lambda(S^{k+1}). & \text{(36f)}\end{cases}$$

对于此算法, Gu 等人 [45] 通过适当选取初始点 $(\overline{U}^0,\overline{V}^0,\overline{S}^0)$, 在真实低秩矩阵 $\overline{L}$ 的不相关条件以及线性映射 $\mathcal{A}$ 的限制同构条件下建立了其局部线性收敛速率.

值得指出的是, 为了克服凸松弛法所面临的计算瓶颈, Netrapalli [77] 等人对无噪声全采的秩加零模极小化问题, 通过将适当损失残量交替向秩约束集与零模约束集的投影, 提出一种非凸交替投影法, 并在真实低秩阵的不相关条件下建立其精确恢复保证.

6 低秩张量优化问题的模型与算法

前面介绍的低秩稀疏矩阵优化问题的模型和算法为二维数据的降维处理提供了解决办法, 但随着计算机技术的迅速发展, 许多科学工程领域面临着多维数据的处理, 张量因其捕获多元线性结构的能力而成为建模的自然选择. 低秩张量优化问题, 作为存储高阶张量数据的主要压缩策略, 被相继提出. 虽然张量是矩阵的直接推广, 但将低秩矩阵优化的许多结果推广到张量并不显然, 其原因之一在于张量的数值代数中到处是 NP-难问题, 譬如, 张量的核范数计算都是 NP-难的 [35], 从而低秩张量的核范数优化模型并不像在矩阵环境下是可解的凸规划; 其原因之二在于张量秩的定义不唯一, 对应不同的张量分解可以提出不同的张量秩. 较为流行的张量分解有 CP 分解和 Tucker 分解, 分别是矩阵奇异值分解和主成分分析的高阶推广. 下面介绍对应这两种分解的低秩张量优化模型与算法.

6.1 低 Tucker 秩张量优化问题的模型与算法

对于一个 n-阶张量 $\mathcal{X}\in\mathbb{R}^{m_1\times m_2\times\cdots\times m_n}$, 其 Tucker 秩是由下式定义的 n 维向量:

$$\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}):=\big(\mathrm{rank}(\mathcal{X}_{(1)}),\mathrm{rank}(\mathcal{X}_{(2)}),\ldots,\mathrm{rank}(\mathcal{X}_{(n)})\big),$$

其中 $\mathcal{X}_{(i)}\in\mathbb{R}^{m_i\times\Pi_{j\neq i}m_j}$ 是张量 $\mathcal{X}$ 沿模态 i 展开的矩阵. 设 $f:\mathbb{R}^{m_1\times\cdots\times m_n}\to\mathbb{R}_+$ 是定义在张量空间中 $\mathbb{R}^{m_1\times m_2\times\cdots\times m_n}$ 中的损失函数, 低 Tucker 秩张量优化问题可以模型化为

$$\min_{\mathcal{X}\in\mathbb{R}^{m_1\times\cdots\times m_n}}\Big\{\mathrm{rank}_{\mathrm{tc}}(\mathcal{X})\quad \text{s.t.}\quad f(\mathcal{X})\leq\delta\Big\} \tag{37}$$

其中常数 $\delta\geq 0$ 表示噪声水平. 根据 Tucker 秩的定义, 低 Tucker 秩张量优化问题的本质是矩阵向量优化问题. 受低秩矩阵优化问题的核范数代理模型及向量优化问题的常用标量化形式启发, 下述加权核范数凸代理模型在过去几年得到研究者的较多关注:

$$\begin{cases}\displaystyle\min_{\mathcal{X}\in\mathbb{R}^{m_1\times\cdots\times m_n}}\Big\{\sum_{i=1}^{n}\alpha_i\|\mathcal{X}_{(i)}\|_*\quad \text{s.t.}\quad f(\mathcal{X})\leq\delta\Big\}; & \text{(38a)}\\[2ex] \displaystyle\min_{\mathcal{X}\in\mathbb{R}^{m_1\times\cdots\times m_n}}\Big\{f(\mathcal{X})+\sum_{i=1}^{n}\alpha_i\|\mathcal{X}_{(i)}\|_*\Big\}. & \text{(38b)}\end{cases}$$

其中, Tomioka [99] 和 Raskutti 等人 [85] 分别针对低 Tucker 秩张量恢复问题和多响应张量回归问题, 在一定限制强凸条件下建立了模态展开矩阵加权核范数正则模型 (38b) 的统计理论保证. 注意到模型 (38a)-(38b) 是具有可分结构的凸规划, 从而适于采用块坐标下降法、ADMM 和对称 Gauss-Seidel 型的 ADMM 求解.

不难看到, 低 Tucker 秩张量优化问题的处理本质是在处理低秩矩阵优化. 事实上, 由于张量秩的定义大多是藉助张量的矩阵化形式给出的, 所以相应低秩张量优化问题的处理自然回归到低秩矩阵优化的处理, 如低 TT 秩和 Tubal 秩张量恢复问题 [83, 111].

6.2 张量的 CP 因子分解模型与算法

类似于低秩矩阵优化问题的因子分解法, 藉助张量的 CP 分解可以提出寻求低 CP 秩张量估计的因子分解法. 设 $\mathcal{G}$ 是某 n-阶低 CP 秩张量 $\mathcal{M} \in \mathbb{R}^{m_1 \times \cdots \times m_n}$ 的噪声观测. 根据目标张量具有低 CP 秩的信息, 可以通过求解如下非凸因子分解模型

$$\min_{x_1^{(1)},\ldots,x_1^{(n)},\ldots,x_K^{(1)},\ldots,x_K^{(n)}} \left\| \mathcal{G} - \sum_{i=1}^{K} x_i^{(1)} \circ x_i^{(2)} \circ \cdots \circ x_i^{(n)} \right\|_F^2$$

来得到 $\mathcal{M}$ 的满意低 CP 秩张量估计, 其中 "$\circ$" 表示向量外积, $x_1^{(j)},\ldots,x_K^{(j)} \in \mathbb{R}^{m_j}$, 而正整数 K 是对真实张量 $\mathcal{M}$ 的 CP 秩的估计. 对每个 $j \in \{1,2,\ldots,n\}$, 定义因子矩阵 $X^{(j)} = \left[x_1^{(j)}\ x_2^{(j)}\ \cdots\ x_K^{(j)}\right] \in \mathbb{R}^{m_j \times K}$. 根据矩阵 Khatri-Rao 积, 上述优化模型可等价写为

$$\min_{X^{(1)},\ldots,X^{(n)}} \left\| \mathcal{G}_{(i)} - X^{(i)} \left(X^{(n)} \odot \cdots \odot X^{(i+1)} \odot X^{(i-1)} \odot \cdots \odot X^{(1)} \right)^{\mathbb{T}} \right\|_F^2, \tag{39}$$

其中 "$\odot$" 表示矩阵 Khatri-Rao 积. 目前, 比较流行的求解因子分解模型 (39) 的方法是块坐标下降法 [26], 其每步的基本迭代公式为

$$\begin{cases} \textbf{For } i = 1,2,\ldots,n \\ \qquad X^{(i)} \longleftarrow \mathcal{G}_{(i)} \left(X^{(n)} \odot \cdots \odot X^{(i+1)} \odot X^{(i-1)} \odot \cdots \odot X^{(1)} \right) \Big((X^{(n)})^{\mathbb{T}} X^{(n)} \circledast \\ \qquad\qquad \cdots \circledast (X^{(i+1)})^{\mathbb{T}} X^{(i+1)} \circledast (X^{(i-1)})^{\mathbb{T}} X^{(i-1)} \circledast \cdots \circledast (X^{(1)})^{\mathbb{T}} X^{(1)} \Big)^{\dagger} \\ \textbf{End} \end{cases}$$

其中 "$\circledast$" 表示矩阵的 Hadamard 积, 而 "$\dagger$" 代表矩阵的 Moore-Penrose 伪逆. 关于求解非凸模型 (39) 的块坐标下降法的收敛性分析, 感兴趣的读者可参阅文献 [104].

7 关键问题与挑战

低秩稀疏矩阵优化的模型研究在于根据问题的应用背景、藉助恰当的数据 "稀疏" 表示建立合理的秩与零模矩阵优化模型; 其算法研究在于为所建立的秩与零模矩阵优化问题提供实际有效且有理论保证的计算方法; 而其理论研究不仅要为算法产生的解提供统计支撑, 还要分析算法的全局收敛性及局部收敛速率. 本节将从这三个方面来分别阐述目前低秩稀疏矩阵优化的模型、算法与理论研究中遇到的关键问题与挑战.

7.1 数据的恰当 "稀疏" 表示

数据的恰当 "稀疏" 表示是低秩稀疏矩阵优化模型研究的关键, 但目前在此方向上的研究并不充分, 主要体现在如下两个方面: **(1)** 复合结构的 "稀疏" 表示探索不足. 对许多问题, 藉助单一的低秩、元素稀疏或列稀疏并不能建立具有良好泛化能力的秩与零模矩阵优化模型, 往往需要藉助像同时低秩稀疏、非负低秩、对角低秩等复合结构 "稀疏" 来建模, 如多任务机器学习、二

次压缩感知、稀疏相位恢复以及社交网络图的刻画等; 详见文献 [80] 中的讨论. 此外, 还有些问题的 "稀疏" 性需藉助决策矩阵的适当变换的低秩或元素稀疏来表示, 如 fused 稀疏 [97]、组稀疏 [107]、低 Tucker 秩张量优化的有效标量化. **(2)** 缺乏对近似 "稀疏" 的探讨. 实际数据都是带有噪声的, 近似 "稀疏" 的目标矩阵可能比精确 "稀疏" 的更接近实际要求, 这就提出如何度量近似低秩和近似稀疏的问题?

7.2 基于非凸代理的凸松弛法的理论保证

秩与零模矩阵优化问题的非凸代理尤其是等价非凸代理比其核范数和/或 ℓ_1-范数凸代理更接近原来带有组合性的矩阵优化问题, 但是这些非凸代理模型的全局最优解几乎是得不到的, 一般的求解算法只能产生它们的稳定点. 这样, 就需要从模型角度分析什么样的条件能 (以高概率) 保证这些非凸代理模型的稳定点是好的, 而实现此目标的关键是刻画非凸代理模型的全局最优解集或某类稳定点集到真实解的误差界. 对求解一般非凸非光滑优化问题的算法, 往往容易得到子列的收敛性而不是整个序列的收敛性, 而为了建立算法的全局收敛性以及在稳定点处的局部快速收敛速率, 就需对目标函数的适当正则性质, 如指数为 $[0,1/2]$ 的 KL 性质 [3, 4], 进行深入研究. 此外, 秩与零模矩阵优化问题的现有凸松弛法的理论研究大多是针对光滑损失函数进行的, 而对涉及非光滑损失的秩与零模矩阵优化问题的凸松弛法的理论研究更具有挑战性.

7.3 因子分解模型的理论保证

低秩矩阵优化问题的因子分解法在数值计算上占有优势, 但由于双因子的引入额外增加了很多非局部最优的稳定点, 从而给其理论分析带来很多困难. 根据第 5.2 节的讨论知, 当前因子分解法的理论研究大多是针对秩约束矩阵优化问题 (即假设目标矩阵的秩或其上紧估计 κ 已知) 开展的, 而对于秩正则矩阵优化问题, 仅限于研究核范数凸松弛的因子模型. 虽然核范数凸松弛的因子模型具有满意光滑性, 但其降秩能力是弱的. 因此, 研究秩优化问题本身的因子模型或其非凸代理因子模型的理论保证是十分必要的, 尽管其非凸非光滑性给研究带来了极大挑战. 另外, 低秩矩阵优化的现有因子分解模型研究大多是针对光滑损失函数, 而根据文献 [8] 知, 鲁棒型的低秩矩阵优化模型一般都会涉及非光滑的损失函数. 这样, 对秩约束或正则的非光滑损失极小化问题的因子分解法的理论研究也是有重要价值的. 特别地, 对张量的 CP 因子分解模型, 不仅没有探索其非凸目标函数的几何蓝图, 而且对所提出的一些求解算法尚未研究统计意义下的收敛速率.

7.4 低秩矩阵优化问题的有效计算

直接求解秩优化问题的凸或非凸代理模型, 每步迭代都会涉及对一个满秩矩阵进行特征值或奇异值分解; 对于无约束或简单约束的凸或非凸代理模型, 利用最优解是低秩的特点, 可以采用前 κ 个最大特征值或奇异值分解来缩减每步迭代的计算工作量, 其中正整数 κ 是最优解的秩的上紧估计; 目前面临的挑战是对于复杂尤其是 hard 型约束的凸与非凸代理模型, 如何有效运用特征值或奇异值分解来降低计算量. 因子分解模型的每步迭代的工作量主要集中在矩阵乘积的计算上; 对低秩矩阵填充问题, 文献 [101] 通过巧妙缩减最小二乘子问题中的矩阵乘积工作量, 提出了

求解其秩约束优化因子分解模型的非线性过松弛算法-LMaFit, 而文献 [48] 藉助正交子空间迭代法提出了适于解大规模的核范数凸松弛因子模型的交替子空间算法; 但对一般秩约束优化和秩正则极小化问题的因子模型, 如何利用其结构特点缩减矩阵乘积工作量仍是极具挑战的计算问题.

8 总结

低秩稀疏矩阵优化问题是一类与数据处理密切相关的新型优化问题, 只有与其它应用学科 (如生物、物理、化学、工程) 紧密结合才能从具体应用中挖掘新的 “稀疏” 性, 只有与统计和机器学习等学科交叉才能提出新的低秩稀疏矩阵优化模型和求解算法. 对这类重要且困难的矩阵优化问题, 可以尝试从如下几个方面继续开展其模型与算法研究:

(1) 从深度学习中发掘低秩矩阵优化问题. 为了学习大数据中的特征, 一些大规模深度学习和深度计算模型被相继提出来, 而这些模型中的参数与特征映射的冗余是导致其高计算量和高内存消耗的关键因素. 参数与特征映射的冗余主要反映在加权矩阵和特征映射的结构性质上, 如何利用 “稀疏” 矩阵和张量建立适宜的学习模型除去这些冗余是大规模神经网络压缩的关键. 这意味着, 整个神经网络结构的压缩蕴藏了许多新型低秩稀疏矩阵/张量优化问题, 而如何有效训练压缩后的神经网络又对低秩矩阵/张量优化问题的求解算法提出新的挑战.

(2) 基于数据类型开展低秩稀疏矩阵优化的研究. 基于图的数据 (如搜索引擎查询、社交网络的检测和垃圾邮件过滤) 和实时数据流 (如在线广告、推荐系统和电子商务中的点击和查询记录) 的分析, 建立合理的低秩稀疏矩阵优化模型; 针对快速 MRI 成像、低剂量 CT 成像、优质 PET 成像、高维多模态图像分析、医学图像检索、相位恢复以及低温电子显微镜和三维重构中的若干反问题, 建立合适的低秩稀疏矩阵优化模型, 并通过分析和研究模型的性质, 发展快速有效的算法.

(3) 优化理论与随机分析相结合的分析策略. 现有的非凸非光滑优化理论 (见专著 [91]) 关注的是优化问题在某点邻域中的确定性质, 而要验证该点在某个方向区域以高概率具有某种性质, 则需藉助随机分析技术. 事实上, 近几年通过将随机分析与优化理论相结合, 为一些 NP-难低秩稀疏矩阵优化问题提供了相应的统计保证, 但这种分析策略主要针对的是有明显结构的低秩稀疏优化模型, 如何将其推广应用于一般低秩稀疏矩阵优化问题的模型与算法分析, 尚需进行更深入研究.

参考文献

[1] A. Agarwal, S. Negahban and M. J. Wainwright, *Fast global convergence of gradient methods for high-dimensional statistical recovery*, The Annals of Statistics, 40(2012): 2452-2482.

[2] A. Agarwal, S. Negahban and M. J. Wainwright, *Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions*, The Annals of Statistics, 40(2012): 1171-1197.

[3] H. Attouch, J. Bolte, P. Redont and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality*, Mathematics of Operations Research, 35(2010): 438-457.

[4] H. Attouch, J. Bolte and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and reguarlized Gauss-Seidel methods*, Mathematical Programming, 137(2013): 91-129.

[5] H. Attouch and A. Cabot, *Convergence rate of inertial forward-backward algorithm*, SIAM Journal on Optimization, 28(2018): 849-874.

[6] O. Banerjee, L. E. Ghaoui and A. d́aspremont, Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary Data, Journal of Machine Learning Research, 9(2008): 485-516.

[7] H. H. Bauschke, D. R. Luke, H. M. Phan and X. F. Wang, *Restricted normal cones and sparsity optimization with affine constraints*, Foundations of Computational Mathematics, 14(2014): 63-83.

[8] D. Bertsimas and M. S. Copenhaver, *Characterization of the equivalence of robustification and regularization in linear and matrix regression*, European Journal of Operational Research, 2017: 1-12.

[9] S. Bhojanapalli, B. Neyshabur and N. Srebro, *Global optimality of local search for low rank matrix recovery*, Advances in Neural Information Processing Systems. 2016: 3873-3881.

[10] P. Biswas and Y. Y. Ye, *Semidefinite programming for ad hoc wireless sensor network localization*, In Proceedings of the 3rd international symposium on Information processing in sensor networks, pp. 46-54, 2004.

[11] S. J. Bi and S. H. Pan, *Error bounds for rank constrained optimization problems and applications*, Operations Research Letters, 44(2016): 336-341.

[12] S. J. Bi and S. H. Pan, *Multistage convex relaxation approach to rank regularized minimization problems based on equivalent mathematical program with a generalized complementarity constraint*, SIAM Journal on Control and Optimization, 55(2017): 2493-2518.

[13] P. S. Bradley and O. L. Mangasarian, *Feature selection via concave minimization and support vector machines*, In Proceeding of international conference on machine learning ICML, 1998.

[14] S. Burer and R. D. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs with low-rank factorization*, Mathematical Programming, 95(2003): 329-357.

[15] T. Cai, W. D. Liu and X. Luo, *A constrained ℓ_1 minimization approach to sparse precision matrix estimation*, Journal of the American Statistical Association, 106(2011): 594-607.

[16] J. F. Cai, E. J. Candès and Z. W. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization, 20(2010): 1956-1982.

[17] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics, 9(2009): 717-772.

[18] E. J. Candès, X. D. Li, Y. Ma and J. Wright, *Robust principal component analysis*, Journal of the ACM, 58(2009): 1-37.

[19] E. J. Candès and Y. Plain, *Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements*, IEEE Transactions on Information Theory, 57(2011): 2342-2359.

[20] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo and A. S. Willsky *Rank-sparsity incoherence for matrix decomposition*, SIAM Journal on Optimization, 21(2009): 572-596.

[21] R. Chartrand, *Exact reconstruction of sparse signals via nonconvex minimization*, IEEE Signal Processing Letters, 14(2007): 707-710.

[22] X. J. Chen, F. M. Xu and Y. Y. Ye, *Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization*, SIAM Journal on Scientific Computing, 32(2010): 2832-2852.

[23] Y. D. Chen, *Incoherence-optimal matrix completion*, IEEE Transactions on Information Theory, 60(2015): 2909-2923.

[24] Y. X. Chen, Y. J. Chi, J. Q. Fang, C. Ma and Y. L. Yan, *Noisy matrix completion: understanding statistical guarantees for convex relaxation via nonconvex optimization*, arXiv:1902.07698v2, 2019.

[25] L. Chen, D. F. Sun and K. C. Toh, *An effcient inexact symmetric Gauss-Seidel based majorized ADMM for high-dimensional convex composite conic programming*, Mathematical Programming, 161(2017): 237-270.

[26] A. Cichocki, N. Lee, I. V. Oseledets, A-H. Phan, Q. B. Zhao and D. P. Mandic, *Low-Rank Tensor Networks for Dimensionality Reduction and Large-Scale Optimization Problems: Perspectives and Challenges PART1*, arXiv:1609.00893v3.

[27] R. M. Corless, P. M. Gianni, B. M. Trager and S. M. Watt, *The singular value decomposition for polynomial systems*, Proceedings of the 1995 international symposium on Symbolic and algebraic computation. ACM, 1995: 195-207.

[28] A. Daniilidis, A. S. Lewis, J. Malick and H. Sendov, *Prox-regularity of spectral functions and spectral sets*, Journal of Convex Analysis, (15)2008: 547-560.

[29] D. DeCoste, *Collaborative prediction using ensembles of maximum margin matrix factorizations.* In Proceedings of the 23rd International Conference on Machine learning, pp. 249-256, 2006.

[30] D. Donoho and V. Stodden, *When does non-negative matrix factorization give a correct decomposition into parts*, Proc. Adv. Neural Inf. Process. Syst., 2004, pp. 1141-1148.

[31] M. Fazel, *Matrix Rank Minimization with Applications*, PhD thesis, Stanford University, 2002.

[32] M. Fazel, T. K. Pong, D. F. Sun and P. Tseng, *Hankel matrix rank minimization with applications to system identification and realization.* SIAM Journal on Matrix Analysis and Applications, 34(2013): 946-977.

[33] J. Q. Fan and R. Z. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of American Statistics Association, 96(2001): 1348-1360.

[34] J. Friedman, T. Hastie and R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics, 9(2008): 432-441.

[35] S. Friedland and L. H. Lim, *Nuclear norm of higher-order tensors*, Mathematics of Computation, 87(2018): 1255-1281.

[36] X. Fu, K. Huang and N. D. Sidiropoulos, *On identifiability of nonnegative matrix factorization*, IEEE Transactions on Signal Processing, 25(2018): 2306-2320.

[37] X. Fu, K. Huang, N. D. Sidiropoulos and W. K. Ma, *Nonnegative matrix factorization for signal and data analytics: identifiability, algorithms, and applications*, arXiv preprint arXiv:1803.01257.

[38] Y. Gao and D. F. Sun, A majorized penalty approach for calibrating rank constrained correlation matrix problems, Technical report, Department of Mathematics, National University of Singapore, Singapore, 2010.

[39] R. Ge, J. D. Lee and T. Ma, *Matrix completion has no spurious local minimum*, Advances in Neural Information Processing Systems, 2016: 2973-2981.

[40] R. Ge, C. Jin and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, Proceedings of the 34th International Conference on Machine Learning, 70(2017): 1233-1242.

[41] R. Glowinski, *Lectures on numerical methods for nonlinear variational problems.* In: Tata Institute of Fundamental Research Lectures on Mathematics and Physics, vol. 65. Tata Insti-

tute of Fundamental Research, Bombay. Notes by M. G. Vijayasundaram and M. Adimurthi (1980).

[42] R. Glowinski, *On alternating direction methods of multipliers: a historical perspective*, Modeling, Simulation and Optimization for Science and Technology, Springer Netherlands, pp. 59-82, 2014.

[43] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite element approximation*, Computers and Mathematics with Applications, vol. 2, pp. 17-40, 1976.

[44] D. Gross, Y. K. Liu, S. T. Flammia, S. Becker and J. Eisert, *Quantum state tomography via compressed sensing*, Physical Review Letters, 105(15):150401, 2010.

[45] Q. Q. Gu, Z. R. Wang and H. Liu, *Low-rank and sparse structure pursuit via alternating minimization*, International Conference on Artificial Intelligence and Statistics, 2016: 600-609.

[46] B. D. Haeffele, E. Yang and R. Vidal, *Structured low-rank matrix factorization: optimality, algorithm and applications to image processing.* In: Proceedings of the 31st International Conference on Machine Learning (ICML), pp. 2007-2015(2014)

[47] L. Han, S. J. Bi and S. H. Pan, *Two-stage convex relaxation approach to least squares loss constrained low-rank plus sparsity optimization problems*, Computational Optimization and Applications, 64(2016): 119-148.

[48] T. Hastie, R. Mazumder, J. D. Lee and R. Zadeh, *Matrix completion and low-rank SVD via fast alternating least squares* Journal of Machine Learning Research, 16(2015): 3367-3402.

[49] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Presss, Cambridge, 1991.

[50] K. Huang, N. Sidiropoulos and A. Swami, *Non-negative matrix factorization revisited: uniqueness and algorithm for symmetric decomposition*, IEEE Transactions on Signal Processing, 62(2014): 211-224.

[51] D. Hsu, S. M. Kakade and T. Zhang, *Robust matrix decomposition with sparse corruptions*, IEEE Transactions on Information Theory, 57(2011): 7221-7234.

[52] Y. H. Hu, C. Li, K. W. Meng, J. Qin and X. Q. Yang, *Group sparse optimization via $\ell_{p,q}$ regularization*, Journal of Machine Learning Research, 18(2017): 1-52.

[53] C. J. Hsieh, M. A. Sustik, I. S. Dhillon and P. Ravikumar, *QUIC: Quadratic approximation for sparse inverse covariance estimation*, Journal of Machine Learning Research, 15(2014): 2911-2947.

[54] P. Jain, P. Netrapalli and S. Sanghavi, *Low-rank matrix completion using alternating minimization*, ACM Symposium on Theory of Computing, 2013: 665-674.

[55] P. Jain, A. Tewari and P. Kar, *On iterative hard thresholding methods for high-dimensional M-Estimation*, Proceedings of the 27th International Conference on Neural Information Processing Systems, 1(2014): 685-693.

[56] J. G. Kim, Y. L. He and H. Park, *Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework*, Journal of Global Optimization, 58(2014): 285-319.

[57] Y. Koren, R. Bell and C. Volinsky, *Matrix factorization techniques for recommender systems*, Computer, 8(2009): 30-37.

[58] L. C. Kong and N. H. Xiu, *Exact low-rank matrix recovery via nonconvex schatten p-minimization*, Asia-Pacific Journal of Operational Research, 30(2013): 1340010.

[59] M. J. Lai, Y. Y. Xu and W. T. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed minimization*, SIAM Journal on Numerical Analysis, 51(2013): 927-957.

[60] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen and S. Jensen, *Theorems on positive data: on the uniqueness of NMF*, Comput. Intell. Neurosci., 2008, Art. no. 764206.

[61] H. Y. Le, Generalized subdifferentials of the rank function, Optimization Letters, 7(2013): 731-743.

[62] D. Lee and H. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature, 401(1999): 788-791.

[63] A. S. Lewis and H. S. Sendov, *Nonsmooth analysis of singular values. Part I: Theory*, Set-Valued Analysis, 2005(13): 213-241.

[64] A. S. Lewis and H. S. Sendov, *Nonsmooth analysis of singular values. Part II: Applications*, Set-Valued Analysis, 2005(13): 243-264.

[65] Q. W. Li, Z. H. Zhu and G. G. Tang, *The non-convex geometry of low-rank matrix optimization*, Information and Inference: A Journal of the IMA, 8(2018): 51-96.

[66] G. Y. Li and T. K. Pong, *Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods*, Foundations of Computational Mathematics, 18(2018): 1199-1232.

[67] X. D. Li, D. F. Sun and K. C. Toh, *A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions*, Mathematical Programming, 155(2016): 333-373.

[68] Y. J. Liu, D. F. Sun and K. C. Toh, *An implementable proximal point algorithmic framework for nuclear norm minimization*, Mathematical Programming, 133(2012): 399-436.

[69] Y. L. Liu, S. J. Bi and S. H. Pan, *Equivalent Lipschitz surrogates for zero-norm and rank optimization problems*, Journal of Global Optimization, 72(2018): 679-704.

[70] T. X. Liu, T. K. Pong and A. Takeda, *A refined convergence analysis of* $pDCA_e$ *with applications to simultatneous sparse recovery and outlier detection*, Computation Optimization and Applications, 2019, `https://doi.org/10.1007/s10589-019-00067-z`.

[71] S. Q. Ma, D. Goldfarb and L. F. Chen, Fixed point and Bregman iterative methods for matrix rank minimization, Mathematical Programming, 128(2011): 321-353.

[72] O. L. Mangasarian, *Machine learning via polyhedral concave minimization*, In H. Fischer, B. Riedmueller & S. Schaeffler (Eds.), Applied mathematics and parallel computing-Festschrift for Klaus Ritter (pp. 175-188), 1996.

[73] I. Markovsky, *Recent progress on variable projection methods for structured low-rank approximation*, Signal Processing，96(2014): 406-419.

[74] K. R. Min, Z. D. Zhang, J. Wright and Y. Ma, *Decomposing background topics from keywords by principal component pursuit*, ACM on Information and Knowledge Management, 2010: 269-278.

[75] S. Negahban and M. J. Wainwright, *Estimation of (near) low-rank matrices with noise and high-dimensional scaling*, The Annals of Statistics, 39(2011): 1069-1097.

[76] S. Negahban and M. J. Wainwright, *Restricted strong convexity and weighted matrix completion: Optimal bounds with noise*, Journal of Machine Learning Research, 13(2012): 1665-1697.

[77] P. Netrapalli, U. N. Niranjan, S. Sanghavi and A. Anandkumar, *Provable non-convex robust PCA*, Advances in neural information processing systems, 2(2014): 1107-1115.

[78] G. Obozinski, M. J. Wainwright and M. I. Jordan, *Union support recovery in high-dimensional multivariate regression*, The Annals of Statistics, 39(2011): 1-47.

[79] G. Obozinski, B. Taskar and M. I. Jordan, *Joint covariate selection and joint subspace selection for multiple classification problems*, Statistical Computing, 20(2010): 231-252.

[80] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar and B. Hassibi, *Simultaneously structured models with application to sparse and low-rank matrices*, IEEE Transactions on Information Theory, 61(2015): 2886-2908.

[81] D. Park, A. Kyrillidis, C. Caramanis and S. Sanghavi, *Finding low-rank solution via non-convex matrix factorization efficiently and provably*, SIAM Journal on Imaging Sciences, 11(2018): 2165-2204.

[82] D. Park, A. Kyrillidis, C. Caramanis and S. Sanghavi, *Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach*, In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54(2017): 65-74.

[83] H. N. Phien, H. D. Tuan, J. A. Bengua and M. J. Do, *Efficient tensor completion: Low-rank tensor train*, arXiv:1601.01083, 2016.

[84] R. Pietersz and P. J. F. Groenen, *Rank reduction of correlation matrices by majorization*, Quantitative Finance, 4(2004): 649-662.

[85] G. Raskutti, M. Yuan and H. Chen, *Convex regularization for high-dimensional multiresponse tensor regression*, arXiv:1512.01215.

[86] P. Ravikumar, M. J. Wainwright, G. Raskutti and B. Yu, *High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence noise*, Electronic Journal of Statistics, 5(2011): 935-980.

[87] B. Recht, M. Fazel and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Review, 52(2010): 471-501.

[88] F. Rinaldi, F. Schoen and M. Sciandrone, *Concave programming for minimizing the zero-norm over polyhedral sets*, Computation Optimization and Applications, 46(2010): 467-486.

[89] J. D. M. Rennie and N. Srebro, *Fast maximum margin matrix factorization for collaborative prediction*, In Proceedings of the 22nd International Conference on Machine Learning, 2005: 713-719.

[90] A. Rohde and A. B. Tsybakov, *Estimation of high-dimensional low-rank matrices*, The Annals of Statistics, 39(2011): 887-930.

[91] R. T. Rockafellar and R. J-B. Wets, *Variational Analysis*, Springer, 1998.

[92] N. Srebro, *Learning with matrix factorizations*, PhD thesis, Massachusetts Institute of Technology, 2004.

[93] S. Shalev-Shwartz, N. Srebro and T. Zhang, *Trading accuracy for sparsity in optimization problems with sparsity constraints*, SIAM Journal on Optimization, 20(2010): 2807-2832.

[94] E. Soubies, L. Blang-Fraud and G. Aubert, *A unified view of exact continuous penalities for ℓ_2-ℓ_0 minimization*, SIAM Journal on Optimization, 8(2017): 1067-1639.

[95] R. Y. Sun and Z. Q. Luo, *Guaranteed matrix completion via non-convex factorization*, IEEE Transactions on Information Theory, 62(2016): 6535-6579.

[96] A. Tasissa and R. J. Lai, *Exact reconstruction of Euclidean distance geometry problem using low-rank matrix completion*, IEEE Transactions on Information Theory, 65(2019): 3124-3144.

[97] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu and K. Knight, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society (B), 67(2005): 91-108.

[98] K. C. Toh and S. Yun, *An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems*, Pacific Journal of Optimization, 6(2010): 615-640.

[99] R. Tomioka, T. Suzuki, K. Hayashi and H. Kashima, *Statistical performance of convex tensor decomposition*, Advances in Neural Information Processing Systems, 2011: 972-980.

[100] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi and B. Recht, *Low-rank solution of linear matrix equations via procrustes flow*, In International Conference on Machine Learning, 48(2016): 964-973.

[101] Z. W. Wen, W. T. Yin and Y. Zhang, *Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm*, Mathematical Programming Computation, 4(2012): 333-361.

[102] J. Weston, A. Elisseef, B. Schölkopf and M. Tipping, *Use of the zero norm with linear models and kernel methods*, Journal of Machine Learning Research, 3(2003): 1439-1461.

[103] S. J. Wright, *Coordinate descent algorithms*, Mathematical Programming, 151(2015): 3-34.

[104] Y. Y. Xu and W. T. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM Journal on Imaging Science, 3(2013): 1758-1789.

[105] J. F. Yang and X. M. Yuan, *Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization*, Mathematics of Computation, 82(2013): 301-329.

[106] M. Yuan and Y. Lin, *Model Selection and Estimation in the Gaussian Graphical Model*, Biometrika, 94(2007): 19-35.

[107] L. Yuan, J. Liu and J. P. Ye, *Efficient methods for overlapping group Lasso*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(2013): 2104-2116.

[108] M. C. Yue and A. M. C. So, *A perturbation inequality for concave functions of singular values and its applications in low-rank matrix recovery*, Applied and Computational Harmonic Analysis, 40(2016): 396-416.

[109] C. H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Annals of Statistics, 38(2010): 894-942.

[110] Y. Zhang and Q. Yang, *An overview of multi-task learning*, National Science Review, 5(2018): 30-43.

[111] Z. Zhang and S. Aeron, *Exact tensor completion using t-SVD*, IEEE Transcations on Signal Processing, 65(2017): 1511-1526.

[112] X. Zhang, L. X. Wang, Y. D. Yu and Q. Q. Gu, *A Primal-Dual Analysis of Global Optimality in Nonconvex Low-Rank Matrix Recovery*, In International conference on machine learning, 80(2018): 5862-5871.

[113] T. Zhao, Z. Wang and H. Liu, *A nonconvex optimization framework for low rank matrix estimation*, Advances in Neural Information Processing Systems, 1(2015): 559-567.

[114] Q. Zheng and J. Lafferty, *A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements*, Advances in Neural Information Processing Systems, 2015: 109-117.

[115] Z. H. Zhou, X. D. Li, J. Wright, E. J. Candès and Y. Ma, *Stable principal component pursuit*, IEEE International Symposium on Information Theory, Austin, Texas, U.S.A., 2010: 1518-1522.

[116] Z. H. Zhu, Q. W. Li, G. G. Tang and M. B. Wakin, *Global optimization in low-rank matrix optimization*, IEEE Transactions on Signal Processing, 66(2018): 3614-3628.

[117] Z. H. Zhu, Q. W. Li, G. G. Tang and M. B. Wakin, *The global optimization geometry in low-rank matrix optimization*, arXiv:1703.01256v2, 2018.