大数据学习的 slow kill

Yiyuan She、Jiahui Shen 和 Adrian Barbu，佛罗里达州立大学统计学系

arXiv:2305.01726v1 [stat.ML] 2 May 2023

摘要

大数据应用通常涉及大量观测和特征，这给变量选择和参数估计带来了新的挑战。本文提出一种名为“slow kill”的新技术，该技术利用非凸约束优化、自适应 ℓ_2-收缩以及递增的学习率。在 slow kill 迭代过程中，问题规模能够逐渐减小，这使其在大规模变量筛选中尤为有效。统计与优化之间的相互作用为如何控制分位数、步长和收缩参数提供了有价值的见解，从而放宽达到预期统计准确性所需的正则性条件。在真实数据和合成数据上的实验结果表明，在各种情形下，slow kill 的表现均优于最先进的算法，同时在处理大规模数据时具有较高的计算效率。

索引词

自顶向下算法 、稀疏性、非凸优化、非渐近分析、亚奈奎斯特频谱感知

I. 引言

本文研究如何在大数据应用中构建简约且具有预测能力的模型，其中预测变量的数量和观测值的数量都可能非常庞大。设 $y \in \mathbb{R}^n$ 为包含 n 个样本的响应向量， $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n\times p}$ 为由 p 个特征或预测变量构成的设计矩阵。考虑一个一般的学习问题，其损失函数为 $l_0(X\beta; y)$，用于衡量 $X\beta$ 与 y之间的差异。由于 p 可能远大于 n，因此通常使用促进稀疏性的正则化项来实现模型简约性

$$\min_{\beta\in\mathbb{R}^p} l_0(X\beta; y) + P(\beta; \lambda), \tag{1}$$

其中 λ 为正则化参数。正则化项和损失函数有许多可选形式，二者都不一定是凸的。在许多情况下， l_0 可能是负对数似然函数，但我们将考虑一个不一定基于似然的更一般设定。

过去十年中，惩罚问题（1）的极小化器相关统计理论取得了重大进展。然而，现代科学家经常会遇到大数据带来的挑战，即使存在凸性，也难以获得全局最优的估计量。本文旨在将计算方面的考量纳入统计建模，从而构建一个具有理论保证的新型大数据学习框架。在应对大规模变量选择中的这些挑战时，理想的算法应具备以下特征：

（a）易于调参。在实践中，人们通常希望获得具有预定基数（或特定变量数量，以 q 表示）的解。然而，使用针对惩罚问题（1）设计的算法可能需要过多的计算，而且在试图实现这一目标时，正则化参数 λ 可能不够直观。许多实践者会对 λ 进行网格搜索。然而，在处理大数据时，网格必须足够细密，以涵盖潜在有用的候选模型，从而带来沉重的计算负担。

(b) 可扩展性。除了高效之外，理想的算法还应易于实现。由于临时设计的过程可能不可靠，因此，与依赖启发式方法相比，采用基于优化的算法更为可取。如果算法能够根据可用的计算资源自适应地调整其参数，也将具有优势，而这需要理解算法的迭代复杂度和每次迭代的成本。

(c) 统计保证。人们普遍认为，当设计矩阵具有低相干性且信号足够强时，LASSO 在变量选择方面表现出色 [1, 2]。一些更简单、更快速的方法，例如用于变量筛选的方法 [3], ，基于特征相互独立（或仅存在轻微相关性）的假设。尽管这些弱相关性假设允许大幅剔除特征，但对于现实世界中的高维数据而言，它们往往限制过多。求解带有 ℓ_0 型惩罚项 [4] 的式 (1) 的全局最优解

无论相干性如何，都具有统计上可靠的保证，但在计算上仅适用于小型数据集。因此，更迫切的挑战是设计一种迭代过程，以放宽达到最优统计准确性所需的严格正则性条件。

本文提出了一种称为 slow kill 的新方法，以应对上述挑战。该算法的主要特点如下。

- 有趣的是，slow kill 的工作方向与正向路径方法和提升法算法相反，后者都是从零模型 [5–9]逐步构建模型。
- Slow kill 采用自适应的 ℓ_2-收缩和不断增大的学习率，以处理相干设计并减轻计算负担。其源于优化，因此具有良好的计算可扩展性，且参数易于调节。
- 从理论上看，slow kill 在统计意义下具有严格且可证明的精度保证和线性收敛性。特别是，我们的理论支持后向分位数控制和快速学习。

本文其余部分的组织如下。第 II 节研究回归环境下的混合正则化估计，以阐释 slow kill 的一些基本要素，并将其与相关工作进行比较。第 III 节针对可微损失函数介绍一般的 slow kill 过程，并分析随着循环推进统计误差的变化。第 IV 节开展大量模拟和真实数据实验，从效率和精度两个方面将 slow kill 与一些最先进的方法进行比较。我们在第 V 节总结研究结果。更多技术细节见附录。

记号和符号。以下记号和符号将在下文中使用。令 $[n] = \{1, \ldots, n\}$ 和 $\lfloor x \rfloor$ 为不大于 x的最大整数。定义 $a \vee b = \max(a, b)$ 和 $a \wedge b = \min(a, b)$。我们使用 $a \lesssim b$表示对于某个正常数 c有 $a \leq cb$，且用 c 或 C 表示的常数在每次出现时不一定相同。给定任意 $\beta \in \mathbb{R}^p$，我们使用 $\mathcal{J}(\beta) \subset [p]$ 表示其支集，即 $\mathcal{J}(\beta) = \{j : \beta_j \neq 0\}$，以及$J(\beta) = |\mathcal{J}(\beta)| = \|\beta\|_0 = \sum_{j=1}^p 1_{\beta_j \neq 0}$。给定 $I \subset [p]$，我们使用 X_I表示由 X中属于 I的列构成的子矩阵，并用 β_I 表示与 I相关联的子向量。特别地，对于任意$j \in [p]$，x_j 表示 X 的第 j列。当 A 为对称矩阵时，我们使用 A_I 表示由 A 中索引属于 I的列和行构成的子矩阵，并分别用 $\lambda_{\max}(A)$ 和 $\lambda_{\min}(A)$表示其最大和最小特征值。

给定 $X \in \mathbb{R}^{n \times p}$，受限等距数 $\rho_+(s)$、$\rho_-(s)$ [10] 分别是满足以下条件的最小值和最大值

$$\rho_-(s)\|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq \rho_+(s)\|\beta\|_2^2, \ \forall \beta \in \mathbb{R}^p : \|\beta\|_0 \leq s, \tag{2}$$

其对 X 的依赖关系略去不写。显然，$0 \leq \rho_-(s) \leq \rho_+(s) \leq \rho_+(p) = \|X\|_2^2$，其中 $\|X\|_2$ 表示 X 的谱范数。

为便于说明，我们引入分位数阈值算子 $\Theta^{\#}$，它同时执行阈值处理和 ℓ_2 收缩 [11]。给定任意 $s = [s_1, \ldots, s_p]^T \in \mathbb{R}^p$，满足 $t_{(j)} = s_{(j)}/(1+\eta)$（当 $1 \leq j \leq q$ 时）且其他情况下为 0 的 $\Theta^{\#}(s; q, \eta) = [t_1, \ldots, t_p]^T$，其中 $s_{(1)}$、.....、$s_{(p)}$ 是满足$|s_{(1)}| \geq \cdots \geq |s_{(p)}|$的 s_1、.....、s_p 的顺序统计量，而 $t_{(1)}$、.....、$t_{(p)}$ 的定义类似。为避免歧义，本文在执行 $\Theta^{\#}(s; q, \eta)$ 时作出 $\Theta^{\#}$ 唯一性假设：$|s_{(q)}| > |s_{(q+1)}|$ 或 $s_{(q)} = s_{(q+1)} = 0$ 必有其一成立。对于任意 $S = [s_1, \ldots, s_p]^T \in \mathbb{R}^{p \times m}$，多变量分位数阈值函数 $\vec{\Theta}^{\#}(S; q, \eta)$ 定义为一个 $p \times m$ 矩阵 $T = [t_1, \ldots, t_p]^T$，当 $\|s_j\|_2$ 是 $\{\|s_j\|_2$ 中最大的 q 个元素之一时，令 $t_j = s_j/(1+\eta)$；否则令其为 0。

II. 为什么选择后向选择?

本节旨在基本的回归设置中阐明一种“自上而下”的算法设计思路。二次损失是强凸损失中的一个重要情形，研究这一情形将为进一步探讨受限强凸性下的更一般问题奠定基础。

假设 $y = X\beta^* + \epsilon$，其中 $\beta^* \in \mathbb{R}^p, \|\beta^*\|_0 \leq s$ 满足 $s \leq p \wedge n$。首先，我们考虑一个受 ℓ_0 约束、采用 ℓ_2 惩罚的优化问题，以估计高维情形下的系数向量，

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \frac{\eta_0}{2}\|\beta\|_2^2 \equiv f(\beta) \ \text{ s.t. } \|\beta\|_0 \leq q. \tag{3}$$

当 X, y 未经中心化时，应在损失函数中加入截距项 1α，且 α 不受正则化约束。(1) 中的混合正则化不同于弹性网中常用的 ℓ_1 和 ℓ_2 惩罚项的线性组合

弹性网 [12]。与常规的 ℓ_1 惩罚项和其他非凸惩罚项相比， $\|\cdot\|_0$ 可以说是实现稀疏性的理想选择，而且不会带来任何不必要的偏差。约束参数 q（$\leq p$）直接控制最终模型中的变量数量，因此比惩罚参数 λ 更便于使用。同时采用 ℓ_2 惩罚项是为了补偿共线性和较大的噪声，之后还将用于克服后向消除中的一些障碍。相关的正则化参数 η_0 易于调整，且在实验中不太敏感。我们的理论分析将揭示，精心设计的收缩序列对数值稳定性和统计准确性均有益处。

问题 (3) 是非凸的，并且包含离散约束。虽然计算求解这类问题可能颇具挑战，但我们可以使用可扩展的迭代优化算法找到一个局部极小值。此外，在大数据时代，为了在“常规”问题中取得良好的统计性能，未必需要彻底求解 (3)；而分析由算法驱动的非全局估计量，对于发现提升非凸优化统计性能的新型、经济高效的方法至关重要。具体而言，为了引入一个原型算法，我们首先为 (3) 构造一个替代函数 $g(\beta, \beta^-)$，

$$g(\beta, \beta^-) = \frac{1}{2}\|y - X\beta^-\|_2^2 + \langle X^T(X\beta^- - y), \beta - \beta^- \rangle + \frac{\rho}{2}\|\beta - \beta^-\|_2^2 + \frac{\eta_0}{2}\|\beta\|_2^2,$$

其中 $\rho > 0$ 将在稍后选定，然后定义一个迭代序列：

$$\beta^{(t+1)} = \arg\min_{\beta:\|\beta\|_0 \leq q} g(\beta, \beta^{(t)}). \tag{4}$$

回顾第 I 节末尾定义的分位数阈值算子 $\Theta^{\#}$。经过一些简单的代数运算（细节略去），我们得到一个迭代分位数阈值算法

$$\beta^{(t+1)} = \Theta^{\#}\Big\{\beta^{(t)} - \frac{1}{\rho}X^T(X\beta^{(t)} - y); q, \frac{\eta_0}{\rho}\Big\}. \tag{5}$$

第一步相当于在 $\beta^{(0)} = 0$ 时进行确定性独立筛选 [3]。然而，(5) 通过迭代降低贪心性，同时保持较低的单次迭代成本。

(5) 中的更新规则具有一些令人称道的计算性质。例如，只要 ρ 足够大（更具体地说，若 $\rho \geq \rho_+(2q)$，其中 $\rho_+(\cdot)$ 在 (2) 中定义），那么无论问题的维度、相干性和信号强度如何，该算法都表现出最坏情况下的次线性收敛速率。所得解（虽不一定是最优解）可刻画为 (4) 中定义的算法映射的不动点。更多结果和技术细节，请参阅定理 A.1。

这类过程已应用于信号处理和信息处理 [11, 13], ; 在 $\eta_0 = 0$ 这一特殊情形下，(5) 中的简单更新规则属于迭代硬阈值（IHT）算法[14, 15] ，其性能仅属一般水平（参见注释 3 和第 IV 节）。事实上，通过自适应调整 (5) 中的三个关键参数 ρ、η_0、q，仍有很大的改进空间，但文献中尚未对此进行系统探索。

A. 统计误差分析：能力与局限性

虽然优化误差对于分析算法十分重要，但我们的主要关注点是统计误差。本小节将研究原型算法（5），以此为后续章节中的新技术提供动机。为了获得该算法精确的非渐近结果，需要注意的是，阈值会随着迭代次数的变化而变化，并且最终的估计量可能并非全局最优。

回忆 $y = X\beta^* + \epsilon$ with $\|\beta^*\|_0 \leq s$。令

$$\vartheta := q/s$$

在全文中使用 $\vartheta > 1$。满足以下方程、与 (5) 相关的固定点 β 称为 $\Theta^{\#}$-估计量，

$$\hat{\beta} = \Theta^{\#}\big\{\hat{\beta} - \frac{1}{\rho}X^T(X\hat{\beta} - y); q, \bar{\eta}_0\big\}, \text{ with } \bar{\eta}_0 = \eta_0/\rho. \tag{6}$$

定理 1 研究了这些估计量的统计准确性。

定理 1。假设 ϵ 是一个零均值且尺度由 σ 界定的次高斯随机向量（参见附录中的定义 ^^ A.1）。令 β 为满足条件 (6) 的任意估计量，其中 $\eta_0 \geq 0$ 满足 $\|\beta\|_0 = q$，并令 $\rho > 0$ 取为使得

$$\frac{\rho - \{(2-\varepsilon)\sqrt{\vartheta} - 1\}\eta_0}{\sqrt{\vartheta}}\|\beta\|_2^2 \leq (2-\delta)\|X\beta\|_2^2 \quad \forall \beta : \|\beta\|_0 \leq (1+\vartheta)s \tag{7}$$

对于某些 $\varepsilon, \delta > 0$。那么，以至少 $1 - Cp^{-c}$ 的概率，

$$\|X(\hat{\beta} - \beta^*)\|_2^2 \vee \frac{\eta_0 \varepsilon}{\delta}\|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{1}{\delta^2}\sigma^2 \vartheta s \log\frac{ep}{\vartheta s} + \frac{\eta_0}{\delta\varepsilon}\|\beta^*\|_2^2, \tag{8}$$

其中 $C, c > 0$ 是常数。

根据误差界，在假设 (7) 成立且将 $\vartheta, \delta, \varepsilon$ 视为常数时，(5) 可以达到极小极大最优误差率 $\mathcal{O}(\sigma^2 s \log(ep/s))$ [16], 。该结果不要求 η_0 恰好为零。事实上，正的 η_0 反而有助于满足条件 (7)（例如 $\rho = (1.9\sqrt{\vartheta} - 1)\eta_0 + 1.9\sqrt{\vartheta}\rho_-(q+s)$ 和 $\varepsilon = \delta = 0.1$，适用于 $q > n$）。另一个有趣的观察是，为了获得良好的统计准确性，应将 ρ 选得适当小；这与前面提到的数值收敛误差界 $\rho \geq \rho_+(2q)$ 相反。下面的评论将进一步扩展并比较这些结果。

注 1（估计误差界与忠实变量选择）。定理 1 的 ℓ_2-恢复结果至关重要，在适当的正则性条件下，可用于推导其他范数下的估计误差界。

定理 2. 在定理 1 的设定下，假设将正则性条件 (7) 替换为

$$\Big\{\frac{\rho - (2\sqrt{\vartheta} - 1)\eta_0}{\sqrt{\vartheta}} + \delta\rho_+((1+\vartheta)s)\Big\}\|\beta\|_2^2 \leq 2\|X\beta\|_2^2, \ \forall \beta : \|\beta\|_0 \leq (1+\vartheta)s \tag{9}$$

对于某些 $\delta > 0$。然后

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{1}{\delta^2\rho_+((1+\vartheta)s)}\sigma^2\vartheta s \log\frac{ep}{\vartheta s} + \frac{\eta_0^2}{\delta^2\rho_+((1+\vartheta)s)}\|\beta^*\|_2^2 \tag{10}$$

以至少 $1 - Cp^{-c}$ 的概率成立，其中 C、$c > 0$ 为某些常数。此外，在.....之下

$$\nu\|\beta\|_\infty \leq \|(X^TX + \eta_0 I)\beta\|_\infty / n, \quad \beta : \|\beta\|_0 \leq (1+\vartheta)s \tag{11}$$

^ 对于某个 $\nu > 0$，任意不动点 β 均满足

$$\|\hat{\beta} - \beta^*\|_\infty \leq \frac{(\rho + \eta_0)}{n\nu\sqrt{\vartheta - 1}}\frac{\|\hat{\beta} - \beta^*\|_2}{\sqrt{s}} + \frac{\|X^T\epsilon\|_\infty}{n\nu} + \frac{\eta_0}{n\nu}\|\beta^*\|_\infty, \tag{12}$$

and

$$\|(\hat{\beta} - \beta^*)_{\mathcal{J}^*}\|_\infty + (1 - \frac{\rho + \eta_0}{n\nu})\|(\hat{\beta} - \beta^*)_{\hat{\mathcal{J}}\setminus\mathcal{J}^*}\|_\infty \leq \frac{\|X^T\epsilon\|_\infty}{n\nu} + \frac{\eta_0}{n\nu}\|\beta^*\|_\infty, \tag{13}$$

^^ 其中 $\mathcal{J}^* = \mathcal{J}(\beta^*)$、 $\mathcal{J} = \mathcal{J}(\beta)$。

与 (7) 相比，(9) 的条件将 $\delta\|X\beta\|_2^2$ 替换为 $\delta\rho_+((1+\vartheta)s)\|\beta\|_2^2$。当 q 和 s 较小时，$\rho_+((1+\vartheta)s)$ 的阶为 $\mathcal{O}(n)$。因此，假设 δ、ϑ 为常数且 η_0 足够小，(10) 变为 $\|\hat{}\beta - \beta^*\|_2^2 \lesssim \{\sigma^2 s \log(ep/s)\}/n$。

此外，逐元素误差界 (12) 蕴含了正则性条件 (11) 下的忠实变量选择（与之前的正则性条件一样，该条件偏好低相干性，即 X^TX/n 的非对角线元素的幅度应相对较小）。具体而言，假设 ϑ、ν、δ 为常数，且 $\|x_j\|_2 \lesssim \sqrt{n}$、 $\rho + \eta_0 \lesssim n$ 以及 beta-min 条件 $\min_{j\in\mathcal{J}^*}|\beta_j^*| > c\sigma\{\log(ep)/n\}^{1/2}$（其中 c 为足够大的常数）成立，则 (12) 表明， $|\beta_j|$ 中最大的 $s\hat{}$ 个元素以高概率对应于 $\mathcal{J}^* = \{j : \ \beta_j^* \neq 0\}$。

备注 2（不动点与全局最优解的比较）。统计准确性结果 (8)、(10) 和 (12) 是针对由 (6) 定义的所有非全局不动点估计量证明的。我们的证明可以稍作修改，以表明如果能够计算出一个全局最优解，统计误差率保持不变，但

(7) 的左侧变为 0，这表明对于任意 $\delta \leq 2$，正则性条件始终成立。然而，依赖多个起始点来获得全局最优解，并以此提高统计性能，对于大型数据集而言可能效率低下。

注释 3（与一些理论研究的比较）。上述 IHT算法可能涉及使用硬阈值化 $\Theta_H(s;\ \lambda) = [s_i 1_{|s_i|\geq\lambda}]$，其中采用固定阈值 λ，或者在固定 q *[14, 15]* 的情况下，将 $|s_i|$ $(1 \leq i \leq p)$ 的 q/p 分位数作为变化阈值。相比之下，不应忽略（5）中的 ℓ_2 分量，在存在高相干性且 p 较大的情况下，它可能导致不同的稀疏模式。客观而言，IHT 的性能无法与一些标准统计方法和软件包（如 LASSO）相媲美。这正是我们开展理论分析的原因，希望借此发现并开发新的技术。

在一项理论研究中，*[17]* 在函数值方面获得了一个收敛结果，条件是

$$\vartheta > \rho_+^2(2q)/\rho_-^2(2q),$$

这改善了 *[18]* 中的条件

$$\vartheta > 32\rho_+^2(2q)/\rho_-^2(2q).$$

我们在定理 1 中的条件限制更少。例如，满足 (7) 的一个充分条件是

$$\vartheta > \{\rho_+(2q) + \eta_0\}^2/[4\{\rho_-(q+s) + \eta_0\}^2],$$

or

$$\vartheta > \{\rho_+(2q) + \eta_0\}^2/[4\{\rho_-(2q) + \eta_0\}^2)]$$

由于 $\rho_-(q+s) \geq \rho_-(2q)$，在 $\eta_0 = 0$ 的最坏情况下，这变为 $\vartheta > \rho_+^2(2q)/\{4\rho_-^2(2q)\}$。总之，$32\rho_+^2(2q)/\rho_-^2(2q) \geq \rho_+^2(2q)/\rho_-^2(2q) \geq \rho_+^2(2q)/\{4\rho_-^2(2q)\} \geq \rho_+^2(2q)/[4\{\rho_-(q+s)\}^2] \geq \{\rho_+(2q)+\eta_0\}^2/[4\{\rho_-(q+s) + \eta_0\}^2]$，而我们得到的误差率 $\sigma^2 s\log(ep/s)$ 是极小极大最优的。

感兴趣的读者也可参阅 *[19–21, 9, 17, 22],*，例如，其中分析了各种惩罚项和混合阈值规则，误差率为 $\sigma^2 s\log(ep)$。由于我们的目的是为特征数量预先确定的问题设计一种新的向后选择算法，因此不讨论其技术假设。第 IV 节中的实验对不同方法在各种情形下进行了全面比较。

B. 大规模数据改进的新方法

为预测、估计和变量选择提供可证明的保证令人安心。但真正的挑战在于，找到创新技术，在确保良好统计准确性的同时，放宽所需的正则性条件，并且比采用多次随机初始化更具成本效益。为了获得更多见解，我们可以使用受限等距数（定义见 (2)）为 (7) 提供一个充分条件：

$$\rho < 2\sqrt{\vartheta}\rho_-(q+s) + (2\sqrt{\vartheta} - 1)\eta_0 \quad \text{or} \quad 4\vartheta > \frac{(\rho+\eta_0)^2}{(\rho_-(q+s) + \eta_0)^2}. \tag{14}$$

1) “快速”学习：第 II-A 节中的结果表明，逆学习率 ρ 至关重要。在机器学习领域，训练非凸模型时通常建议采用“较慢”的学习率。这有助于获得良好的计算性能，定理 A.1 中 ρ 的下界便体现了这一点。然而，需要注意的是，根据 (14)，ρ 的取值过大可能会损害模型的统计保证。

事实上，(7) 表明更倾向于采用较小的 ρ 值，而结合统计分析与数值分析可得出 ρ 的以下取值范围：

$$\rho_+(2q) \leq \rho \leq 2\sqrt{\vartheta}\rho_-(q+s) + (2\sqrt{\vartheta} - 1)\eta_0. \tag{15}$$

在凸规划中，只要算法收敛，步长的选择就不会影响解的最优性。然而，在我们所研究的非凸约束优化中，选择足够大的 $1/\rho$ 不仅对实现快速收敛至关重要，也有助于确保统计准确性。据我们所知，这是一个新颖的发现。由于实践中可能难以确定理论上的受限等距数，因此可以通过常规线搜索来选择步长。具体而言，根据附录 A 中的证明，可以采用主化条件 $f(\beta^{(t+1)}) \leq g(\beta^{(t+1)}, \beta^{(t)})$ 或 $\|X(\beta^{(t+1)} - \beta^{(t)})\|_2^2 \leq \rho\|\beta^{(t+1)} - \beta^{(t)}\|_2^2$，以防止 ρ 变得过大，同时保持定理 A.1 中所述的收敛性质。使用随迭代变化的序列 ρ_t 这一概念将在下一节中发挥重要作用。

2）“反向”选择：另一个重要发现是基数控制的影响。如果我们采用保守的逆学习率 $\rho = \rho_+(2q)$，那么式（14）会对设计矩阵的受限条件数施加限制：

$$\vartheta > [\rho_+(2q) + \eta_0]^2/\{4[\rho_-(q+s) + \eta_0]^2)\}. \tag{16}$$

这表明 通过提高 ϑ 的值来放宽正则性条件的一种有前景的方法 .

图 1 在假设随机设计的情况下证实了这一点： q 的值越大，式 (14) 越有可能在大规模数据上成立。随机矩阵理论也支持这一观点。

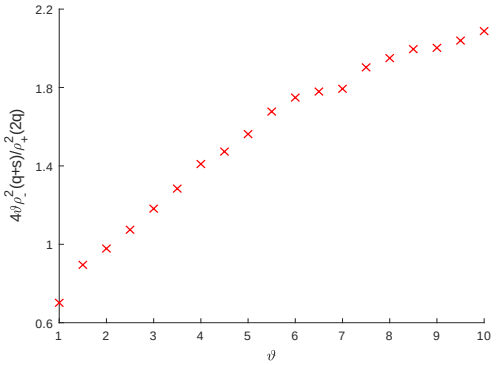


图 1：$4\vartheta\rho_-^2(q+s)/\rho_+^2(2q)$随 ϑ 增加而变化的示意图。此处， X 的各行均独立采样自均值为零、协方差为 $\Sigma = [0.5^{|i-j|}]$, $n = 2{,}000, p = 4{,}000, s = 4$的多元高斯分布。为确定给定矩阵 X对应的 $\rho_\pm$ ， 我们进行随机抽样。结果取自由相同分布生成的 100 个独立 X的平均值。

定理 3. 假设随机矩阵 $X \in \mathbb{R}^{n\times p}$ 的各行独立同分布于 $N(0, \Sigma)$，其中 $\Sigma_{ii} \le 1$。令 $\lambda_{\max}^{(2q)}$ 为所有满足 $|I| \le 2q$ 的 $I \subset [p]$ 对应的 Σ_I 的最大特征值，令 $\lambda^{(q+s)}$ 最小值 为所有满足 $|I| \le q+s$ 的 $I \subset [p]$ 对应的 Σ_I 的最小特征值。那么对于任意 $0 < c < 1$,

$$\frac{\rho_+(2q)}{\rho_-(q+s)} \le \left\{ \frac{(1+c)\sqrt{\lambda_{\max}^{(2q)}} + \sqrt{\{2\lambda_{\max}^{(2q)} q \log(ep/q)\}/n} + \sqrt{2q/n}}{(1-c)\sqrt{\lambda_{\min}^{(q+s)}} - \sqrt{\{\lambda_{\min}^{(q+s)}(q+s)\log(ep/q)\}/n} - \sqrt{(q+s)/n}} \right\}^2 \tag{17}$$

以至少为 $1 - 2\exp(-nc^2/2)$ 的概率，假设 $n \ge \{2(q+s)/(1-c)^2\}\{1/\lambda^{(q+s)}$ 最小 $+ \log(ep/q)\}$。

这些结果可以推广到次高斯设计（例如，利用 [23] 的定理6.2和 Weyl定理）。考虑满足 $0 \le \tau < 1$ 的托普利茨设计 $\Sigma = [\tau^{|i-j|}]$ 。根据交错定理,

$$(1-\tau)/(1+\tau) = \lambda_{\min}(\Sigma) \le \lambda_{\min}^{(q+s)} \le \lambda_{\max}^{(2q)} \le \lambda_{\max}(\Sigma) = (1+\tau)/(1-\tau),$$

因此，(17) 的右侧以高概率被常数 $n \gg q\log(ep/q)$ 所界定。因此，取适当大的 ϑ，即可满足正则性条件。

当然，式 (8) 中的误差界也会随着 q 值的增大而增大。为解决这一问题，我们提出采用 q_t 的递减序列，以逐步收紧基数约束。基于前面的讨论，在迭代过程中，建议使用递增的学习率 $1/\rho_t$ （例如 $1/\rho_+(2q_t)$）。调整收缩参数为序列 η_t 也可能有所帮助，尤其是在 $q_t > n$ 时。由此得到的算法结合了渐进式分位数、 ℓ_2-收缩和学习率，将被称为“slow kill”。它不同于具有固定 q 的纯优化算法 (5)，也不同于文献中常用的各种自底向上提升和贪心算法。

本节旨在为 slow kill 技术的某些方面提供有说服力的理论依据。我们将在更一般的设定下给出结果，包括迭代序列的快速收敛，以及当 q_t 趋近于 q时 slow kill 如何提高初始估计的质量，从而进一步放宽正则性条件。

III. 分位数、学习率和 ℓ_2-收缩的自适应控制

给定一般的 损失函数，基于上一节的讨论，我们通过 β 追求稀疏性

$$\min_{\beta\in\mathbb{R}^p} l_0(X\beta;y)+\frac{\eta_0}{2}\|\beta\|_2^2 \equiv l(\beta)+\frac{\eta_0}{2}\|\beta\|_2^2 \equiv f(\beta) \text{ s.t. } \|\beta\|_0\leq q, \tag{18}$$

其中，为简化记号， $l_0(X\beta;y)$通常简写为 $l(\beta)$。同样，采用混合正则化旨在解决共线性和较大的 p问题。我们假设在算法设计和理论分析中，正则化参数 q、η_0 均已给定。（当然，给定 q后，可以使用AIC等方法轻松调节 η_0 的取值；至于 q的选择，附录H中给出了一种信息准则。）对于可微的 l ，广义Bregman函数是我们用来处理各种损失函数的主要工具之一：

$$\boldsymbol{\Delta}_l(\beta_1,\beta_2) := l(\beta_1)-l(\beta_2)-\langle\nabla l(\beta_2),\beta_1-\beta_2\rangle, \tag{19}$$

其中，在统计计算 [22]中，为分析广泛的算法，可以将可微性替换为方向可微性。若 l 也是严格凸的，则 $\boldsymbol{\Delta}_l$ 变为标准的Bregman散度 [24, 25]。当 $l(\cdot)=\|\cdot\|_2^2/2$时， $\boldsymbol{\Delta}_l(\beta_1,\beta_2)=\|\beta_1-\beta_2\|_2^2/2$，它是对称的，我们将其简写为$\mathbf{D}_2(\beta_1,\beta_2)$。将 $\boldsymbol{\Delta}_l(\beta_1,\beta_2)$的对称化版本定义为 $\bar{\boldsymbol{\Delta}}_l(\beta_1,\beta_2):=\{\boldsymbol{\Delta}_l(\beta_1,\beta_2)+\grave{\boldsymbol{\Delta}}_l(\beta_1,\beta_2)\}/2$，其中 $\grave{\boldsymbol{\Delta}}_l(\beta_1,\beta_2)=\boldsymbol{\Delta}_l(\beta_2,\beta_1)$。作为式（2）的扩展，我们引入两个满足以下条件的广义受限等距数 $\rho_+^l(s_1,s_2)$、 $\rho_-^l(s_1,s_2)$：

$$\boldsymbol{\Delta}_l(\beta_1,\beta_2)\leq\rho_+^l(s_1,s_2)\mathbf{D}_2(\beta_1,\beta_2),\ \forall\beta_i:\|\beta_i\|_0\leq s_i, i=1,2 \tag{20}$$

$$\boldsymbol{\Delta}_l(\beta_1,\beta_2)\geq\rho_-^l(s_1,s_2)\mathbf{D}_2(\beta_1,\beta_2),\ \forall\beta_i:\|\beta_i\|_0\leq s_i, i=1,2. \tag{21}$$

我们区分 s_1,s_2 ，因为 $\boldsymbol{\Delta}_l$ 可能不是对称的。这些数值在理论分析中很方便且有用；例如，定理 4 和定理5将分别使用正的 $\rho_+^l(q,q)$ 和 $\rho_+^l(q,s)$，而定理 6 将使用非负的 ρ_-^l。当 $l(\beta)=\|X\beta-y\|_2^2/2$、 $\boldsymbol{\Delta}_l(\beta_1,\beta_2)=\|X\beta_1-X\beta_2\|_2^2/2$ 且$\rho_+^l(s_1,s_2)=\rho_+(s_1+s_2)$。更一般地，如果 $l_0(\cdot;y)$ 的梯度是 L-利普希茨连续的（回归或逻辑回归就是这种情况），

$$\|\nabla l_0(\xi_1;y)-\nabla l_0(\xi_2;y)\|_2\leq L\|\xi_1-\xi_2\|_2, \tag{22}$$

对所有 $\xi_1,\xi_2\in\mathbb{R}^n$，那么很容易证明

$$\rho_+^l(s_1,s_2)\leq L\rho_+(s_1+s_2)\ (\leq L\|X\|_2^2). \tag{23}$$

*A. N*通用优化算法的数值收敛性和统计准确性

首先，我们将先前的迭代分位数阈值算法扩展到可处理非二次损失函数的情况。构造如下替代函数

$$g(\beta,\beta^-)=l_0(X\beta;y)+\frac{\eta_0}{2}\|\beta\|_2^2+(\rho\mathbf{D}_2-\boldsymbol{\Delta}_l)(\beta,\beta^-), \tag{24}$$

其方法是将 损失函数（仅）线性化。随后，与第 II 节中的推导类似，(24) 得出一种算法

$$\beta^{(t+1)}=\Theta^{\#}\Big\{\beta^{(t)}-\frac{1}{\rho}X^T\nabla l_0(X\beta^{(t)};y);q,\frac{\eta_0}{\rho}\Big\}. \tag{25}$$

一些基本的数值性质总结如下。

定理 4。假设 $\inf_{\xi,y} l_0(\xi;y)>-\infty$。考虑从任意可行 $\beta^{(0)}$ 出发的 (25)。那么
$\rho\geq\rho_+^l\ q\ q$ $t\geq 0\ f\ \beta^{(t+1)}\leq g\ \beta^{(t+1)}\ \beta^{(t)}$ $\rho-\rho_+^l\ q\ q$ $_2\ \beta^{(t+1)}\ \beta^{(t)}\leq$(,) 保证对所有 , ( )
$f\ \beta^{(t)}-f\ \beta^{(t+1)}$, $t\to\infty$ $\rho>\rho_+^l\ q\ q$ $\eta_0>0$
(,) 和 ((,))D (,) ( ) ( )，因此目标函数值收敛于 。假设 (,)，且 ∇l_0 连续。那么 $\beta^{(t)}$ 的每个聚点 $\hat{}\beta$ 都满足不动点方程

$$\hat{\beta}=\Theta^{\#}\{\hat{\beta}-X^T\nabla l_0(X\hat{\beta};y)/\rho;q,\eta_0/\rho\}. \tag{26}$$

此外，若 $l_0(\cdot;y)$ 是凸函数，$\lim_{t\to\infty}\beta^{(t)}=\hat{}\beta$，且在 $\|\hat{}\beta\|_0=q$ 条件下，$\hat{}\beta$ 是问题 (18) 的一个局部极小值，并且 $\beta^{(t)}$ 的支撑集经过有限次迭代后会稳定下来。

接下来，我们讨论由 (26) 定义的估计量的统计准确性。为克服损失函数不一定与概率密度函数相关联这一障碍，我们定义相对于统计真值 β^* 的有效噪声概念如下:

$$\epsilon = -\nabla l_0(X\beta^*; y), \tag{27}$$

本节中，我们将 X 视为固定量，将 y 视为随机量。(27) 中有效噪声的定义与正则化项无关。对于累积量函数为 b 、典范连接函数为 $g = (b')^{-1}$的广义线性模型这一特殊情形，损失函数为 $l(\beta) = l_0(X\beta; y) = -\langle y, X\beta\rangle + \langle 1, b(X\beta)\rangle$，因此 $\epsilon = y - g^{-1}(X\beta^*) = y - \mathbb{E}y$。对于回归，有效噪声项 ϵ 等价于原始噪声，而原始噪声通常假设服从高斯分布。在使用逻辑偏差进行分类时， ϵ 有界，因此是次高斯的。事实上，无论 y的分布如何，任何导数有界的损失函数，例如 Huber损失、Hampel损失或合页损失，都会产生次高斯 ϵ。本节假设有效噪声是零均值的次高斯随机向量，且其尺度上界为 σ。不过，我们的证明技巧也可应用于更一般的情形。以下定理给出了由 (25) 得到的估计量的风险界，并展示了起始点的质量对正则性条件的影响。

定理5。设 $\hat{}\ \beta$: $\|\hat{}\beta\|_0 = q$是利用 (25) 得到的估计量，其起始点 $\beta^{(0)}$ 是可行的，即 $\hat{}\beta \in \min_{\|\beta\|_0 \le q} g(\beta, \hat{}\beta)$ 且 $f(\hat{}\beta) \le f(\beta^{(0)})$，其中 $\|\beta^{(0)}\|_0 \le q_\circ$ 定义

$$P_o(q) = q\log(ep/q). \tag{28}$$

假设 $\beta^{(0)}$ 满足

$$\mathbb{E}\mathbf{D}_2(\beta^{(0)}, \beta^*) = \mathcal{O}(M)\frac{\sigma^2 P_o(q) + \sigma^2}{n} \text{ for some } M: 1 \le M \le +\infty. \tag{29}$$

令 $Q = \{\rho_+(q+s)M/n\}^{1/2} + \{\rho_+^l(q,s) + \eta_0\}M/n_\circ$ 假设对于某个 $\delta > 0$、$0 < \varepsilon \le 1$ 以及足够大的 $K \ge 0$,

$$\begin{aligned} &K\sigma^2 P_o(\vartheta s) + \Big\{2(1 - \frac{1}{M})\bar{\mathbf{\Delta}}_{l_0} + \frac{C}{M(Q\delta \vee 1)}\mathbf{\Delta}_{l_0} - \delta\mathbf{D}_2\Big\}(X\beta, X\beta') \\ &\ge \frac{1 - 1/M}{\sqrt{\vartheta}}\big[\rho - \{(2-\varepsilon)\sqrt{\vartheta} - 1\}\eta_0\big]\mathbf{D}_2(\beta, \beta'), \forall \beta, \beta': \|\beta\|_0 \le \vartheta s, \|\beta'\|_0 \le s, \end{aligned} \tag{30}$$

其中 C 是某个正常数。那么

$$\mathbb{E}\big\{\mathbf{D}_2(X\hat{\beta}, X\beta^*) \vee \frac{\eta_0\varepsilon}{\delta}\mathbf{D}_2(\hat{\beta}, \beta^*)\big\} \lesssim \frac{K\delta \vee 1}{\delta^2}\Big\{\sigma^2\vartheta s\log\big(\frac{ep}{\vartheta s}\big) + \sigma^2\Big\} + \frac{\eta_0}{\delta\varepsilon}\|\beta^*\|_2^2. \tag{31}$$

因此，只要 K、δ、ϑ 是常数，且 η_0 不是过大，我们就能达到所需的统计准确性。当 $M = +\infty$ 时（对$\beta^{(0)}$没有要求），正则性条件 (30) 变为

$$\frac{\rho - \{(2-\varepsilon)\sqrt{\vartheta} - 1\}\eta_0}{\sqrt{\vartheta}}\mathbf{D}_2(\beta, \beta') \le \big(2\bar{\mathbf{\Delta}}_{l_0} - \delta\mathbf{D}_2\big)(X\beta, X\beta') + K\sigma^2 P_o(\vartheta s), \forall \beta, \beta': \|\beta\|_0 \le \vartheta s, \|\beta'\|_0 \le s,$$

其中包括作为特殊情形的（7）。但是，当使用一个合适的起始点时，（30）会宽松得多。在 $M = 1$ 的极端情形下，（30）的右端变为 0，因此，对于 $\|\beta\|_0 \le \vartheta s$、$\|\beta'\|_0 \le s$，具有 μ-受限强凸性$^-$（$\mathbf{\Delta}_{l_0} - \mu\mathbf{D}_2$）（$X\beta$、$X\beta'$）的 ≥ 0 总能满足（30）。

B. Slow kill：算法设计与序贯分析

对于大规模数据，使用多起点策略为 $\beta^{(0)}$ 选择高质量初始值在计算上可能不可行。幸运的是，我们将看到，设计迭代变化阈值化和收缩能够有效放宽统计正则条件，并提高迭代序列的统计准确性。

更具体地说，slow kill 通过引入三个辅助序列ρ_{t+1}、q_{t+1}、η_{t+1}来修改优化算法 (25)。

$$\beta^{(t+1)} = \Theta^{\#}\Big\{\beta^{(t)} - \rho_{t+1}^{-1}X^T\nabla l_0(X\beta^{(t)}; y); q_{t+1}, \bar{\eta}_{t+1}\Big\}, \text{ with } \bar{\eta}_{t+1} = \eta_{t+1}/\rho_{t+1} \tag{32}$$

其中 $q_t \to q$, $\eta_t \to \eta_{0\circ}$ 在后续分析中，使用缩放收缩序列 $\bar{\eta}_t$ 会比使用原始序列η_t 更加方便。我们希望了解，在迭代过程中调整逆学习率、基数和 ℓ_2-收缩参数,

是否能够提升性能。具体而言，我们旨在研究随着 t 增加， $\beta^{(t)}$ 的统计准确性如何变化，以及统计误差在什么条件下会以几何速度快速收敛。定理 6 关注的是 $\beta^{(t)}$ 相对于统计真值 β^*的统计误差，而非它们相对于特定极小化器 β^o的优化误差。我们将看到，原则上，slow kill 会受益于降低 q_t 和 ρ_t。还值得注意的是，式 (35) 中的误差界对 $\vartheta_t, \rho_t, \eta_t$均无要求。

定理 6. 令迭代序列 $\beta^{(t)}$： $\|\beta^{(t)}\|_0 = q_t$ 由式（32）生成，并具有可行的 $\beta^{(0)}$。给定任意$t \geq 1$，定义

$$h_t^{-1} = (1 - 1/\sqrt{\vartheta_t})(\rho_t + \eta_t) + (1 - \varepsilon)(\rho_-^l(q_t, s) + \eta_t), \tag{33}$$

$$\kappa_t = (\rho_t - \rho_-^l(s, q_t))h_t, \tag{34}$$

其中， ε 是区间（$0, 1$]中的任意数。则以下递归统计误差界

$$\begin{aligned}
&\mathbf{D}_2(\beta^*, \beta^{(T+1)}) + \sum_{t=0}^{T} \big(\Pi_{\tau=t}^{T} h_{\tau+1}\big)(\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta^{(t+1)}, \beta^{(t)}) \\
&\leq \sum_{t=0}^{T} \big(\kappa_{t+1} \cdots \kappa_{T+1}\big) \Bigg\{ \frac{A\sigma^2}{\varepsilon} \frac{\rho_+(q_{t+1} + s)}{\big(\frac{\rho_-^l(q_{t+1}, s)}{\rho_{t+1}} \vee \bar{\eta}_{t+1}\big)\big(1 - \frac{\rho_-^l(s, q_{t+1})}{\rho_{t+1}}\big)\rho_{t+1}^2} \cdot \vartheta_{t+1} s \log\big(\frac{ep}{\vartheta_{t+1} s}\big) \\
&\qquad + \frac{\bar{\eta}_{t+1}}{\big(1 - \frac{\rho_-^l(s, q_{t+1})}{\rho_{t+1}}\big)\varepsilon} \|\beta^*\|_2^2 \Bigg\} + \Big(\Pi_{t=0}^{T} \kappa_{t+1}\Big) \mathbf{D}_2(\beta^*, \beta^{(0)}).
\end{aligned} \tag{35}$$

成立　对所有 $T \geq 0$均成立，其概率至少为 $1 - Cp^{-cA}$，其中 C、c 为正的常数。

下面的推论展示了该定理在算法配置中的实用性。

推论 1. 在定理 6 的设定下，给定任意 $\varepsilon \in (0, 1]$, ，若 ρ_t 和 η_t 的选取满足

$$\rho_{t+1} \geq \rho_+^l(q_{t+1}, q_t) \tag{36}$$

$$\bar{\eta}_t \geq 0 \vee \frac{(1/\sqrt{\vartheta_t} + \varepsilon) - 2(\rho_-^l(s, q_t) \wedge \rho_-^l(q_t, s))/\rho_t}{2 - 1/\sqrt{\vartheta_t} - \varepsilon} \tag{37}$$

使得 $(\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta^{(t+1)}, \beta^{(t)}) \geq 0$ 且 $\kappa_t \leq (1+\varepsilon)^{-1}$，则以至少 $1 - Cp^{-cA}$ 的概率， $\{\beta^{(t)}\}$ 的统计误差以几何速度衰减，

$$\mathbf{D}_2(\beta^*, \beta^{(T+1)}) \leq \left(\frac{1}{1+\varepsilon}\right)^{T+1} \mathbf{D}_2(\beta^*, \beta^{(0)}) + \frac{1}{\varepsilon} \sum_{t=0}^{T} \left(\frac{1}{1+\varepsilon}\right)^{T-t+1} E_{t+1} \tag{38}$$

对于所有 $T \geq 0$，其中

$$E_{t+1} = \big\{1 - \frac{\rho_-^l(s, q_{t+1})}{\rho_+^l(q_{t+1}, q_t)}\big\}^{-1} \Bigg\{ \frac{A\sigma^2}{\frac{\rho_-^l(q_{t+1}, s)}{\rho_+^l(q_{t+1}, q_t)} \vee \bar{\eta}_{t+1}} \frac{\rho_+(q_{t+1} + s)}{(\rho_+^l(q_{t+1}, q_t))^2} \vartheta_{t+1} s \log\big(\frac{ep}{\vartheta_{t+1} s}\big) + \bar{\eta}_{t+1} \|\beta^*\|_2^2 \Bigg\}. \tag{39}$$

理论结果为 slow kill 的三个主要要素的设计提供了宝贵的见解。我们首先应用定理 6，分析采用固定分位数 $q_t \equiv q$ 和通用值$\rho_t \equiv \rho$、 $\bar{\eta}_t \equiv \bar{\eta}$的基本优化算法。(38) 随后表明，统计误差线性收敛，右侧的第一项表示初始点的影响。由于 $\Sigma_{t=0}^{T}\{1/(1+\varepsilon)\}^{T-t+1} \leq 1/\varepsilon$，最终误差的量级为

$$\frac{\rho_+(q+s)}{(\rho_+^l(q,q))^2} \sigma^2 \vartheta s \log\big(\frac{ep}{\vartheta s}\big) + \bar{\eta}\|\beta^*\|_2^2, \tag{40}$$

其中，受限条件数 $\rho_+^l(q,q)/\{\rho_-^l(s,q) \wedge \rho_-^l(q,s)\}$ 以及 ε均假设为常数。(37) 中推导出的下界有助于减小偏差，并表明在这方面使用较大分位数的益处。

另一方面，较大的分位数可能导致（39）中的方差项 $\vartheta_{t+1} s \log\{ep/(\vartheta_{t+1} s)\}$ 膨胀，这促使我们采用递减分位数，这是 slow kill 最显著的特点。事实上，更仔细地考察（38）可以发现，因子 $1/(1+\varepsilon)^{T-t+1}$ 允许在较早的迭代步骤中使用大得多的 q_t。这是

因为对于较小的 t，相关误差 E_{t+1} 将在最终界中被更大幅度地收缩。尽管理论上可能难以推导出序列 q_t的最优冷却方案，但各种方案在实践中似乎都表现良好，例如 $q_{t+1}=\lfloor q+(T-t)/(aTt+bT)\rfloor$ (反函数) 或 $\lfloor q+(p-q)/1+a\exp(bt/T)^c\rfloor$ (S形) 等。

给定 q_t 后，可以根据 (36) 理论确定 ρ_{t+1} 的取值： $\rho_{t+1}\geq\rho_+^l(q_{t+1},q_t)$，这给出了步长的上界，以防止 slow kill 发散。在实际实现中， $\rho_+^l(q_{t+1},q_t)$ 通常未知。对于正则设计矩阵（例如托普利茨矩阵），假设 ∇l_0 是 L-利普希茨连续的，可以根据附录 D 的证明中的 (A.17)，采用 $L\{n+q_{t+1}\log(ep/q_{t+1})\}$ 的常数倍。更一般地，从 (35) 左侧的第二项可以看出，我们可以使用满足以下准则的线搜索

$$(\rho_{t+1}\mathbf{D}_2-\mathbf{\Delta}_l)(\beta^{(t+1)},\beta^{(t)})\geq 0. \tag{41}$$

有关线搜索的一些实现细节，请参见附录 I。(41) 在 ($\beta^{(t+1)}$, $\beta^{(t)}$) 处强制满足主化条件，因此得到的 ρ_{t+1} 甚至可能小于 $\rho_+^l(q_{t+1},q_t)$。此前已在第 II-B 节讨论过，对于 ℓ_0-约束回归，限制 ρ_t 的大小十分重要。同样，较小的 ρ_{t+1} 有助于实现较大的 ε，进而带来更快的收敛速度和更小的误差，如 (33) 和 (37) 所示。

较低　界可改写为：Corollary 1 中缩放 ℓ_2-收缩序列 $\bar{\eta}_t$ 的下界

$$2\sqrt{\vartheta}_t>\frac{\rho_{t+1}+\bar{\eta}_t\rho_t}{\rho_-^l(s,q_t)\wedge\rho_-^l(q_t,s)+\bar{\eta}_t\rho_t}. \tag{42}$$

它类似于受限条件数条件，并将 (16) 推广到了一般损失函数。具体而言，当$2q_t>n$时，(37) 意味着 $\bar{\eta}_t>(1/\sqrt{\vartheta_t})/(2-1/\sqrt{\vartheta_t})=1/(2\sqrt{\vartheta_t}-1)$，因此我们建议使用如下定义的缩放收缩序列：

$$\bar{\eta}_t=1/(2\sqrt{q_t/\bar{s}}-1), \tag{43}$$

其中 $\bar{s}=q\wedge nL^2/\log(ep)$（根据附录 F，这是 s的替代量），而 L 是 ∇l_0的利普希茨参数。(43) 在 slow kill 的早期迭代中发挥重要作用，且与学习率无关。

我们的分析支持使用 ℓ_2辅助的后向分位数控制来逐步收紧约束。slow kill 中采用的更新公式 (32) 有坚实的优化理论基础，因此优于基于启发式的多阶段过程。定理 6 中确立的快速几何收敛，结合较强的信号强度，表明 $\beta^{(t)}$ 中的零值以很高的概率对应无关预测变量（参见注 1 和附录 G）。因此，我们可以偶尔使用 $\mathcal{J}(\beta^{(t+1)})$压缩设计矩阵（例如，当 q_{t+1} 达到 $p/2^k$时），以缩小问题规模（附录 I）。这样便能在早期移除明显的无用特征，节省计算成本；而对于较难识别的无关特征，则只需在接近找到最优解时再加以处理。这一特点使 slow kill 尤其适用于大数据学习。slow kill 在组变量选择 [11] 和低秩矩阵估计 [26]中也具有类似优势。

相较之下，前向逐步算法和提升算法 [5, 27, 6–8, 19, 9] 以自底向上的方式从零模型开始逐步构建模型。这类算法在每次迭代时都必须考虑几乎所有特征，因此计算量很大，因为它们往往需要数百或数千次提升迭代。受 ℓ_0 优化视角的启发，我们还可以研究一类“稳步增长”过程，其中 q_t 在式 (32) 中从 0 增至 q 。与提升法相比，更新和选择除了梯度外，还会纳入先前估计的影响。在稳步增长过程中可以引入一种保留选项，其作用与 slow kill 中的压缩操作相反。保留与压缩的研究，以及 slow kill 与稳步增长的结合，将留待未来研究。

最后，如何获得具有指定基数的稀疏模型，是本文始终关注的问题。不过，如果要确定 q 的最佳取值，我们建议使用预测信息准则 [28]，它能够以非渐近的方式保证最优预测误差率（详见附录 H）。

IV. 实验

A. 模拟

在本部分中，我们通过模拟研究比较 slow kill（在下文的表格和图中简称为 SK）与一些常用稀疏学习方法在预测准确率、选择方面的表现。

一致性和计算效率，将 slow kill（在下文的表格和图中简称为 SK）与一些常用的稀疏学习方法进行比较。除非另有说明，预测变量矩阵 $X = [\tilde{x}_1, \ldots, \tilde{x}_n]^T \in \mathbb{R}^{n \times p}$ 的各行 $\tilde{x}_i^T$ 均独立生成，服从协方差矩阵为 Σ 的多元正态分布，其中 Σ 具有托普利茨结构 $[\tau^{|i-j|}]$ 或等相关结构 $[\tau 1_{i \neq j}]$。实验中将包含较高的相关强度，例如 $\tau = 0.9$。我们同时考虑稀疏 β^* 下的回归和分类：$\beta_j^* = 1$，若 $j = 10k+1$，则 $0 \leq k < s$，因此 $s = \|\beta^*\|_0$。在回归实验中，$y = X\beta^* + \epsilon$ 服从 $\epsilon_i \sim N(0, 1)$；在分类实验中，若 $\tilde{x}_i^T \beta^* > 0$，则 $y_i = 1$，否则为 0。

除 slow kill 外，比较中还包括以下方法：LASSO [29], 弹性网（ENET） [12],MCP [4], SCAD [30], 以及 IHT 和 NIHT（[15, 31], 仅用于回归）。（我们还在模拟中评估了 picasso [9] 的性能，它是 [19]的改进版本。然而，其逐路径计算在合成数据上的误差率和缺失率都高于标准非凸优化。因此，我们未展示相关结果。后续章节将把该算法纳入真实数据实验。）回归采用二次损失，分类采用逻辑偏差。对于 slow kill，我们采用一个简单的单一起始点 $\beta^{(0)} = 0$ 和 $\eta_0 = 50$；使用反向冷却计划 $q_{t+1} = \lfloor q + (T-t)/\{tT/(p-q) + 2T/(p-2q)\} \rfloor (0 \leq t \leq T)$，使得 $q_T = q$ 和 $q_1 = p/2$；为方便和提高效率，我们在所有实验中均设定 $T = 100$。我们使用 R 包 glmnet 实现 LASSO 和弹性网，使用 ncpen 包 [32] 实现上述非凸惩罚项，并使用 sparsify 包实现 IHT 方法。（glmnet 的核心使用 Fortran 子程序实现，而 ncpen 主要基于 C++。通过使用 C 或 Fortran，我们的 slow kill 实现有望变得更高效、占用更少内存；但正如后文表格和图示所示，它的表现已经与其他方法相当或更好。）为确保比较公平，并消除不同参数调优方案的影响，我们为每种方法选取非零系数数目为 1.5s 的估计结果。为校准偏差，我们仅使用选定变量重新拟合每个得到的模型。所有其他算法参数均设为默认值。

对于每种模拟设定，我们重复实验 50 次，并根据以下指标评估各算法的性能：漏检率 $\times 100\%$ 和预测误差。具体而言，漏检率是未检测到的真实变量所占的比例；在回归中，预测误差使用真实信号，通过 10 倍 $(\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*)$ 计算；在分类中，预测误差指的是在一个独立测试集上的误分类错误率 $\times 100\%$，该测试集包含与训练数据集相同数量的观测值。此外，还纳入总计算时间（以秒为单位）来描述计算成本。由于惩罚方法的实现通常采用热启动，因此当算法得到具有指定基数的估计结果时，我们便终止算法。

表 I 展示了回归设定下的一些实验结果。图 2 绘制了若干代表性方法在改变稀疏度 s 和相关强度 τ 时的更多结果（不包括弹性网和 IHT，因为它们的表现分别与 LASSO 相近和较差）。可以看出，在统计准确性和计算时间两方面，slow kill 都优于其他方法，尤其是在更具挑战性的情形下，例如相关特征更多、设计更相干时。

表 I：回归：采用不同相关结构时，在预测误差、漏检率和计算时间方面的性能比较。更具体地说，$p = 5{,}000, n = 150, s = 10$ 和 $\Sigma = [\tau^{|i-j|}]$ 或 $[\tau 1_{i \neq j}]$，其中 $\tau = 0.9$

	托普利兹结构			等相关		
	错误	Miss	Time	错误	Miss	Time
LASSO	16	32	5	15	83	13
ENET	16	31	13	14	82	34
IHT	85	68	55	16	88	57
NIHT	12	22	4	17	80	18
MCP	12	23	34	18	78	24
SCAD	12	23	13	16	85	6
SK	2	2	1	12	50	1

对于分类，表 II 和图 3 比较了具有不同相关结构和问题维度的各种方法，可以得出类似的结论。值得注意的是，与其他方法相比，slow kill 出色的统计准确性并未以牺牲计算时间为代价。事实上，如图 3 所示，slow kill 能够大幅节省时间，尤其是在 n 较大时，同时在变量选择和预测方面表现出色。

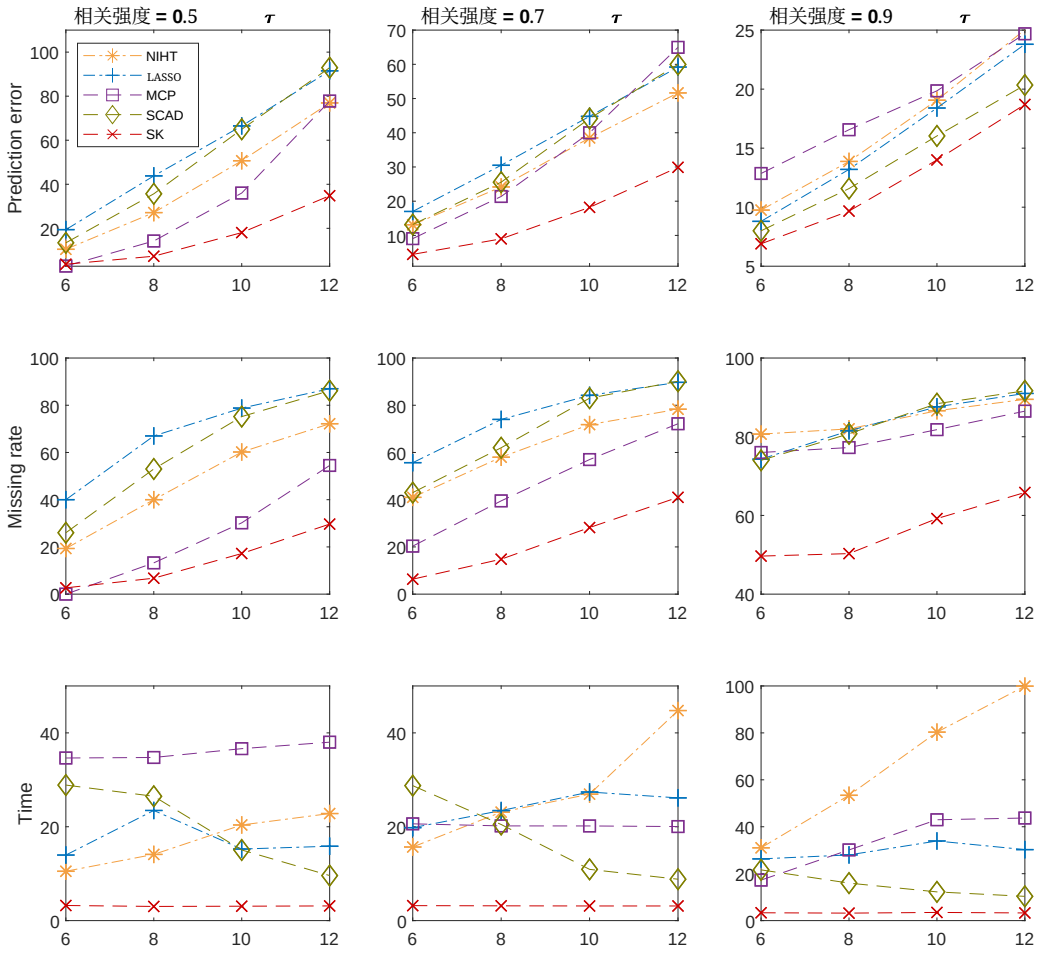


图 2：回归：改变模型的稀疏性和相关强度时，在预测误差、漏检率和计算时间方面的性能比较。具体而言，$p = 10{,}000, n = 150, s = 6$, 8、10、12 和 $\Sigma = [\tau 1_{i\neq j}]$，其中 $\tau = 0.5$、0.7、0.9。

表 II：分类：在不同相关结构下，从预测误差、漏检率和计算时间方面进行性能比较。更详细地说，$p = 2{,}000, n = 500, s = 10$ 和 $\Sigma = [\tau^{|i-j|}]$ 或 $[\tau 1_{i\neq j}]$ 与 $\tau = 0.9$

	托普利兹结构			等相关		
	错误	Miss	Time	错误	Miss	Time
LASSO	8.0	24	10	5.1	95	49
ENET	8.0	25	31	4.7	95	135
MCP	6.9	23	15	5.0	93	20
SCAD	7.0	22	22	5.1	94	16
SK	2.2	2	4	3.9	78	4

接下来，我们展示一些信号强度变化时的实验。回顾一下，在回归设定中，我们对 $\beta_j^* = 1$（β^*）设置为 $j \in \mathcal{J}$。对于 $n = 100$、$p = 5000$、$\sigma = 1$，极小极大最优速率约为$\sigma\sqrt{(\log p)/n}$（≈ 0.292）（忽略常数因子，因为其精确值可能难以推导）。我们通过设置 $\beta_j^* = 0.8$、0.6、0.4、0.2 进行了额外实验。不同方法的比较结果如图 4 所示。当信号强度较低时（例如 $\beta_j^* = 0.2$、0.4），所有方法的表现都很差。信号强度较高时，slow kill 的表现则远胜其他方法。

我们进行了另一项实验，以探索更大的 $\|\beta^*\|_2^2$ 值。（提醒一下，在先前的设置中，where$s = 10$ and $\beta_j^* = 1, \forall j \in \mathcal{J}(\beta^*)$, 我们有 $\|\beta^*\|_2^2 = 10$。）我们通过按相应比例放大每个 β_j^*，测试了 $\|\beta^*\|_2^2 = 50$、100、150、200。该实验的结果如图 5 所示。随着 $\|\beta^*\|_2^2$ 增大，NIHT、MCP 和 slow kill 展现出明显优势，其中后两者的预测误差和缺失率相近。

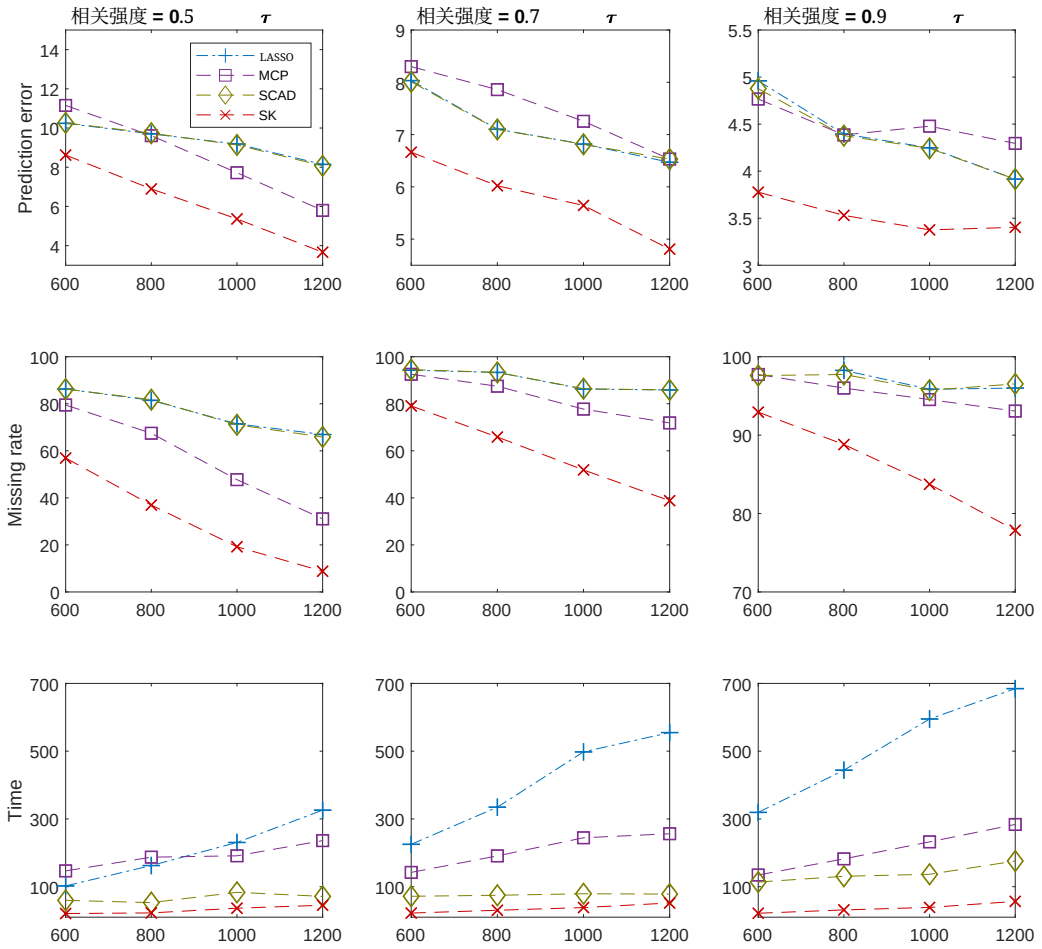


图3：分类：在不同相关结构和样本量下，从预测误差、漏检率和计算时间三个方面进行性能比较。具体而言，$p = 10{,}000$、$n = 600$、800、1000、1200、$s = 15$ 以及 $[\tau 1_{i\neq j}]$，相关强度为 $\tau = 0.5$、0.7、0.9。

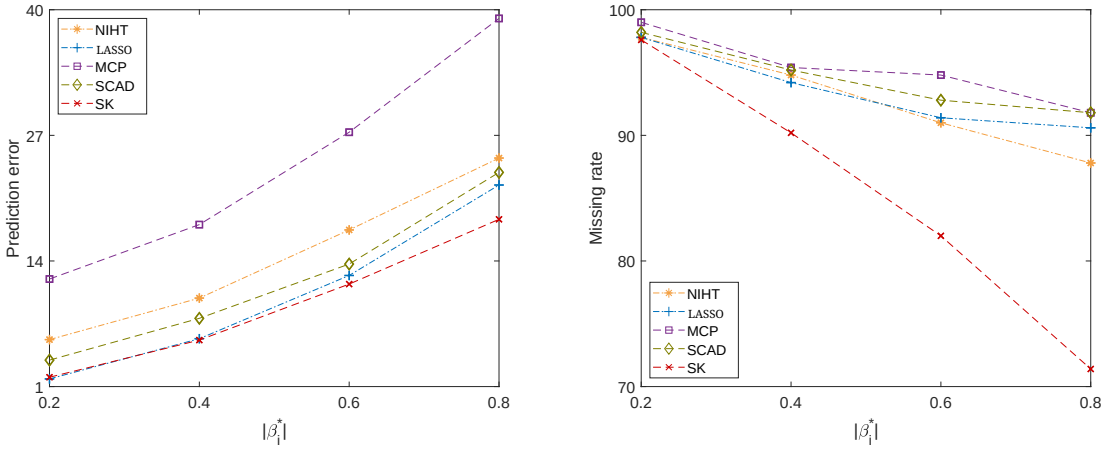


图 4：不同信号强度下各方法的预测误差（左）和缺失率（右）对比。回归设定的详情见第 IV-A 节，我们将 $p = 5{,}000, n = 100, s = 10,$ $\tau = 0.8$ 和 $\beta_j^* = 0.2$ 设为 0.4、0.6、0.8，用于 $j \in \mathcal{J}(\beta^*)$。

B. 手写数字分类

Gisette 数据集 [33] 旨在对手写数字识别中容易混淆的数字 4 和 9 进行分类。该数据集包含 5,000 个预测变量，包括根据像素构造的各种特征，以及一些预测能力较弱的“探针”。由于相关特征的确切数量未知，我们在相同模型基数的条件下评估不同方法的性能，以进行公平比较。我们随机划分这 7,000

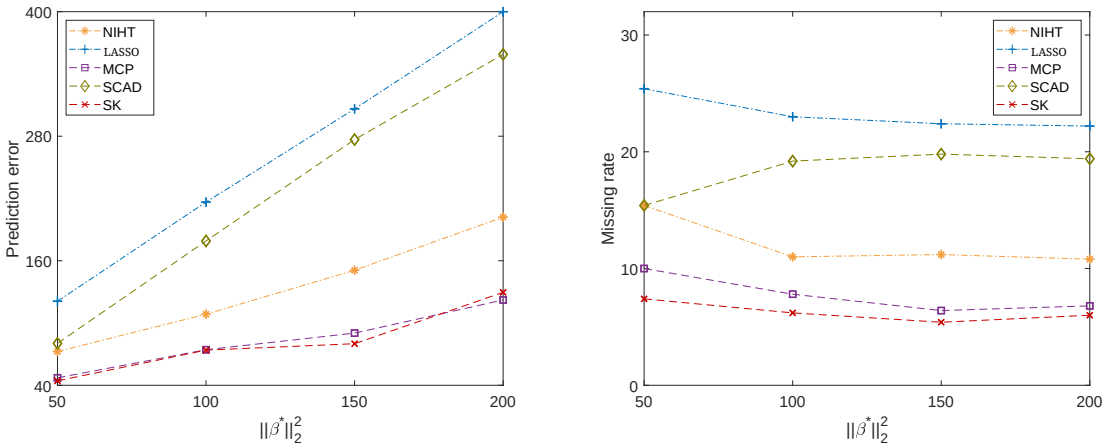


图 5：大信号下不同方法的预测误差（左）和缺失率（右）对比。回归设定的详情见第 IV-A 节，我们将 $p = 5{,}000$、$n = 100$、$s = 10$、$\tau = 0.8$ 和 $\|\beta^*\|_2^2 = 50$,100、150、200（通过逐步放大每个 β_j^*）设为.....

重复 20 次，将样本划分为包含 3,000 个样本的训练子集和包含 4,000 个样本的测试子集，并报告平均误分类错误率和总计算时间。

由于数据规模相对较大，计算效率是一个主要问题。许多统计软件包无法在合理的时间内给出有意义的结果。这里，我们比较 glmnet [34],logitboost [35, 36], picasso（采用 MCP 选项 [37], ）和具有不同选定特征数的 slow kill。

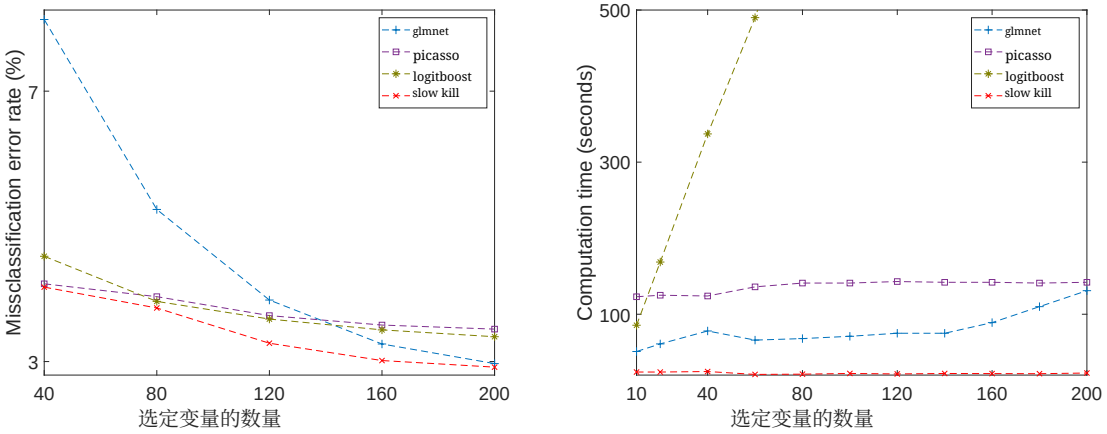


图6：Gisette 数据。左图：误分类错误率均值；右图：总计算时间，展示不同数量的选定特征。与其他方法相比，picasso 的成本过高，因此仅展示其成本曲线的一部分。

根据图6，logitboost 和 picasso 在该数据集上的误分类错误率优于 glmnet，但 slow kill 始终表现最佳。在计算成本方面，glmnet 和 slow kill 具有极强的可扩展性；即使仅使用 $q = 40$，logitboost 的成本也相当高，而 picasso 在$q \geq 60$时也存在类似问题。

C. 乳腺癌微阵列数据

来自精选微阵列数据库的乳腺癌微阵列数据集 [38] 包含 35,981 个基因表达水平数据，涵盖 143 名乳腺癌患者的肿瘤样本以及 146 份配对的邻近正常乳腺组织样本。其目标是识别一些差异表达基因，以帮助对正常组织和肿瘤组织进行分类。我们将该数据集随机划分为训练子集（60%）和测试子集（40%），重复 20 次，并在表III中报告了不同方法的误分类错误率和总计算时间。

根据表 III，logitboost 的计算复杂度最高，而 picasso 在该数据集上的整体分类表现最差。相比之下，glmnet 和 slow kill 的误分类错误率更低；根据我们的实验，后者的性价比要高得多。

表III：乳腺癌微阵列数据：误分类错误率（×100%）和总计算时间（秒）

	$q = 60$		$q = 80$		$q = 100$		$q = 120$		$q = 140$	
	错误	Time	错误	Time	错误	Time	错误	Time	错误	Time
GLMNET	10.9	19	10.7	19	10.5	50	10.2	50	10.2	50
PICASSO	11.4	43	11.3	43	11.1	48	11.3	48	11.2	42
logitboost	11.2	500	11.2	680	10.9	860	10.6	1080	10.8	1220
SK	10.8	10	10.2	11	10.2	11	10.1	11	9.8	11

D. 亚奈奎斯特频谱感知与学习

基于亚奈奎斯特采样的毫米波宽带频谱感知，是下一代无线通信系统中的一个重要课题。采用多余集采样器 [39], ，可以在信号处理中将多测量向量模型表述为 $Y = XB^* + \mathcal{E}$，其目标是利用 B^* 的联合（逐行）弱稀疏性来重构频谱。此处所有矩阵均为复数矩阵（例如 $Y \in \mathbb{C}^{n\times m}$、 $X \in \mathbb{C}^{n\times p}$），预测变量矩阵 X 的大小由余集数量和信道数量决定；详情请参阅 [40] 。值得注意的是，使用厄米特内积 $\langle A 、 B\rangle$，以 tr $\{A^{\mathrm{H}}B\}$ 替代实内积，并将广义Bregman函数重新定义为 $\mathbf{\Delta}_l(B_1, B_2) = l(B_1) - l(B_2) - \langle \nabla l(B_2)\ B_1 - B_2\rangle/2 - \langle B_1 - B_2, \nabla l(B_2)\rangle/2$ 后，我们所有的定理和算法都可以扩展到复数群稀疏追踪。

我们将我们的方法与两种流行的方法 SOMP [41] 和 JB-HTP [42], 在 [43]中的基准时域ˆ数据集上进行了比较。表 IV 展示了每种方法的归一化均方误差 $\|B - B^*\|_F/\|B^*\|_F$，其中我们改变了 q（所选信道的数量）。图 7 展示了频谱恢复示例，其中使用附录 H 中的预测信息准则在 slow kill 中进行模型选择。

表 IV：以归一化均方误差表示的频谱重构误差

	$q = 3$	$q = 4$	$q = 5$	$q = 6$	$q = 7$	$q = 8$
SOMP	0.83	0.93	0.82	0.91	0.92	0.94
JB-HTP	0.94	1.00	0.99	0.95	1.07	0.96
SK	0.74	0.65	0.53	0.38	0.42	0.50

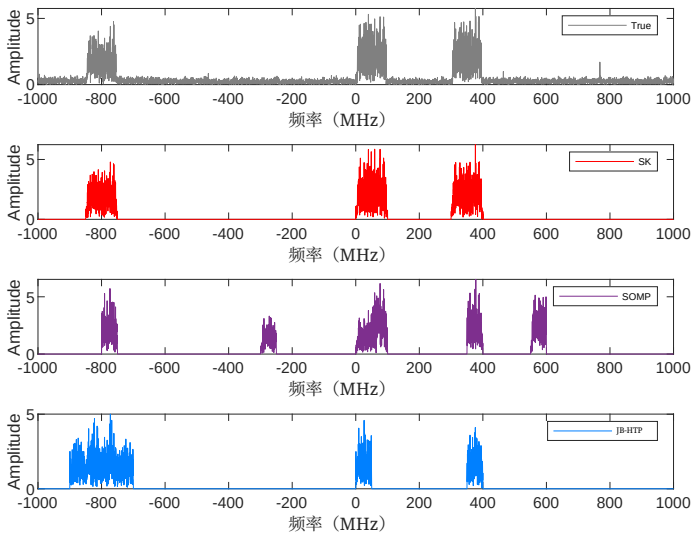


图 7：不同方法的频谱感知结果。

五、总结

本文提出了一种用于大规模变量选择的新型 slow kill 方法。这是一种可扩展的优化型算法，采用三组经过精心设计且具有理论依据的序列：阈值序列、收缩序列和学习率序列。

直观而言，slow kill 采用一种新颖的后向分位数控制，结合自适应的 ℓ_2 收缩和递增的学习率，以放宽正则性条件并克服后向消除中的障碍。该方法与现有文献中的提升法及许多前向逐步法显著不同。我们的理论研究揭示了如何设计渐进式混合正则化，以达到最优误差率并实现快速收敛。该技术适用于一般损失函数，不一定是负对数似然函数；而且，它能够在整个迭代过程中缩减问题规模，因此适用于大数据。

附录

子高斯变量的定义　　高斯分布随机变量或次高斯随机向量在文献中是标准的。　　.

定义 A.1. 如果 ξ 具有零均值，且 ξ的尺度（ψ_2-范数），定义为 $\inf\{\sigma>0:\ \mathbb{E}[\exp(\xi^2/\sigma^2)]\le 2\}$，有限，则称其为次高斯随机变量。如果 $\langle \xi,\alpha\rangle$ 的所有一维边缘分布都是满足 $\|\langle \xi,\alpha\rangle\|_{\psi_2}\le \sigma\|\alpha\|_2$的次高斯分布，其中尺度不超过 σ，则称 $\xi\in\mathbb{R}^p$ 为次高斯随机向量。类似地，如果 vec (ξ) 是次高斯的，则称随机矩阵 ξ 为次高斯。

A. 定理 A.1 和定理 4

首先，对于在第 II 节的设定中定义的算法 (5)，我们有以下数值性质。

定理 A.1. 对于任意 X、y 和 $\beta^{(0)}$，由 (5) 生成的迭代序列 $\beta^{(t)}$ 满足 $f(\beta^{(t)})-f(\beta^{(t+1)})\ge \rho\|\beta^{(t+1)}-\beta^{(t)}\|_2^2/2-\|X(\beta^{(t+1)}-\beta^{(t)})\|_2^2/2,\quad \forall t\ge 0$ 因此当 $\rho\ge\rho_+(2q)$ 时，$f(\beta^{(t)})$ 收敛，且 $\beta^{(t)}$ 满足

$$\min_{0\le t\le T}\|\beta^{(t+1)}-\beta^{(t)}\|_2^2\le\frac{1}{T+1}\frac{2f(\beta^{(0)})}{\rho-\rho_+(2q)}.$$

此外，只要 $\rho>\rho_+(2q)$ 且 $\eta_0>0$，$\beta^{(t)}$ 就有唯一的极限点 $\hat{}\ \beta$，该极限点满足“不动点”方程

$$\beta=\Theta^{\#}\{\beta-X^T(X\beta-y)/\rho;q,\eta_0/\rho\},$$

$\hat{}\ \hat{}$ 且当 $\|\beta\|_0=q$ 时，β 也是问题 (3) 的一个局部极小值。

为证明定理 A.1 的第一个结论，请注意，在回归情形下，

$$g(\beta^{(t+1)},\beta^{(t)})-f(\beta^{(t+1)})=\rho\|\beta^{(t+1)}-\beta^{(t)}\|_2^2/2-\|X(\beta^{(t+1)}-\beta^{(t)})\|_2^2/2,$$

因此

$$f(\beta^{(t)})-f(\beta^{(t+1)})\ge\frac{\rho}{2}\|\beta^{(t+1)}-\beta^{(t)}\|_2^2-\frac{1}{2}\|X(\beta^{(t+1)}-\beta^{(t)})\|_2^2,\quad\forall t\ge 0.$$

对 $t=0$ 到 $t=T$ 求和，并利用 $\|X(\beta^{(t+1)}-\beta^{(t)})\|_2^2\le\rho_+(2q)\|\beta^{(t+1)}-\beta^{(t)}\|_2^2$ 这一事实，可得

$$\frac{(\rho-\rho_+(2q))}{2}\sum_{t=0}^{T}\|\beta^{(t+1)}-\beta^{(t)}\|_2^2\le f(\beta^{(0)})-f(\beta^{(T+1)}),$$

这导致

$$\min_{0\le t\le T}\|\beta^{(t+1)}-\beta^{(t)}\|_2^2\le\frac{2}{(T+1)(\rho-\rho_+(2q))}f(\beta^{(0)}).$$

接下来，我们考虑一般问题并证明定理 4，这意味着定理 A.1 的第二部分。由 $\inf_{\xi,\ y} l_0(\xi;y)>-\infty$，不失一般性地假设 $l_0(\xi;y)\ge 0$。回顾一下，$l_0(X\beta;y)$ 缩写为 $l(\beta)$，因此根据链式法则，$\nabla l(\beta)=X^T\nabla l_0(X\beta)$。

由构造 $g(\beta,\beta^{(t)})=f(\beta)+(\rho\mathrm{D}_2-\mathbf{\Delta}_{l_0})(\beta,\beta^{(t)})$，我们得到

$$(\rho\mathbf{D}_2-\mathbf{\Delta}_{l_0})(\beta^{(t+1)},\beta^{(t)})+f(\beta^{(t+1)})\le g(\beta^{(t)},\beta^{(t)})=f(\beta^{(t)}).$$

当 $\rho\ge\rho_+^l(q,q)$ 时，$(\rho\mathrm{D}_2-\mathbf{\Delta}_{l_0})(\beta^{(t+1)},\beta^{(t)})\ge 0$，由此可知，$f(\beta^{(t)})$ 序列单调不增且收敛。事实上，只需

$$f(\beta^{(t+1)})\le g(\beta^{(t+1)},\beta^{(t)})\tag{A.1}$$

即可享有函数值收敛性质，该性质可用于线搜索。

此外，我们还得到

$$(\rho - \rho_+^l(q,q))\mathbf{D}_2(\beta^{(t+1)},\beta^{(t)}) \le f(\beta^{(t)}) - f(\beta^{(t+1)}).$$

最后，我们来研究迭代序列的极限点。首先注意到，在 $\eta_0 > 0$ 下， $\{\beta^{(t)}\}_{t=0}^{\infty}$ 一致有界，因为

$$\eta_0\|\beta^{(t)}\|_2^2/2 \le f(\beta^{(t)}) \le f(\beta^{(0)}).$$

由 $\lim_{t\to\infty}\{f(\beta^{(t)}) - f(\beta^{(t+1)})\} = 0$、$\lim_{t\to\infty}(\rho\mathbf{D}_2 - \mathbf{\Delta}_{l_0})(\beta^{(t+1)},\beta^{(t)}) = 0$，且由于 $\rho > \rho_+^l(q,q)$，

$$\lim_{t\to\infty}(\beta^{(t+1)} - \beta^{(t)}) = 0.$$

设 $\hat{\beta}$ 是 $\beta^{(t)}$ 的任意一个极限点，满足 $\hat{\beta} = \lim_{k\to\infty}\beta^{(j_k)}$，其中某个序列为 j_k。该 n

$$\begin{aligned} 0 = \lim_{k\to\infty}(\beta^{(j_k+1)} - \beta^{(j_k)}) &= \lim_{k\to\infty}\Theta^{\#}\{\beta^{(j_k)} - \nabla l(\beta^{(j_k)})/\rho; q, \eta_0/\rho\} - \hat{\beta} \\ &= \Theta^{\#}\{\hat{\beta} - \nabla l(\hat{\beta})/\rho; q, \eta_0/\rho\} - \hat{\beta}, \end{aligned}$$

其中第二个等号源于 $\nabla l(\beta)$ 的连续性以及 $\Theta^{\#}$-唯一性假设。

定义 $\hat{\mathcal{J}} = \{j : \hat{\beta}_j \neq 0\}$。于是得到

$$\hat{\beta}_{\hat{\mathcal{J}}} = \hat{\beta}_{\hat{\mathcal{J}}}/(1+\eta_0/\rho) - X_{\hat{\mathcal{J}}}^T\nabla l_0(X_{\hat{\mathcal{J}}}\hat{\beta}_{\hat{\mathcal{J}}}; y)/(\rho+\eta_0),$$

或者等价地，

$$\eta_0\hat{\beta}_{\hat{\mathcal{J}}} + X_{\hat{\mathcal{J}}}^T\nabla l_0(X_{\hat{\mathcal{J}}}\hat{\beta}_{\hat{\mathcal{J}}}; y) = 0.$$

因此，给定 $\mathcal{J}$, $\beta^{\mathcal{J}}$是一个驻点

$$\min_{\gamma} l_0(X_{\hat{\mathcal{J}}}\gamma; y) + \eta_0\|\gamma\|_2^2/2. \tag{A.2}$$

当 $l_0(\cdot; y)$ 是凸函数且 $\eta_0 > 0$ 时，(A.2) 是强凸的，因此 $\beta^{\mathcal{J}}$是唯一的极小化器。

根据 Ostrowski 收敛定理， $\beta^{(t)}$ 的极限点集合必定是连通的。另一方面，所有受限最优解的集合$\{\beta^{\mathcal{J}}\}$ 是有限的，因此

$$\lim_{t\to\infty}\beta^{(t)} = \hat{\beta}.$$

在 $\|\mathcal{J}\|_0 = q$下，很容易看出，邻域 $\{\beta\colon \|\beta-\beta\|_\infty < \epsilon,\ J(\beta) \le q\}$，其中 $0 < \epsilon < \min_{j\in\hat{\mathcal{J}}}|\beta_j|$，恰好就是 $\{\beta\colon \mathcal{J}(\beta) = \hat{\mathcal{J}}, |\beta_j - \hat{\beta}_j| < \epsilon, \forall j \in \hat{\mathcal{J}}\}$。因此，$\hat{\beta}$的局部最优性以及 $\beta^{(t)}$的支撑稳定性随之成立。

B. 定理 1 的证明

我们首先介绍一些有助于证明该定理的引理。第一个是对 [44]中引理 9 的推广。

引理 A.1. 令 $\mathcal{J}(B)$表示矩阵 B的行支撑集，并定义 $J(B) = \|B\|_{2,0} = |\mathcal{J}(B)|$。考虑以下关于 $0 \le q \le p, \eta \ge 0$ 的问题:

$$\min_{B\in\mathbb{R}^{p\times m}} \frac{1}{2}\|Y-B\|_F^2 + \frac{\eta}{2}\|B\|_F^2 = l(B) \quad \textit{subject to } \|B\|_{2,0} \le q.$$

那么，$\hat{B} = \vec{\Theta}^{\#}(Y; q, \eta)$（回忆一下在第 I 节中定义的 $\vec{\Theta}^{\#}$）给出了一个全局最优解；对于任意满足$J(B) \le s$的 B，我们有

$$l(B) - l(\hat{B}) \ge (1 - \mathcal{L}(\mathcal{J},\hat{\mathcal{J}}))(1+\eta)\frac{\|\hat{B}-B\|_F^2}{2} \tag{A.3}$$

其中 $\mathcal{J} = \mathcal{J}(B)$、$\hat{\mathcal{J}} = \mathcal{J}(\hat{B})$ 和 $\mathcal{L}(\mathcal{J},\hat{\mathcal{J}}) = \sqrt{|\mathcal{J}\setminus\hat{\mathcal{J}}|/|\hat{\mathcal{J}}\setminus\mathcal{J}|}$。当 $J(\hat{B}) = q$与 $\vartheta(\equiv q/s) \ge 1$时，$\mathcal{L}(\mathcal{J},\hat{\mathcal{J}}) \le \sqrt{|\mathcal{J}|/|\hat{\mathcal{J}}|} \le 1/\sqrt{\vartheta}$。在上述陈述中， $0/0$ 视为 1。

引理 A.2. 存在常数 A、C、$c>0$，使得对于任意 $a>0$，以下事件

$$\sup_{\beta_1,\beta_2}\langle\epsilon, X(\beta_1-\beta_2)\rangle-\frac{1}{2a}\|X(\beta_1-\beta_2)\|_2^2-\frac{a}{2}A\sigma^2\{J(\beta_1)\vee J(\beta_2)\}\log\Big\{\frac{ep}{J(\beta_1)\vee J(\beta_2)}\Big\}\geq\frac{a}{2}\sigma^2 t \tag{A.4}$$

发生的概率至多为 $C\exp(-ct)p^{-cA}$，其中 $t\geq 0$。

ˆ 首先，根据定义，很容易证明 β 满足

$$\hat\beta\in\operatorname*{argmin}_{\beta} g(\beta,\hat\beta),$$

其中 $g(\beta,\beta^-)=\|y-X\beta^-\|_2^2/2+\langle X^T(X\beta^--y)\ \beta-\beta^-\rangle+\rho\|\beta-\beta^-\|_2^2/2+\eta_0\|\beta\|_2^2/2$。由 $g(\hat\beta,\hat\beta)\leq g(\beta^*,\hat\beta)$ 和引理 A.1,

$$\begin{aligned}\frac{1}{2}\|\beta^*-\hat\beta+\frac{1}{\rho}X^T(X\hat\beta-y)\|_2^2-\frac{1}{2}\|\frac{1}{\rho}X^T(X\hat\beta-y)\|_2^2+\frac{\eta_0}{2\rho}\|\beta^*\|_2^2-\frac{\eta_0}{2\rho}\|\hat\beta\|_2^2\\ \geq(1+\frac{\eta_0}{\rho})\frac{1-\mathcal{L}(\mathcal{J}^*,\hat{\mathcal{J}})}{2}\|\hat\beta-\beta^*\|_2^2,\end{aligned}$$

其中 $\mathcal{J}^*=\mathcal{J}(\beta^*)$、$\hat{\ }\mathcal{J}=\mathcal{J}(\hat{\ }\beta)$ 和 $\mathcal{L}(\mathcal{J}^*,\hat{\ }\mathcal{J})\leq 1/\sqrt{\vartheta}$。

由模型 $y=X\beta^*+\epsilon$ 可知。

$$\|X\hat\beta-X\beta^*\|_2^2+\frac{\eta_0}{2}\|\hat\beta\|_2^2\leq\frac{\rho-(\sqrt{\vartheta}-1)\eta_0}{2\sqrt{\vartheta}}\|\hat\beta-\beta^*\|_2^2+\frac{\eta_0}{2}\|\beta^*\|_2^2+\langle X\hat\beta-X\beta^*,\epsilon\rangle,$$

这给出的是

$$\begin{aligned}&\|X\hat\beta-X\beta^*\|_2^2+\frac{\eta_0}{2}\|\hat\beta-\beta^*\|_2^2\\ \leq&\frac{\rho-(\sqrt{\vartheta}-1)\eta_0}{2\sqrt{\vartheta}}\|\hat\beta-\beta^*\|_2^2+\eta_0\langle\hat\beta-\beta^*,-\beta^*\rangle+\langle X\hat\beta-X\beta^*,\epsilon\rangle\\ \leq&\frac{\rho-(\sqrt{\vartheta}-1)\eta_0}{2\sqrt{\vartheta}}\|\hat\beta-\beta^*\|_2^2+\frac{b\eta_0}{2}\|\hat\beta-\beta^*\|_2^2+\frac{\eta_0}{2b}\|\beta^*\|_2^2+\langle X\hat\beta-X\beta^*,\epsilon\rangle\end{aligned} \tag{A.5}$$

对于任意 $b>0$。应用引理 A.2 以及 $t=0$，我们可以证明，对于任意 $a>0$，以下事件

$$\langle X\hat\beta-X\beta^*,\epsilon\rangle\leq\frac{1}{2a}\|X\hat\beta-X\beta^*\|_2^2+\frac{a}{2}A\sigma^2\vartheta s\log\frac{ep}{\vartheta s} \tag{A.6}$$

发生，其发生的概率至少为 $1-Cp^{-c}$，其中 A、C、$c>0$ 是一些普适常数 ts.

结合 (A.5)、(A.6) 和正则性条件 (7)，可得

$$\frac{\eta_0(\varepsilon-b)}{2}\|\hat\beta-\beta^*\|_2^2+\Big(\frac{\delta}{2}-\frac{1}{2a}\Big)\|X\hat\beta-X\beta^*\|_2^2\leq\frac{\eta_0}{2b}\|\beta^*\|_2^2+\frac{a}{2}A\sigma^2\vartheta s\log\frac{ep}{\vartheta s}$$

其发生的概率至少为 $1-Cp^{-c}$。通过选择 $a=2/\delta$ 和 $b=\varepsilon/2$，我们得到预测误差的界为

$$\begin{aligned}\|X\hat\beta-X\beta^*\|_2^2+\frac{\eta_0\varepsilon}{\delta}\|\hat\beta-\beta^*\|_2^2&\leq\frac{4\eta_0}{\delta\varepsilon}\|\beta^*\|_2^2+\frac{4}{\delta^2}A\sigma^2\vartheta s\log\frac{ep}{\vartheta s}\\ &\lesssim\frac{\eta_0}{\delta\varepsilon}\|\beta^*\|_2^2+\frac{1}{\delta^2}\sigma^2\vartheta s\log\frac{ep}{\vartheta s},\end{aligned}$$

该结论以至少为 $1-Cp^{-c}$ 的概率成立。

引理 A.1 的证明 在此证明中，给定矩阵 $B\in\mathbb{R}^{p\times m}$ 和指标集 $\mathcal{I}\subset[p]$，我们用 $B_{\mathcal{I}}$ 表示通过提取 B 中以 $\mathcal{I}$ 为索引的行得到的子矩阵。令 $\mathcal{J}_1=\mathcal{J}\cap\mathcal{J}$、$\mathcal{J}_2=\mathcal{J}\setminus\mathcal{J}$ 和 $\mathcal{J}_3=\mathcal{J}\setminus\mathcal{J}$。于是 $\hat{\ }\mathcal{J}=\mathcal{J}_1\cup\mathcal{J}_3$ 和 $\mathcal{J}=\mathcal{J}_1\cup\mathcal{J}_2$。

容易证明，$B_{\mathcal{J}_1}=Y_{\mathcal{J}_1}/(1+\eta)$ 和 $B_{\mathcal{J}_2}=Y_{\mathcal{J}_2}/(1+\eta)$。将 $B_{\mathcal{J}_1}=Y_{\mathcal{J}_1}/(1+\eta)$ 写作 $+\Delta_{\mathcal{J}_1}$，并将 $B_{\mathcal{J}_3}=Y_{\mathcal{J}_3}/(1+\eta)$ 写作 $+\Delta_{\mathcal{J}_3}$，可得

$$l(B)-l(\hat{B})=\frac{1+\eta}{2}\|\Delta_{\mathcal{J}_1}\|_F^2+\frac{1}{2(1+\eta)}\|Y_{\mathcal{J}_2}\|_F^2+\frac{1+\eta}{2}\|\Delta_{\mathcal{J}_3}\|_F^2-\frac{1}{2(1+\eta)}\|Y_{\mathcal{J}_3}\|_F^2,$$
$$\frac{1+\eta}{2}\|\hat{B}-B\|_F^2=\frac{1+\eta}{2}\|\Delta_{\mathcal{J}_1}\|_F^2+\frac{1}{2(1+\eta)}\|Y_{\mathcal{J}_2}\|_F^2+\frac{1+\eta}{2}\|\frac{1}{1+\eta}Y_{\mathcal{J}_3}+\Delta_{\mathcal{J}_3}\|_F^2.$$

令 $K\le 1$ 满足

$$l(B)-l(\hat{B})\ge\frac{K}{2}(1+\eta)\|\hat{B}-B\|_F^2,$$

这可由以下条件推出：

$$\begin{aligned}&\frac{1}{2(1+\eta)}\|Y_{\mathcal{J}_2}\|_F^2+\frac{1+\eta}{2}\|\Delta_{\mathcal{J}_3}\|_F^2-\frac{1}{2(1+\eta)}\|Y_{\mathcal{J}_3}\|_F^2\\ \ge&\frac{K}{2(1+\eta)}\|Y_{\mathcal{J}_2}\|_F^2+\frac{K(1+\eta)}{2}\|\frac{1}{1+\eta}Y_{\mathcal{J}_3}+\Delta_{\mathcal{J}_3}\|_F^2.\end{aligned}\tag{A.7}$$

(A.7) 等价于

$$(1-K)\|Y_{\mathcal{J}_2}\|_F^2+(1+\eta)^2\|\Delta_{\mathcal{J}_3}\|_F^2\ge(1+\eta)^2K\|\frac{1}{1+\eta}Y_{\mathcal{J}_3}+\Delta_{\mathcal{J}_3}\|_F^2+\|Y_{\mathcal{J}_3}\|_F^2.\tag{A.8}$$

根据构造，对于任意 $i\in\mathcal{J}_2$ 和 $j\in\mathcal{J}_3$，都有 $\|y_i\|_2\ge\|y_j\|_2$。因此 $\|Y_{\mathcal{J}_2}\|_F^2\ge J_2\|Y_{\mathcal{J}_3}\|_F^2/J_3$，由此可知，(A.8) 可由以下条件推出：

$$\{(1-K)(J_2/J_3)-(1+K)\}\|Y_{\mathcal{J}_3}\|_F^2+(1-K)(1+\eta)^2\|\Delta_{\mathcal{J}_3}\|_F^2\ge 2K(1+\eta)\langle Y_{\mathcal{J}_3},\Delta_{\mathcal{J}_3}\rangle.$$

因此，将 K 限制为 $(1+K)/(1-K)\le J_2/J_3$ 或 $K\le(J_2-J_3)/(J_2+J_3)\le 1$，最大可能的 K 应满足

$$\{(1-K)(J_2/J_3)-(1+K)\}\cdot(1-K)=|K|^2$$

或 $(1-K)^2=J_3/J_2$，或 $K=1-\sqrt{J_3/J_2}(\le(J_2-J_3)/(J_2+J_3))$。由此可得

$$\mathcal{L}=1-K=(J_3/J_2)^{1/2}.$$

Not需要注意，当 $\mathcal{J}_2=\emptyset$、K 可以分别取 $-\infty$（对应 $\mathcal{J}_3\neq\emptyset$）和 0（对应 $\mathcal{J}_3=\emptyset$），以确保 (A. 8).

现在假设 $J(\hat{B})=q$，其中 $\vartheta\ge 1$。如果 $\mathcal{J}_2\neq\emptyset$，则 $\mathcal{L}\le\sqrt{(J_3+J_1)/(J_2+J_1)}=\sqrt{J/\hat{J}}\le 1/\sqrt{\vartheta}$。否则，我们必有 $\mathcal{J}_3=\emptyset$，$\mathcal{J}=\hat{\mathcal{J}}$ 且 $\vartheta=1$。证明完毕。

该引理可用于分析 ℓ_0 约束的（逐元素）稀疏性追求，以及组变量选择（参见第 IV-D 节）。

引理 A.2 的证明 给定一个矩阵 A，记 $\mathcal{P}_A$ 为其值域上的正交投影，记 $\mathcal{P}_A^{\perp}$ 为其正交补。在证明中，对于任意 $J\subset[p]$，$\mathcal{P}_{\mathcal{J}}$ 用作证明中 $\mathcal{P}_{X_{\mathcal{J}}}$ 的简写。令 $\mathcal{J}_1=\mathcal{J}(\beta_1)$ $\mathcal{J}_2=\mathcal{J}(\beta_2)$ $J_1=|\mathcal{J}_1|, J_2=|\mathcal{J}_2|$。

首先请注意，式（A.4）中使用的是项 $\{J(\beta_1)\vee J(\beta_2)\}\log[ep/\{J(\beta_1)\vee J(\beta_2)\}]$，而不是 $J(\beta_1-\beta_2)\log\{ep/J(\beta_1-\beta_2)\}$。尽管 $J(\beta_1-\beta_2)\le J(\beta_1)+J(\beta_2)$，$J(\beta_1)+J(\beta_2)$ 仍可能大于 p。为解决这一问题，我们采用一种分解技巧

$$\begin{aligned}X\beta_1-X\beta_2&=\mathcal{P}_{\mathcal{J}_1}X(\beta_1-\beta_2)+\mathcal{P}_{\mathcal{J}_1}^{\perp}X(\beta_1-\beta_2)\\&=\mathcal{P}_{\mathcal{J}_1}X(\beta_1-\beta_2)+\mathcal{P}_{\mathcal{J}_1}^{\perp}\mathcal{P}_{\mathcal{J}_2}X(\beta_1-\beta_2).\end{aligned}$$

令 $\Delta=\beta_1-\beta_2$。那么

$$\langle\epsilon,X\Delta\rangle=\langle\epsilon,P_{\mathcal{J}_1}X\Delta\rangle+\langle\epsilon,\mathcal{P}_{\mathcal{J}_1}^{\perp}\mathcal{P}_{\mathcal{J}_2}X\Delta\rangle.\tag{A.9}$$

我们来界定 (A.9) 右侧的第一项。定义 $P_o(J) = \sigma^2 J\log(ep/J)$（其中 $0 \le J \le p$），这是一个递增函数；并且 $\Gamma_J = \{\alpha \in \mathbb{R}^p : \ \|\alpha\|_2 \le 1, \alpha \in \mathcal{P}_{\mathcal{J}}$，其中某些 $\mathcal{J} \subset [p], |\mathcal{J}| \le J\}$。那么对于任意 $a, b > 0$

$$
\begin{aligned}
&\langle \epsilon, \mathcal{P}_{\mathcal{J}_1} X\Delta\rangle - \frac{1}{a}\|\mathcal{P}_{\mathcal{J}_1} X\Delta\|_2^2 - bLP_o(J_1) \\
\le& \|\mathcal{P}_{\mathcal{J}_1} X\Delta\|_2 \langle \epsilon, \frac{\mathcal{P}_{\mathcal{J}_1} X\Delta}{\|\mathcal{P}_{\mathcal{J}_1} X\Delta\|_2}\rangle - 2\|\mathcal{P}_{\mathcal{J}_1} X\Delta\|_2 \sqrt{\frac{b}{a} LP_o(J_1)} \\
\le& \frac{1}{a}\|\mathcal{P}_{\mathcal{J}_1} X\Delta\|_2^2 + \frac{a}{4} \sup_{J_1 \le p} \sup_{\Delta \in \Gamma_{J_1}} \left\{\langle \epsilon, \Delta\rangle - 2\sqrt{(b/a)LP_o(J_1)}\right\}_+^2 \\
\equiv& \frac{1}{a}\|\mathcal{P}_{\mathcal{J}_1} X\Delta\|_2^2 + \frac{a}{4} \sup_{J_1 \le p} R_{J_1}^2,
\end{aligned}
$$

其中 $R_{J_1} := \sup_{\Delta \in \Gamma_{J_1}} \left\{\langle \epsilon, \Delta\rangle - 2\sqrt{(b/a)LP_o(J_1)}\right\}_+$，$L$ 为一个充分大的常数。当 $J_1 = 0$时，$R_{J_1} = 0$。当 $J_1 \ge 1$时，对于任意 $t \ge 0$，若 $4b/a$ 是一个大于 1 的常数，则有

$$
\begin{aligned}
&\mathbb{P}(\sup_{1 \le J_1 \le p} R_{J_1} \ge t\sigma) \\
\le& \sum_{J_1=1}^{p} \mathbb{P}\left(\sup_{\Delta \in \Gamma_{J_1}} \langle \epsilon, \Delta\rangle - \sqrt{LP_o(J_1)} \ge t\sigma + 2\sqrt{\frac{b}{a} LP_o(J_1)} - \sqrt{LP_o(J_1)}\right) \\
\le& C\exp(-ct^2) \sum_{J_1=1}^{p} \exp[-c(2\sqrt{b/a} - 1)^2 LP_o(J_1)/\sigma^2] \\
\le& C\exp(-ct^2)\exp(-cL\log p) \sum_{J_1=1}^{p} \exp(-cLJ_1) \\
\le& C\exp(-ct^2) p^{-cL}.
\end{aligned}
\tag{A.10}
$$

第二个不等式由 [21], 的引理 6 得出，并且在第三个不等式中，对于任意 $J \in [p]$，我们使用了 $J\log(ep/J) \ge J + \log p$。因此，对于任意 a、$b > 0$、$4b > a$ 和 $t \ge 0$，有

$$
\mathbb{P}\left\{\langle \epsilon, \mathcal{P}_{\mathcal{J}_1} X\Delta\rangle - \frac{2}{a}\|\mathcal{P}_{\mathcal{J}_1} X\Delta\|_2^2 - bLP_o(J_1) \ge \frac{a}{4} t\sigma^2\right\} \le C\exp(-ct) p^{-Lc}. \tag{A.11}
$$

类似地，对于（A.9）中的第二项，我们可以使用 [13] 的引理 7 证明，对于任意 $t \ge 0$,

$$
\mathbb{P}\left[\langle \epsilon, \mathcal{P}_{\mathcal{J}_1}^{\perp} \mathcal{P}_{\mathcal{J}_2} X\Delta\rangle - \frac{2}{a}\|\mathcal{P}_{\mathcal{J}_1}^{\perp} \mathcal{P}_{\mathcal{J}_2} X\Delta\|_2^2 - bL\{P_o(J_1) + P_o(J_2)\} \ge \frac{a}{4} t\sigma^2\right] \le C\exp(-ct) p^{-Lc}. \tag{A.12}
$$

结合（A.11）和（A.12），并利用 $\|\mathcal{P}_{\mathcal{J}_1} X\Delta\|_2^2 + \|\mathcal{P}_{\mathcal{J}_1}^{\perp} P_{\mathcal{J}_2} X\Delta\|_2^2 = \|X\Delta\|_2^2$这一事实，对于任意 a、$b > 0$、$4b > a$ 和 $t \ge 0$，我们得到

$$
\mathbb{P}\left[\langle \epsilon, X\Delta\rangle - \frac{4}{a}\|X\Delta\|_2^2 - 3bL\{P_o(J_1) \vee P_o(J_2)\} \ge \frac{a}{2} t\sigma^2\right] \le C\exp(-ct) p^{-Lc}. \tag{A.13}
$$

最后，利用 $P_o(J)$ 关于 $J \in [0, p]$ 的递增性，我们有 $P_o(J_1) \vee P_o(J_2) \le (J_1 \vee J_2)\log\{ep/(J_1 \vee J_2)\}$。对 (A.13) 重新参数化即可得到结论。

C. 定理 2 的证明

由定理 1 的证明可知，以概率 $1 - Cp^{-c}$,

$$
\begin{aligned}
\|X\hat{\beta} - X\beta^*\|_2^2 + \frac{\eta_0(1-b)}{2}\|\hat{\beta} - \beta^*\|_2^2 \le& \frac{\rho - (\sqrt{\vartheta} - 1)\eta_0}{2\sqrt{\vartheta}}\|\hat{\beta} - \beta^*\|_2^2 + \frac{\eta_0}{2b}\|\beta^*\|_2^2 + \\
& \frac{1}{2a}\|X\hat{\beta} - X\beta^*\|_2^2 + \frac{a}{2} A\sigma^2 \vartheta s \log\frac{ep}{\vartheta s},
\end{aligned}
$$

由此得到

$$\|X\hat{\beta} - X\beta^*\|_2^2 - \frac{\eta_0 b}{2}\|\hat{\beta} - \beta^*\|_2^2 \le \frac{\rho - (2\sqrt{\vartheta} - 1)\eta_0}{2\sqrt{\vartheta}}\|\hat{\beta} - \beta^*\|_2^2 + \frac{\eta_0}{2b}\|\beta^*\|_2^2 + \frac{\rho_+((1+\vartheta)s)}{2a}\|\hat{\beta} - \beta^*\|_2^2 + \frac{a}{2}A\sigma^2\vartheta s \log\frac{ep}{\vartheta s}.$$

在正则性条件 (9) 下，选取 $a = 2/\delta$ 和 $b = \delta\rho_+((1+\vartheta)s)/(4\eta_0)$ 可得 (10)。（该结果同样适用于 $\eta_0 = 0$ 。）

为证明第二个结果，注意到根据定义，$\hat{\beta}$ 必须满足 $\hat{\beta} = \Theta^{\#}\{\hat{\beta} - \frac{1}{\rho}X^T\nabla l_0(X\hat{\beta}), \eta_0/\rho\}$，这意味着

$$\begin{aligned}
&\left\|\hat{\beta}(1+\eta_0/\rho) - \hat{\beta} + \frac{1}{\rho}X^T\nabla l_0(X\hat{\beta})\right\|_\infty \le (1+\eta_0/\rho)\min_{j\in\hat{\mathcal{J}}}|\hat{\beta}_j| \\
\Longrightarrow &\left\|\eta_0\hat{\beta} + X^T(\nabla l_0(X\hat{\beta}) - \nabla l_0(X\beta^*)) - X^T\epsilon\right\|_\infty \le (\rho+\eta_0)\min_{j\in\hat{\mathcal{J}}}|\hat{\beta}_j| \\
\Longrightarrow &\left\|X^T(\nabla l_0(X\hat{\beta}) - \nabla l_0(X\beta^*)) + \eta_0(\hat{\beta} - \beta^*)\right\|_\infty \le \|X^T\epsilon\|_\infty + \eta_0\|\beta^*\|_\infty + (\rho+\eta_0)\min_{j\in\hat{\mathcal{J}}}|\hat{\beta}_j|.
\end{aligned}$$

接下来，我们引入一个引理。

引理 A.3。设 $\tilde{\beta}$, $\beta \in \mathbb{R}^p$ 满足 $\|\beta\|_0 = q > s \ge \|\beta\|_0$，为简便起见，分别用 $\tilde{\mathcal{J}}$ 和 $\mathcal{J}$ 表示 $\mathcal{J}(\tilde{\beta})$ 和 $\mathcal{J}(\beta)$。那么

$$\min_{j\in\tilde{\mathcal{J}}}|\tilde{\beta}_j| \le \min_{j\in\tilde{\mathcal{J}}\setminus\mathcal{J}}|\tilde{\beta}_j| \le \frac{\|(\tilde{\beta}-\beta)_{\tilde{\mathcal{J}}\setminus\mathcal{J}}\|_2}{\sqrt{|\tilde{\mathcal{J}}\setminus\mathcal{J}|}} \le \frac{\|(\tilde{\beta}-\beta)_{\tilde{\mathcal{J}}\setminus\mathcal{J}}\|_2}{\sqrt{q-s}} \le \frac{\|\tilde{\beta}-\beta\|_2}{\sqrt{q-s}} \tag{A.14}$$

$$\min_{j\in\tilde{\mathcal{J}}}|\tilde{\beta}_j| \le \max_{j\in\tilde{\mathcal{J}}\setminus\mathcal{J}}|\tilde{\beta}_j| = \|(\tilde{\beta}-\beta)_{\tilde{\mathcal{J}}\setminus\mathcal{J}}\|_\infty \le \|\tilde{\beta}-\beta\|_\infty. \tag{A.15}$$

证明很简单，故略去。现在，结合正则性条件 (11) 与 (A.14) 或 (A.15)，即可得到所需结果。

D. 定理 3 的证明

根据定义，我们有

$$\rho_+(2q) = \sup_{I\in[p]:|I|=2q}\lambda_{\max}(X_I^TX_I),$$

并且在 $q+s \le n$ 下，

$$\rho_-(q+s) = \inf_{I\in[p]:|I|=q+s}\lambda_{\min}(X_I^TX_I).$$

根据 [23], 的定理 6.1，我们有

$$\mathbb{P}\left\{\sqrt{\frac{\lambda_{\max}(X_I^TX_I)}{n}} \ge (1+c_0)\sqrt{\lambda_{\max}(\Sigma_I)} + \sqrt{\frac{\mathrm{tr}(\Sigma_I)}{n}}\right\} \le \exp(-nc_0^2/2), \quad \forall I : |I| = 2q$$

and

$$\mathbb{P}\left\{\sqrt{\frac{\lambda_{\min}(X_I^TX_I)}{n}} \le (1-c_0)\sqrt{\lambda_{\min}(\Sigma_I)} - \sqrt{\frac{\mathrm{tr}(\Sigma_I)}{n}}\right\} \le \exp(-nc_0^2/2), \quad \forall I : |I| = q+s$$

对所有 $c_0 > 0$ 均成立。应用并集界可得

$$\mathbb{P}\left\{\sqrt{\frac{\rho_+(2q)}{n}} \ge (1+c_0)\sqrt{\lambda_{\max}^{(2q)}} + \sqrt{\frac{2q}{n}}\right\} \le \binom{p}{2q}\exp(-nc_0^2/2). \tag{A.16}$$

设 $nc^2 = nc_0^2 - \log\binom{p}{2q}$，再利用 $\log\binom{p}{2q} \le 2q\log(ep/q)$，$c_0 \le c + \sqrt{2q\log(ep/q)/n}$。因此，对于任意 $c > 0$,

$$\mathbb{P}\left\{\sqrt{\frac{\rho_+(2q)}{n}} \ge (1+c)\sqrt{\lambda_{\max}^{(2q)}} + \sqrt{\frac{2q\log(ep/q)}{n}}\sqrt{\lambda_{\max}^{(2q)}} + \sqrt{\frac{2q}{n}}\right\} \le \exp(-nc^2/2). \tag{A.17}$$

同理,

$$\mathbb{P}\left\{\sqrt{\frac{\rho_-(q+s)}{n}} \le (1-c)\sqrt{\lambda_{\min}^{(q+s)}} - \sqrt{\frac{(q+s)\log(ep/q)}{n}}\sqrt{\lambda_{\min}^{(q+s)}} - \sqrt{\frac{q+s}{n}}\right\} \le \exp(-nc^2/2).$$

令 $c \in (0,1)$，并假设 $n \ge \{2(q+s)/(1-c)^2\}\{1/\lambda^{(q+s)}$ 最小 $+\log(ep/q)\}$. 该 n

$$\frac{\rho_+(2q)}{\rho_-(q+s)} \le \left\{\frac{(1+c)\sqrt{\lambda_{\max}^{(2q)}} + \sqrt{\{2\lambda_{\max}^{(2q)} q\log(ep/q)\}/n} + \sqrt{2q/n}}{(1-c)\sqrt{\lambda_{\min}^{(q+s)}} - \sqrt{\{\lambda_{\min}^{(q+s)}(q+s)\log(ep/q)\}/n} - \sqrt{(q+s)/n}}\right\}^2$$

以至少 $1-2\exp(-nc^2/2)$ 的概率成立。

E. 定理5的证明

令 $E := \sigma^2 P_o(q) + \sigma^2$。类似于定理 1 的证明，根据 g 的构造和引理 A.1，我们有 e

$$\rho(1-1/\sqrt{\vartheta})(1+\eta_0/\rho)\mathbf{D}_2(\beta^*,\hat{\beta}) + g(\hat{\beta},\hat{\beta}) \le g(\beta^*,\hat{\beta}),$$

因此

$$2\bar{\mathbf{\Delta}}_{l_0}(X\hat{\beta},X\beta^*) + \frac{\eta_0}{2}\|\hat{\beta}\|_2^2 \le \frac{\rho-(\sqrt{\vartheta}-1)\eta_0}{\sqrt{\vartheta}}\mathbf{D}_2(\hat{\beta},\beta^*) + \frac{\eta_0}{2}\|\beta^*\|_2^2 + \langle\epsilon, X\hat{\beta}-X\beta^*\rangle. \tag{A.18}$$

应用引理 A.2 得到

$$\langle\epsilon, X\hat{\beta}-X\beta^*\rangle \le \delta\mathbf{D}_2(X\hat{\beta},X\beta^*) + \frac{1}{\delta}A\sigma^2P_o(q) + R \tag{A.19}$$

对于任意 $\delta>0$，其中 $R :=$ 上确界$_{\beta_1,\ \beta_2}\{\langle\epsilon, X\beta_1 - X\beta_2\rangle - \delta\mathrm{D}_2(X\beta_1, X\beta_2) - A\sigma^2P_o(q)/\delta\}_+$ 且

$$\mathbb{P}(\delta R > \sigma^2 t) \le C\exp(-ct)p^{-cA},$$

其中， A、C、$c>0$ 为一些常数。因此,

$$\mathbb{E}\langle\epsilon, X\hat{\beta}-X\beta^*\rangle \le \mathbb{E}\{\delta\mathbf{D}_2(X\hat{\beta},X\beta^*)\} + \frac{C}{\delta}(\sigma^2P_o(q)+\sigma^2). \tag{A.20}$$

结合（A.18）和（A.20）可得

$$\begin{aligned}&\mathbb{E}\{(2\bar{\mathbf{\Delta}}_{l_0} - \delta\mathbf{D}_2)(X\hat{\beta},X\beta^*) + \eta_0\mathbf{D}_2(\hat{\beta},\beta^*)\}\\ \le&\mathbb{E}\Big\{\frac{\rho-(\sqrt{\vartheta}-1)\eta_0}{\sqrt{\vartheta}}\ \mathbf{D}_2(\hat{\beta},\beta^*) + \eta_0\langle-\beta^*,\hat{\beta}-\beta^*\rangle\Big\} + \frac{C}{\delta}E,\end{aligned} \tag{A.21}$$

因此

$$\mathbb{E}\Big[(2\bar{\mathbf{\Delta}}_{l_0} - \delta\mathbf{D}_2)(X\hat{\beta},X\beta^*) - \frac{\rho-\{(2-\varepsilon)\sqrt{\vartheta}-1\}\eta_0}{\sqrt{\vartheta}}\ \mathbf{D}_2(\hat{\beta},\beta^*)\Big] \le \frac{C}{\delta}E + \frac{\eta_0}{2\varepsilon}\|\beta^*\|_2^2 \tag{A.22}$$

对于任意 $\varepsilon,\delta>0$。

接下来，由 $l_0(X\hat{}\beta) + \eta_0\|\hat{}\beta\|_2^2/2 \le l_0(X\beta^{(0)}) + \eta_0\|\beta^{(0)}\|_2^2/2$，可得

$$\begin{aligned}&\mathbf{\Delta}_{l_0}(X\hat{\beta},X\beta^*) + \eta_0\mathbf{D}_2(\hat{\beta},\beta^*)\\ \le&\mathbf{\Delta}_{l_0}(X\beta^{(0)},X\beta^*) + \eta_0\mathbf{D}_2(\beta^{(0)},\beta^*) + \eta_0\langle-\beta^*,\hat{\beta}-\beta^*\rangle - \eta_0\langle-\beta^*,\beta^{(0)}-\beta^*\rangle\\ &+\langle\epsilon, X\hat{\beta}-X\beta^*\rangle - \langle\epsilon, X\beta^{(0)}-X\beta^*\rangle.\end{aligned} \tag{A.23}$$

因此，对于任意 $\delta',\delta'',\varepsilon'>0$

$$\begin{aligned}&\mathbb{E}\{(\mathbf{\Delta}_{l_0} - \delta'\mathbf{D}_2)(X\hat{\beta},X\beta^*) + \eta_0\mathbf{D}_2(\hat{\beta},\beta^*)\}\\ \le&\mathbb{E}\big\{(\mathbf{\Delta}_{l_0} + \delta''\mathbf{D}_2)(X\beta^{(0)},X\beta^*) + \eta_0\mathbf{D}_2(\beta^{(0)},\beta^*) + \frac{\eta_0}{2\varepsilon}\|\beta^*\|_2^2 + \eta_0\varepsilon\mathbf{D}_2(\hat{\beta},\beta^*)\\ &+\frac{\eta_0}{2\varepsilon'}\|\beta^*\|_2^2 + \eta_0\varepsilon'\mathbf{D}_2(\beta^{(0)},\beta^*)\big\} + CE\big(\frac{1}{\delta'}+\frac{1}{\delta''}\big).\end{aligned}$$

根据起始点 $\mathbb{E}\{\mathbf{D}_2(\beta^{(0)},\beta^*)\}\le CME/n$ 的假设，我们有

$$\mathbb{E}\{\mathbf{D}_2(X\beta^{(0)},X\beta^*)\}\le C\rho_+(q+s)ME/n,\ \mathbb{E}\{\mathbf{\Delta}_{l_0}(X\beta^{(0)},X\beta^*)\}\le C\rho_+^l(q,s)ME/n.$$

取 $1/\delta''=\sqrt{\rho_+(q+s)M/n}$，可得

$$\begin{aligned}&\mathbb{E}\{(\mathbf{\Delta}_{l_0}-\delta'\mathbf{D}_2)(X\hat\beta,X\beta^*)+\eta_0(1-\varepsilon)\mathbf{D}_2(\hat\beta,\beta^*)\}\\ \le& CE\big(\frac{1}{\delta'}+\sqrt{\frac{\rho_+(q+s)M}{n}}+\frac{\rho_+^l(q,s)}{n}M+\frac{\eta_0(1+\varepsilon')}{n}M\big)+\eta_0\big(\frac{1}{\varepsilon}+\frac{1}{\varepsilon'}\big)\frac{\|\beta^*\|_2^2}{2}.\end{aligned}$$

设 $Q_0:=\sqrt{\rho_+(q+s)M/n}+\rho_+^l(q,s)M/n+\eta_0(1+\varepsilon')M/n$。则

$$CE\big(\frac{1}{\delta'}+Q_0\big)\le\frac{C}{c_1\wedge c_2}E\big(\frac{c_1}{\delta'}+c_2Q_0\big)$$

对任意 $c_1,c_2>0$。取 δ'：$\delta^2=\delta'^2/(c_1+c_2Q_0\delta')$ 和 ε'：$1/\varepsilon+1/\varepsilon'=(1/\delta'+Q_0)c_3\delta/\varepsilon$，其中 $c_3>0$ 为某个较大的常数，可得

$$\mathbb{E}\{(\frac{\delta}{\delta'}\mathbf{\Delta}_{l_0}-\delta\mathbf{D}_2)(X\hat\beta,X\beta^*)+\frac{\delta}{\delta'}\eta_0(1-\varepsilon)\mathbf{D}_2(\hat\beta,\beta^*)\}\le\frac{CE}{c_1\wedge c_2}\frac{1}{\delta}+c_3\frac{\eta_0}{2\varepsilon}\|\beta^*\|_2^2.\tag{A.24}$$

将 (A.22) 乘以 $(1-1/M)$，将 (A.24) 乘以 $1/M$，再将两个不等式相加，可得

$$\begin{aligned}&\mathbb{E}\Big[(1-\frac{1}{M})\big\{2\bar{\mathbf{\Delta}}_{l_0}(X\hat\beta,X\beta^*)-\frac{\rho-\{(2-\varepsilon)\sqrt{\vartheta}-1\}\eta_0}{\sqrt{\vartheta}}\mathbf{D}_2(\hat\beta,\beta^*)\big\}\\ &+(\frac{\delta}{M\delta'}\mathbf{\Delta}_{l_0}-\delta\mathbf{D}_2)(X\hat\beta,X\beta^*)+\frac{\delta}{M\delta'}\eta_0(1-\varepsilon)\mathbf{D}_2(\hat\beta,\beta^*)\Big]\\ \le& C\big(\frac{E}{\delta}+\frac{\eta_0}{\varepsilon}\|\beta^*\|_2^2\big).\end{aligned}\tag{A.25}$$

简单计算可得

$$\frac{\delta'}{\delta}=\frac{c_2Q_0\delta+\sqrt{c_2^2Q_0^2\delta^2+4c_1}}{2}\le\frac{\sqrt{2}+1}{2}\{c_2Q_0\delta\vee\sqrt{4c_1}\}\le C(Q_0\delta\vee1).$$

由此可得

$$\varepsilon'\le\frac{\varepsilon}{C(Q_0\delta\vee1)+\delta Q_0-1}\le C\frac{\varepsilon}{Q_0\delta\vee1}\le C\varepsilon$$

对于某个较大的常数 C，因此 $Q_0\lesssim Q$。在.....条件下

$$\begin{aligned}&K\sigma^2P_o(\vartheta s)+\Big\{2(1-\frac{1}{M})\bar{\mathbf{\Delta}}_{l_0}+\frac{C}{M(Q\delta\vee1)}\mathbf{\Delta}_{l_0}-2\delta\mathbf{D}_2\Big\}(X\hat\beta,X\beta^*)\\ \ge&\frac{1-1/M}{\sqrt{\vartheta}}\big[\rho-\{(2-\varepsilon)\sqrt{\vartheta}-1\}\eta_0\big]\mathbf{D}_2(\hat\beta,\beta^*)-\frac{C}{M(Q\delta\vee1)}\eta_0(1-\varepsilon)\mathbf{D}_2(\hat\beta,\beta^*),\end{aligned}\tag{A.26}$$

(A.25) 得出

$$\begin{aligned}\mathbb{E}[\mathbf{D}_2(X\hat\beta,X\beta^*)]\le&\frac{K}{\delta}\sigma^2P_o(\vartheta s)+\frac{CE}{\delta^2}+C\frac{\eta_0}{\varepsilon}\|\beta^*\|_2^2\\ \lesssim&\frac{K\delta\vee1}{\delta^2}E+\frac{\eta_0}{\delta\varepsilon}\|\beta^*\|_2^2.\end{aligned}\tag{A.27}$$

通过重新参数化，正则性条件 (30) 蕴含 (A.26)。

F. 定理 6 的证明

为方便起见，将 $\mathbf{D}_2(X\beta, X\beta')$ 记作 $\mathbf{D}_{2,\ X}(\beta, \beta')$。由引理 A.1，我们有

$$g(\beta^*, \beta^{(t)}) - g(\beta^{(t+1)}; \beta^{(t)}) \geq \rho_{t+1}(1 - \mathcal{L}_{t+1})(1 + \bar{\eta}_{t+1})\mathbf{D}_2(\beta^{(t+1)}, \beta^*), \tag{A.28}$$

其中 $\mathcal{L}_{t+1} = \mathcal{L}(\mathcal{J}(\beta^*)\ , \mathcal{J}(\beta^{(t+1)})) \leq 1/\sqrt{\vartheta_{t+1}}$. （回想 $\vartheta_{t+1} = q_{t+1}/s > 1$，且 $s \geq \|\beta^*\|_0$.）

将 $g(\beta, \beta^{(t)}) = l(\beta) + \eta_{t+1}\mathbf{D}_2(\beta, 0) + (\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta, \beta^{(t)})$ 和 $l(\beta^*) - l(\beta^{(t+1)}) = \langle \epsilon, X\beta^{(t+1)} - X\beta^* \rangle - \grave{\mathbf{\Delta}}_l(\beta^*, \beta^{(t+1)})$ 代入 (A.28) 得到

$$\begin{aligned}
&\{\rho_{t+1}(1 - \mathcal{L}_{t+1})(1 + \bar{\eta}_{t+1})\mathbf{D}_2 + \grave{\mathbf{\Delta}}_l\}(\beta^*, \beta^{(t+1)}) + \eta_{t+1}\mathbf{D}_2(\beta^*, \beta^{(t+1)}) \\
&+ (\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta^{(t+1)}, \beta^{(t)}) \\
\leq &(\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta^*, \beta^{(t)}) + \langle \epsilon, X\beta^{(t+1)} - X\beta^* \rangle + \eta_{t+1}\langle -\beta^*, \beta^{(t+1)} - \beta^* \rangle.
\end{aligned} \tag{A.29}$$

根据引理 A.2，概率至少为 $1 - Cp^{-cA}$

$$\langle \epsilon, X\beta^{(t+1)} - X\beta^* \rangle \leq \delta_{t+1}\mathbf{D}_{2,X}(\beta^*, \beta^{(t+1)}) + \delta_{t+1}^{-1}A\sigma^2 P_o(q_{t+1}), \text{ for all } t \geq 0 \tag{A.30}$$

对于任意 $\delta_{t+1} > 0$，其中 A 是一个常数。此外，对于任意 $\varepsilon_{t+1} > 0$,

$$\langle -\beta^*, \beta^{(t+1)} - \beta^* \rangle \leq \varepsilon_{t+1}\mathbf{D}_2(\beta^*, \beta^{(t+1)}) + \varepsilon_{t+1}^{-1}\mathbf{D}_2(\beta^*, 0). \tag{A.31}$$

将这些界代入（A.29）可得

$$\begin{aligned}
&\big\{\rho_{t+1}(1 - \mathcal{L}_{t+1})(1 + \bar{\eta}_{t+1})\mathbf{D}_2 + \grave{\mathbf{\Delta}}_l + (1 - \varepsilon_{t+1})\eta_{t+1}\mathbf{D}_2 - \delta_{t+1}\mathbf{D}_{2,X}\big\}(\beta^*, \beta^{(t+1)}) \\
&+ (\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta^{(t+1)}, \beta^{(t)}) \\
\leq &(\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta^*, \beta^{(t)}) + \delta_{t+1}^{-1}A\sigma^2 P_o(q_{t+1}) + \varepsilon_{t+1}^{-1}\eta_{t+1}\mathbf{D}_2(\beta^*, 0).
\end{aligned} \tag{A.32}$$

通过 th e （广义）等距数的定义，并使用 $\mathcal{L}_{t+1} \leq 1/\sqrt{\vartheta_{t+1}}$，我们有 e

$$\begin{aligned}
&\Big\{\rho_{t+1}\big(1 - \frac{1}{\sqrt{\vartheta_{t+1}}}\big)(1 + \bar{\eta}_{t+1}) + \rho_-^l(q_{t+1}, s) + (1 - \varepsilon_{t+1})\eta_{t+1} - \delta_{t+1}\rho_+(q_{t+1} + s)\Big\}\mathbf{D}_2(\beta^*, \beta^{(t+1)}) \\
&+ (\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta^{(t+1)}, \beta^{(t)}) \\
\leq &\big\{\rho_{t+1} - \rho_-^l(s, q_{t+1})\big\}\mathbf{D}_2(\beta^*, \beta^{(t)}) + \delta_{t+1}^{-1}A\sigma^2 P_o(q_{t+1}) + \varepsilon_{t+1}^{-1}\eta_{t+1}\mathbf{D}_2(\beta^*, 0).
\end{aligned} \tag{A.33}$$

令 ε_0 为任意数 $\in (0, 1]$。取 $\varepsilon_{t+1} = \varepsilon_0/2, \delta_{t+1} = (\varepsilon_0\rho_-^l(q_{t+1}, s) + \varepsilon_0\eta_{t+1}/2)/\rho_+(q_{t+1} + s)$，我们有

$$\begin{aligned}
&(1 - 1/\sqrt{\vartheta_{t+1}})(1 + \bar{\eta}_{t+1})\rho_{t+1} + \rho_-^l(q_{t+1}, s) + (1 - \varepsilon_{t+1})\eta_{t+1} - \delta_{t+1}\rho_+(q_{t+1} + s) \\
= &(1 - 1/\sqrt{\vartheta_{t+1}})(1 + \bar{\eta}_{t+1})\rho_{t+1} + (1 - \varepsilon_0)\rho_-^l(q_{t+1}, s) + (1 - \varepsilon_0)\eta_{t+1}.
\end{aligned}$$

Let

$$\begin{aligned}
E_{t+1} &= \frac{1}{\rho_{t+1} - \rho_-^l(s, q_{t+1})}\Big\{\frac{A\sigma^2}{\varepsilon_0}\frac{\rho_+(q_{t+1} + s)}{\rho_-^l(q_{t+1}, s) + \eta_{t+1}/2}P_o(q_{t+1}) + \frac{\eta_{t+1}}{\varepsilon_0}\|\beta^*\|_2^2\Big\} \\
&\leq \frac{A\sigma^2}{\varepsilon_0}\frac{\rho_+(q_{t+1} + s)}{(\rho_-^l(q_{t+1}, s)/\rho_{t+1} \vee \bar{\eta}_{t+1})(1 - \rho_-^l(s, q_{t+1})/\rho_{t+1})\rho_{t+1}^2}P_o(q_{t+1}) \\
&\quad + \frac{\bar{\eta}_{t+1}}{\varepsilon_0(1 - \rho_-^l(s, q_{t+1})/\rho_{t+1})}\|\beta^*\|_2^2
\end{aligned}$$

对于任意 $t \geq 0$。根据 κ_t、h_t 的定义，我们可以得到

$$\mathbf{D}_2(\beta^*, \beta^{(t+1)}) + h_{t+1}(\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta^{(t+1)}, \beta^{(t)}) \leq \kappa_{t+1}\mathbf{D}_2(\beta^*, \beta^{(t)}) + \kappa_{t+1}E_{t+1}. \tag{A.34}$$

对 $t = T, \ldots, 0$ 应用递归论证可得

$$\begin{aligned}
&\mathbf{D}_2(\beta^*, \beta^{(T+1)}) + \sum_{t=0}^{T}\big(\Pi_{\tau=t}^{T}h_{\tau+1}\big)(\rho_{t+1}\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta^{(t+1)}, \beta^{(t)}) \\
\leq &\Big(\Pi_{t=0}^{T}\kappa_{t+1}\Big)\mathbf{D}_2(\beta^*, \beta^{(0)}) + \sum_{t=0}^{T}\Big(\Pi_{\tau=t}^{T}\kappa_{\tau+1}\Big)E_{t+1},
\end{aligned}$$

从而得到界（35）。

为确保

$$\frac{\rho_t - \rho_-^l(s,q_t)}{(1-1/\sqrt{\vartheta_t})(1+\bar{\eta}_t)\rho_t + (1-\varepsilon)(\rho_-^l(q_t,s)+\eta_t)} \le \frac{1}{1+\alpha} \tag{A.35}$$

对于某个 $\alpha > 0$，我们需要

$$\bar{\eta}_t \ge \frac{(\alpha + 1/\sqrt{\vartheta_t}) - (2+\alpha-\varepsilon)\{\rho_-^l(s,q_t) \wedge \rho_-^l(q_t,s)\}/\rho_t}{2 - 1/\sqrt{\vartheta_t} - \varepsilon}. \tag{A.36}$$

该推论的结果可通过取 $\alpha = \varepsilon$ 并注意到 $\rho_{t+1} \ge \rho_+^l(q_{t+1}, q_t)$ 蕴含 $(\rho_{t+1}\mathrm{D}_2 - \mathbf{\Delta}_l)(\beta^{t+1}, \beta^t) \ge 0$ 得出。

G. 受限等距性下的递归逐坐标误差界

回顾在 (32) 中定义的一般过程，

$$\beta^{(t+1)} = \Theta^{\#}\left\{\beta^{(t)} - \rho_{t+1}^{-1} X^T \nabla l_0(X\beta^{(t)}; y); q_{t+1}, \bar{\eta}_{t+1}\right\}, \text{ with } \bar{\eta}_{t+1} = \eta_{t+1}/\rho_{t+1}. \tag{A.37}$$

采用与定理 2 针对不动点集合的类似方法，在适当的正则性条件下，可以为 $\beta^{(t+1)}$ 建立 ∞-范数下的误差界。

为便于证明，我们首先回顾式 (21) 中给出的 $\rho_-^l(s_1,s_2)$ 的定义。特别地，在回归设定中，$\rho_-(s_1,s_2)$ 满足

$$\begin{aligned} & \|X(\beta_1-\beta_2)\|_2^2 \ge \rho_-(s_1,s_2)\|\beta_1-\beta_2\|_2^2, \forall \beta_i : \|\beta_i\|_0 \le s_i \\ \iff \quad & (\beta_1-\beta_2)^T(\rho I - X^TX)(\beta_1-\beta_2) \le (\rho - \rho_-(s_1,s_2))\|\beta_1-\beta_2\|_2^2, \forall \beta_i : \|\beta_i\|_0 \le s_i. \end{aligned}$$

格拉姆矩阵 X^TX 中存在正的受限特征值，这意味着矩阵 $\rho I - X^TX$ 的受限特征值存在适当的上界。因此，在考虑 $\beta^{(t+1)}$ 的 ∞-范数误差时，使用矩阵 $\rho I - X^TX$ 似乎比使用 X^TX 更容易处理。

受此启发，给定 l、X 和 s_i，我们引入广义受限等距数 $\upsilon(s_1,s_2)$，使其满足

$$\|\rho(\beta_1-\beta_2) - X^T\{\nabla l_0(X\beta_1) - \nabla l_0(X\beta_2)\}\|_\infty \le (\rho-\upsilon)\|\beta_1-\beta_2\|_\infty, \text{ for all } \beta_i : \|\beta_i\|_0 \le s_i, \rho \ge \upsilon. \tag{A.38}$$

当 $l_0(X\beta) = \|X\beta - y\|_2^2/2$ 时，我们有 $\nabla l_0(X\beta_1) - \nabla l_0(X\beta_2) = X(\beta_1-\beta_2)$，且 $\rho(\beta_1-\beta_2) - X^T(\nabla l_0(X\beta_1) - \nabla l_0(X\beta_2)) = (\rho I - X^TX)(\beta_1-\beta_2)$。因此，(A.38) 可理解为设计矩阵在 ∞-范数情境下低相干性的一种变体。

定理 A.2. 对于由过程 (A.37) 生成的迭代序列，且 υ_t 表示由 (A.38) 定义的 $\upsilon(q_t,s)$，对于任意 $t \ge 0$，以下关于 $\beta^{(t+1)}$ 的递归逐坐标误差界成立：

$$\|\beta^{(t+1)} - \beta^*\|_\infty \le (1 - \frac{\upsilon_t + \eta_{t+1}}{\rho_{t+1}+\eta_{t+1}})\|\beta^{(t)} - \beta^*\|_\infty + \frac{\|X^T\epsilon\|_\infty}{\rho_{t+1}+\eta_{t+1}} + \frac{\eta_{t+1}\|\beta^*\|_\infty}{\rho_{t+1}+\eta_{t+1}} + \frac{1}{\sqrt{\vartheta_{t+1}-1}}\frac{\|\beta^{(t+1)}-\beta^*\|_2}{\sqrt{s}}.$$

证明。该证明与定理 2 的证明思路类似。首先，根据 $\Theta^{\#}$ 的定义，

$$\left\|(1+\bar{\eta}_{t+1})\beta^{(t+1)} - \beta^{(t)} + \frac{1}{\rho_{t+1}} X^T \nabla l_0(X\beta^{(t)})\right\|_\infty \le (1+\bar{\eta}_{t+1}) \min_{j \in \mathcal{J}(\beta_j^{(t+1)})} |\beta_j^{(t+1)}|$$

因此

$$\begin{aligned} \|(\rho_{t+1}+\eta_{t+1})\beta^{(t+1)} - \rho_{t+1}\beta^{(t)} + X^T(\nabla l_0(X\beta^{(t)}) - \nabla l_0(X\beta^*)) - X^T\epsilon\|_\infty \\ \le (\rho_{t+1}+\eta_{t+1}) \min_{j \in \mathcal{J}(\beta_j^{(t+1)})} |\beta_j^{(t+1)}|. \end{aligned}$$

写作

$$(\rho_{t+1}+\eta_{t+1})\beta^{(t+1)} - \rho_{t+1}\beta^{(t)} = (\rho_{t+1}+\eta_{t+1})(\beta^{(t+1)} - \beta^*) - \rho_{t+1}(\beta^{(t)} - \beta^*) + \eta_{t+1}\beta^*$$

并利用 ∞-范数的次可加性，我们得到

$$(\rho_{t+1}+\eta_{t+1})\|\beta^{(t+1)}-\beta^*\|_\infty \le \|\rho_{t+1}(\beta^{(t)}-\beta^*)-X^T(\nabla l_0(X\beta^{(t)})-\nabla l_0(X\beta^*))\|_\infty + \|X^T\epsilon\|_\infty+\eta_{t+1}\|\beta^*\|_\infty+(\rho_{t+1}+\eta_{t+1})\min_{j\in\mathcal{J}(\beta_j^{(t+1)})}|\beta_j^{(t+1)}|.$$

根据引理 A.3 的 (A.14) 以及 υ_t 的定义，我们得到

$$(\rho_{t+1}+\eta_{t+1})\|\beta^{(t+1)}-\beta^*\|_\infty \le (\rho_{t+1}-\upsilon_t)\|\beta^{(t)}-\beta^*\|_\infty + \|X^T\epsilon\|_\infty+\eta_{t+1}\|\beta^*\|_\infty+(\rho_{t+1}+\eta_{t+1})\frac{\|\beta^{(t+1)}-\beta^*\|_2}{\sqrt{q_{t+1}-s}}.$$

此外，通过应用 $(\rho_{t+1}+\eta_{t+1})\|(\beta^{(t+1)}-\beta^*)_{\mathcal{J}^*}\|_\infty \le (\rho_{t+1}-\upsilon_t)\|\beta^{(t)}-\beta^*\|_\infty+\|X^T\epsilon\|_\infty+\eta_{t+1}\|\beta^*\|_\infty$or，我们可以得到

$$\|(\beta^{(t+1)}-\beta^*)_{\mathcal{J}^*}\|_\infty \le (1-\frac{\upsilon_t+\eta_{t+1}}{\rho_{t+1}+\eta_{t+1}})\|\beta^{(t)}-\beta^*\|_\infty+\frac{\|X^T\epsilon\|_\infty}{\rho_{t+1}+\eta_{t+1}}+\frac{\eta_{t+1}\|\beta^*\|_\infty}{\rho_{t+1}+\eta_{t+1}},$$

应用 (A.15) 即可得到。 □

H. 基于预测信息准则的模型选择

尽管参数 q 作为真实模型支撑集大小的上界，通常可以根据领域知识直接指定，但本节将提出一种新的信息准则，用于调节 q，以便在有限样本中实现最佳预测性能。为涵盖第 IV-D 节中的应用，我们假设存在多响应。设 $Y\in\mathbb{R}^{n\times m}$、$X\in\mathbb{R}^{n\times p}$ 分别为响应矩阵和预测变量矩阵，$l_0(XB;Y)$ 为给定的损失函数。我们用 $\mathcal{J}(B)$ 表示 B 的行支撑集，并定义 $J(B)=|\mathcal{J}(B)|$。假设真实的 $B^*\in\mathbb{R}^{p\times m}$ 是行稀疏的，并令 $s^*=J(B^*)$。正文各节中考虑的问题对应于特殊情形$m=1$。为选择最佳（行）支撑集大小，我们建议在预测信息准则中将以下复杂度惩罚项加到损失函数上：

$$P(B)=J(B)m+J(B)\log\{ep/J(B)\}. \tag{A.39}$$

回顾矩阵语境下的 $D_2(A_1,A_2)=\|A_1-A_2\|_F^2/2$。

定理 A.3。设有效噪声 $E=-\nabla l_0(XB^*)$ 为均值为零、尺度有常数上界的次高斯分布，并且 $B^*\in\mathcal{M}$ 和 $B^*\neq 0$。假设存在常数 $\delta>0$ 和 $A_0\ge 0$，使得对于所有 $B,B'\in\mathcal{M}$，$(\Delta_{l_0}-\delta D_2)(XB,XB')+A_0(P(B)+P(B'))\ge 0$。那么，对于充分大的常数 A，使下式最小化的 ˆany B

$$l_0(XB;Y)+AP(B) \tag{A.40}$$

满足约束 $B\in\mathcal{M}$ 的解必须满足

$$\mathbb{E}\{\|X\hat{B}-XB^*\|_F^2\vee P(\hat{B})\}\lesssim ms^*+s^*\log(ep/s^*). \tag{A.41}$$

定理 A.3 不涉及任何正则化参数（如 q,λ），但达到了极小极大最优误差率 (A.41)。此外，对 (A.40) 的论证不需要无限样本量、设计相干性或信噪比条件。

当噪声分布具有离散参数 σ^2 时，定理 A.3 仍然适用，但 (A.40) 中的惩罚项会变为 $A\sigma^2P(B)$，其中包含一个未知因子。可以考虑使用初步尺度估计。不过，对于回归而言，一个颇具吸引力的结果是，可以绕过对 σ 的估计。我们给出预测信息准则的一种无尺度形式：

$$mn\log\{\|Y-XB\|_F^2\}+AP(B), \tag{A.42}$$

其中 A 是一个绝对常数。

定理 A.4. 令 $Y=XB^*+\mathcal{E}$，其中 $E=[\epsilon_{i,k}]$ 的各项相互独立、均值为零且服从次高斯(σ^2)分布，并且$\mathbb{E}\epsilon_{i,\,k}^2\gtrsim\sigma^2$，其中 σ^2 未知。定义 $l_0(XB;Y)=\|XB-Y\|_F^2$。假设真实模型不会过于复杂，即 $P(B^*)\le mn/A_0$，其中某个常数 $A_0>0$。令 $\delta(B)=AP(B)/(mn)$，其中 A 为正

ˆ 常数，满足 $A < A_0$，因此 $\delta(B^*) < 1$。那么，当 A_0 和 A 足够大时，任何使 $\log\, l_0(XB;\, Y) + \delta(B)$ 最小且满足 $\delta(B) < 1$ 的 B，都必须满足 $\mathrm{D}_2(X\hat{B}, XB^*) \lesssim \sigma^2\{s^*m + s^*\log(ep/s^*)\}$，其概率至少为 $1 - Cp^{-c}\exp\{-cm\} - C\exp(-cmn)$，其中 C、$c > 0$ 为某些常数。

$AP(B)$ 的更一般形式可以表示为“$\alpha_1 \times$ 自由度 $+ \alpha_2 \times$ 膨胀”，其中 α_1、α_2 为绝对常数。这两个定理可以通过修改 [28] 中定理 2 和定理 3 的证明来证明。为完整起见，我们在下面给出一些细节。请注意，尽管预测信息准则的对数形式被广泛使用，但还存在其他非渐近形式 [28]。事实上，证明中的一个关键技巧是将这些形式转换为分数形式的无尺度预测信息准则，这对于建立所需性质至关重要。

证明。我们首先在 vec $(\mathcal{E})$ 服从均值为 0、尺度为 σ 的次高斯分布这一假设下证明定理 A.3。根据 $\hat{B}$, $\mathbf{\Delta}_{l_0}(X\hat{B}, XB^*) + A\sigma^2 P(\hat{B}) \le A\sigma^2 P(B^*) + \langle \mathcal{E}, X\hat{B} - XB^* \rangle$ 的定义。类似于引理 A.2 的证明，我们可以证明，对于任意 a、b、$a' > 0$、$4b > a$ 和 $t > 0$,

$$\langle \mathcal{E}, XB - XB^* \rangle \le (\frac{2}{a} + \frac{2}{a'})\mathbf{D}_2(XB, XB^*) + a'\sigma^2 t + 4bL\sigma^2\{P(B^*) + P(B)\}, \forall B \in \mathbb{R}^{p\times m} \tag{A.43}$$

以至少 $1 - Cp^{-c}\exp(-cm)\exp(-ct)$ 的概率发生，其中 L、c、C 是正的常数。（令 L 为足够大的常数，并注意到当 $J \ge 1$ 时，$Jm + J\log(ep/J) \ge m + \log(ep)$ 成立，即可推导出该概率界；此外，如 (A.10) 中的并集界计算无需涵盖 $J = 0$ 的情形。）

现在，代入 $\hat{B}$ 代替 (A.43) 中的 B 并取期望，对于任意 a、b、$a' > 0$、$4b > a$，有

$$\begin{aligned}&\mathbb{E}\{\mathbf{\Delta}_{l_0}(X\hat{B}, XB^*) + A\sigma^2 P(\hat{B})\}\\ \le&\mathbb{E}\Big\{A\sigma^2 P(B^*) + (\frac{2}{a} + \frac{2}{a'})\mathbf{D}_2(X\hat{B}, XB^*) + ca'\sigma^2 + 4bL\sigma^2[P(B^*) + P(\hat{B})]\Big\}.\end{aligned}$$

结合正则性条件可得

$$\mathbb{E}\big\{(\delta - \frac{2}{a} - \frac{2}{a'})\mathbf{D}_2(X\hat{B}, XB^*) + (A - 4bL - C)P(\hat{B})\big\} \le (A + 4bL + C)\sigma^2 P(B^*) + ca'\sigma^2.$$

由于 $P(B^*) \ge c > 0$，选择满足 $(1/a + 1/a')(1 + 1/b') < \delta/2$、$4b > a$ 和 $A > 4bL + C$ 的常数，即可得出结论。

ˆ 接下来，我们证明定理 A.4。首先证明通过预测信息准则的无尺度形式的分数形式选出的 B： $l_0(XB;\, Y)/(1 - \delta(B))$，满足 $\delta(B) \le 1$。令 $h(B;\, A) = 1/\{mn - AP(B)\}$。ˆˆˆ 根据 B 的最优性， $l_0(XB;\, Y)h(B;\, A) \le l_0(XB^*; Y)h(B^*; A)$ 或

$$l_0(X\hat{B}; Y) - l_0(XB^*; Y) \le l_0(XB^*; Y)\Big(\frac{h(B^*; A)}{h(\hat{B}; A)} - 1\Big),$$

ˆ其中我们使用了 $h(B; A) > 0$。利用二次函数的 Bregman 散度，可得

$$\mathbf{D}_2(X\hat{B}, XB^*) \le l_0(XB^*; Y)\Big(\frac{h(B^*; A)}{h(\hat{B}; A)} - 1\Big) + \langle \mathcal{E}, X\hat{B} - XB^* \rangle. \tag{A.44}$$

根据 h 的定义和模型简约性假设，(A.44) 可写为

$$\begin{aligned}&\mathbf{D}_2(X\hat{B}, XB^*)\\ \le\ & l_0(XB^*; Y)\ \frac{AP(B^*) - AP(\hat{B})}{mn - AP(B^*)} + \langle \mathcal{E}, X\hat{B} - XB^* \rangle\\ =\ & \frac{1}{2}\frac{A\|\mathcal{E}\|_F^2}{mn\sigma^2 - A\sigma^2 P(B^*)}\sigma^2 P(B^*) - \frac{1}{2}\frac{A\|\mathcal{E}\|_F^2}{mn - AP(B^*)}\sigma^2 P(\hat{B}) + \langle \mathcal{E}, X\hat{B} - XB^* \rangle\\ \le\ & \frac{1}{2}\frac{A\|\mathcal{E}\|_F^2}{(1 - A/A_0)mn\sigma^2}\sigma^2 P(B^*) - \frac{1}{2}\frac{A\|\mathcal{E}\|_F^2}{mn\sigma^2}\sigma^2 P(\hat{B}) + \langle \mathcal{E}, X\hat{B} - XB^* \rangle.\end{aligned} \tag{A.45}$$

ˆ 随机项 $\langle \mathcal{E}, XB - XB^* \rangle$ 可类似地通过 (A.43) 进行界定：对于满足 $4b_1 > a_1$ 的任意 $a_1, b_1, a_2 > 0$，

$$\langle \mathcal{E}, X\hat{B} - XB^* \rangle \le 2(1/a_1 + 1/a_2)\mathbf{D}_2(X\hat{B}, XB^*) + (b_1)L_1\sigma^2\{P(\hat{B}) + P(B^*)\},$$

以至少 $1-Cp^{-c}\exp\{-cm\}$ 的概率，对某些 $c, C, L_1 > 0$ 成立。将其代入 (A.45) 得到

$$\begin{aligned}&\big(1-\frac{2}{a_1}-\frac{2}{a_2}\big)\mathbf{D}_2(X\hat{B}, XB^*)\\ \le\ &\frac{1}{2}\Big\{\frac{A\|\mathcal{E}\|_F^2}{(1-A/A_0)mn\sigma^2}+2b_1L_1\Big\}\sigma^2P(B^*)-\frac{1}{2}\Big\{\frac{A\|\mathcal{E}\|_F^2}{mn\sigma^2}-2b_1L_1\Big\}\sigma^2P(\hat{B}).\end{aligned}$$

由于 $\epsilon_{i,\ k}$ 相互独立且非退化，$c_1mn\sigma^2 \le \mathbb{E}\|\mathcal{E}\|_F^2 \le c_2mn\sigma^2$ 对于某些常数 c_1、$c_2 > 0$成立。令γ 为满足 $0 < \gamma < 1$的某个常数。在 $\mathcal{E} = \{c_1(1-\gamma)mn\sigma^2 \le \|\mathcal{E}\|_F^2 \le c_2(1+\gamma)mn\sigma^2\}$上，我们有

$$\frac{A\|\mathcal{E}\|_F^2}{(1-A/A_0)mn\sigma^2} \le \frac{c_2(1+\gamma)A_0A}{A_0-A} \quad\text{and}\quad \frac{A\|\mathcal{E}\|_F^2}{mn\sigma^2} \ge c_1(1-\gamma)A.$$

关于该事件发生的概率，我们写作 $\|\mathcal{E}\|_F^2 = \text{vec}\,(\mathcal{E})A\,\text{vec}(\mathcal{E})^T$，其中 $A = I \in \mathbb{R}^{nm\times nm}$，并使用 Hanson-Wright 不等式对其进行界定。事实上，由于 $\text{Tr}(A) = mn, \|A\|_2 = 1, \|A\|_F = \sqrt{mn}$，$\mathcal{E}$的补事件发生的概率至多为 $C'\exp\{-c'mn\}$。

现在，当 A_0、A、a_1、a_2、b_1 足够大，使得 $(1/a_1 + 1/a_2) < 1/2$, $4b_1 > a_1$, $A > 2b_1L_1/\{c_1(1-\gamma)\}$且满足 $A_0 > A$时，我们可以得到分数形式所需的预测误差率。最后，根据对于任意 $0 \le \delta < 1$均有$1/(1-\delta) \ge \exp(\delta) \ge 1/(1-\delta/2)$这一事实，对数形式也具有相同的误差率（更多细节请参见 [28]）。 □

I. 更多实现细节

slow kill 的实现极其简单，下面给出总结。为便于说明，我们基于定理 6 及其讨论定义一个 $\bar{\eta}$ function,

$$\bar{\eta}(q_+, \rho_+) = \begin{cases}\frac{1}{2\sqrt{q_+/\bar{s}-1}}, & \text{if } q_+ > 2q \text{ and } q \ge n/2\\ \frac{\eta_0}{\rho_+}, & \text{if } q_+ \le 2q,\\ \frac{\eta_0}{\rho_+} \wedge \frac{1}{2\sqrt{q_+/\bar{s}-1}}, & \text{otherwise,}\end{cases} \tag{A.46}$$

其中 $\bar{s} = q \wedge nL^2/\log(ep) \ge s$，其中 L 为 ∇l_0 的利普希茨参数，而 η_0 是用户定义的参数。（与q类似，η_0 是可由用户自定义的正则化参数。）我们还定义了一个 β 函数

$$\beta(q_+, \rho_+, \beta^-) = \Theta^{\#}\Big\{\beta^- - \rho_+^{-1}X^T\nabla l_0(X\beta^-; y); q_+, \bar{\eta}(q_+, \rho_+)\Big\}, \tag{A.47}$$

基于 (32)。（通常应包含一个不受正则化约束的截距项（例如 β_1）。我们可以在设计矩阵中添加一列全 1，并重新定义 (A.47) 中的 $\Theta^{\#}$，以保留第一个元素，并对剩余的子向量执行分位数阈值处理。）

回顾试探步长 ρ 的线搜索准则：

$$(\rho\mathbf{D}_2 - \mathbf{\Delta}_l)(\beta(q_{t+1}, \rho, \beta^{(t)}), \beta^{(t)}) \ge 0 \tag{A.48}$$

or

$$\begin{aligned}\frac{\rho}{2}\|\beta(q_{t+1}, \rho, \beta^{(t)}) - \beta^{(t)}\|_2^2 \ge\ & l_0(X\beta(q_{t+1}, \rho, \beta^{(t)})) - l_0(X\beta^{(t)})\\ & - \langle\nabla l_0(X\beta^{(t)}), X\beta(q_{t+1}, \rho, \beta^{(t)}) - X\beta^{(t)}\rangle.\end{aligned}$$

然后，算法可以总结如下。

输入：X, y、分位数参数序列 $q_t \to q \in [p]$、目标 ℓ_2-收缩 $\eta_0 \ge 0$。初始化：$\beta^{(0)}, \rho_0$（分别设为 0 和 $L\|X\|_2^2$）。对每个 q_{t+1} $(t \ge 0)$，执行以下操作：a) 根据准则 (A.48) 通过线搜索找到 ρ_{t+1}。b) 根据 (A.47) 执行 $\beta^{(t+1)} \leftarrow \beta(q_{t+1}, \rho_{t+1}, \beta^{(t)})$。

我们还可以偶尔添加一个压缩操作作为步骤 c)：$X \leftarrow X_{\mathcal{J}(\beta^{(t+1)})}$（例如，当 q_{t+1} 达到 $p/2^k$，且 k 大于某个 k_0时）。此外，当 q_{t+1} 达到 q 且 $\beta^{(t+1)}$ 的稀疏模式趋于稳定后，一个

可使用经典优化方法求解平滑问题，以获得最终估计中的非零项。至于步骤 a)，可以使用许多标准线搜索方法，例如回溯法 [45]。我们采用带热启动的自适应搜索。具体而言，给定 $\alpha \in (0,1)$，我们从 $\rho \leftarrow \rho_t$开始；若 (A.48) 对于$\beta(q_{t+1}, \rho, \beta^{(t)})$成立，则设定 $\rho \leftarrow \alpha\rho$，否则设定 $\rho \leftarrow \rho/\alpha$，重复此过程，直到某个足够小的 ρ_{t+1}使 (A.48) 成立，而 $\alpha\rho_{t+1}$不成立。实践中，最好限制搜索次数（M）。实现时，我们使用 $\alpha = 0.5, M = 5$。

参考文献

[1] P. J. Bickel、Y. Ritov 和 A. B. Tsybakov，“Lasso 与 Dantzig 选择器的同步分析”，The Annals of Statistics，第 37 卷，第 4 期，第 1705–1732 页，2009年 08 月。[2] P. C. Bellec、G. Lecu´e 和 A. B. Tsybakov，“Slope 与 lasso 相遇：改进的预言机界与最优性”，The Annals of Statistics，第 46 卷，第 6B 期，第 3603–3642 页，2018年。[3] J. Fan 和 J. Lv，“超高维特征空间的确定性独立筛选”，Journal of the Royal Statistical Society: Series B (Statistical Methodology)，第 70 卷，第 5 期，第 849–911 页，2008 年。[4] C.-H. Zhang，“最小最大凹惩罚下近乎无偏的变量选择”，The Annals of Statistics，第 38 卷，第 2 期，第 894–942 页，2010年 04 月。[5] P. B¨uhlmann 和 B. Yu，“使用 L_2 损失函数的提升法：回归与分类”，Journal of the American Statistical Association，第 98 卷，第 462 期，第 324–339 页，2003 年。[6] D. Needell 和 J. A. Tropp，“CoSaMP：从不完整且不准确的样本中迭代恢复信号”，Applied and Computational Harmonic Analysis，第 26 卷，第 3 期，第 301–321 页，2009年。[7] T. Zhang，“RIP 条件下使用正交匹配追踪进行稀疏恢复”，IEEE Transactions on Information Theory，第 57 卷，第 9 期，第 6215–6221 页，2011 年。[8] ——，“ 特征选择的多阶段凸松弛 ”，Bernoulli，第 19 卷，第 5B 期，第 2277–2293 页，2013年。[9] T. Zhao、H. Liu 和 T. Zhang，“稀疏学习的路径式坐标优化：算法与理论”，The Annals of Statistics，第 46 卷，第 1 期，第 180–218 页，2018年。[10] E. J. Candes 和 T. Tao，“通过线性规划解码”，IEEE Transactions on Information Theory，第 51 卷，第 12 期，第 4203–4215 页，2005年。[11] Y. She、J. Wang、H. Li 和 D. Wu，“用于超分辨率稀疏谱选择的组迭代谱阈值处理”，IEEE Transactions on signal Processing，第 61 卷，第 24 期，第 6371–6386 页，2013年。[12] H. Zou 和 T. Hastie，“通过弹性网进行正则化与变量选择”，Journal of the Royal Statistical Society: Series B (Statistical Methodology)，第 67 卷，第 2 期，第 301–320 页，2005年。[13] Y. She、Z. Wang 和 J. Shen，“借助渐进分位数提升抗离群值能力：快速算法与理论研究”，Journal of the American Statistical Association，第 117 卷，第 1282–1295 页，2022 年。[14] T. Blu mensath 和 M. E. Davies，“稀疏逼近的迭代阈值处理”，Journal of Fourier analysis and Applications，第 14 卷，第 5-6 期，第 629–654 页，2008 年。[15] ——，“压缩感知的迭代硬阈值处理”，Applied and Computational Harmonic Analysis，第 27 卷，第 3 期，第 265–274 页，2009年。[16] G. Raskutti、M. J. Wainwright 和 B. Yu，“ ℓ_q球上的高维线性回归估计的极小极大速率”，IEEE Transactions on Information Theory，第 57 卷，第 10 期，第 6976–6994 页，2011年。[17] H. Liu 和 R. F. Barber，“介于硬阈值与软阈值之间：最优迭代阈值算法”，Information and Inference: A J ournal of the IMA，第 9 卷，第 4 期，第 899–933 页，2020 年。[18] P. Jain、A. Tewari 和 P. Kar，“关于高维 m 估计的迭代硬阈值方法”，载于第 28 届神经信息处理系统年会（NIPS）论文集，2014 年。[19] Z. Wang、H. Liu 和 T. Zhang，“稀疏非凸学习问题的最优计算与统计收敛速率”，Annals of statistics，第 42 卷，第 6 期，第 2164 页，2014 年。[20] P.-L. Loh 和 M. J. Wainwright，“具有非凸性的正则化 M 估计量：局部最优解的统计与算法理论”，The Journal of Machine Learning Research，第 16 卷，第 1 期，第 559–616 页，2015 年。[21] Y. She，“ 关于 Θ 估计量的有限样本分析 ”，Electronic Journal of Statistics，第 10 卷，第 2 期，第 1874–1895 页，2016 年。[22] Y. She、Z. Wang 和 J. Jin，“非光滑非凸统计学习的广义 Bregman 替代算法分析”，The Annals of Statistics，第 49 卷，第 6 期，第 3434–3459 页，2021年。[23] M. Wainwright，《高维统计：非渐近视角》。剑桥大学出版社，2019年。

[24] L. M. Bregman, 《寻找凸集公共点的松弛法及其在凸规划问题求解中的应用》，苏联计算数学与数学物理，第 7 卷，第 200–217 页，1967年。[25]C. Zhang、Y. Jiang 和 Y. Chai, 《用于高维回归与分类的惩罚Bregman散度》，Biometrika，第 97 卷，第 3 期，第 551–566 页，2010年。[26] Y. She, 《用于特征提取的降秩向量广义线性模型》，Statistics and Its Interface，第 6 卷，第 197–209 页，2013年。[27] B. Efron、T. Hastie、I. Johnstone 和 R. Tibshirani, 《最小角回归》，The Annals of statistics，第 32 卷，第 2 期，第 407–499 页，2004年。[28] Y. She 和 H. Tran, 《关于稀疏降秩回归的交叉验证》，Journal of the Royal Statistical Society: Series B，第 81 卷，第 145–161 页，2019年。[29] R. Tibshirani, 《通过 LASSO 进行回归收缩与变量选择》，Journal of the Royal Statistical Society: Series B (Statistical Methodology)，第 267–288 页，1996年。[30] J. Fan 和 R. Li, 《通过非凹惩罚似然进行变量选择及其预言机性质》，Journal of the American Statistical Association，第 96 卷，第 456 期，第 1348–1360 页，2001年。[31] T. Blumensath 和 M. E. Davies, 《归一化迭代硬阈值：稳定性与性能保证》，IEEE Journal of Selected Topics in Signal Processing，第 4 卷，第 2 期，第 298–309 页，2010年。[32] D. Kim、S. Lee 和 S. Kwon, 《非凸惩罚估计的统一算法：ncpen 软件包》，arXiv 预印本 arXiv:1811.05061，2018 年。[33] Guyon 、 Isabelle 、 G. Steve 、 B.-H. Asa 和 G. Dror, 《 NIPS 2003特征选择挑战赛的结果分析》，载于 Advances in Neural Information Processing Systems，第 545–552 页，2005年。[34] J. Friedman、T. Hastie 和 R. Tibshirani, 《通过坐标下降求解广义线性模型的正则化路径》，Journal of Statistical Software，第 33 卷，第 1 期，第 1–22 页，2010年。[35] ——, 《加性逻辑回归：提升法的统计学观点（附讨论及作者回应）》，The annals of statistics，第 28 卷，第 2 期，第 337–407 页，2000年。[36] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay, 《Scikit-learn：Python 中的机器学习》，Journal of Machine Learning Research，第 12 卷，第 2825–2830 页，2011年。[37] J. Ge、X. Li、H. Jiang、H. Liu、T. Zhang、M. Wang 和 T. Zhao, 《Picasso：用于 R 和 Python 中高维数据分析的稀疏学习库》，Journal of Machine Learning Research，第 20 卷，第 44–1 页，2019年。[38] B. C. Feltes、E. B. Chandelier、B. I. Grisci 和 M. Dorn, 《CuMiDa：一个经过广泛整理的微阵列数据库，用于基准测试和测试癌症研究中的机器学习方法》，Journal of Computational Biology，第 26 卷，第 4 期，第 376–386 页，2019年。[39] M. Mishali 和 Y. C. Eldar, 《盲多频带信号重建：面向模拟信号的压缩感知》，IEEE Transactions on Signal Processing，第 57 卷，第 3 期，第 993–1009 页，2009年。[40] Z. Song、H. Qi 和 Y. Gao, 《面向毫米波的实时多吉赫兹亚奈奎斯特频谱感知系统》，载于第 3 届 ACM毫米波网络与感知系统研讨会论文集，2019年，第 33–38 页。[41] J. A. Tropp、A. C. Gilbert 和 M. J. Strauss, 《通过贪婪追踪实现同时稀疏逼近》，载于 Proceedings.(ICASSP'05). IEEE国际声学、语音与信号处理会议，2005年，第 5 卷。IEEE，2005年，第 v–721 页。[42] H. Qi、X. Zhang 和 Y. Gao, 《宽带空白频谱上的低复杂度子空间辅助压缩频谱感知》，IEEE Transactions on Vehicular Technology，第 68 卷，第 12 期，第 11 762–11777 页，2019年。[43] Y. Gao、Z. Song、H. Zhang、S. Fuller、A. Lambert、Z. Ying、P. M¨ah¨onen、Y. Eldar、S. Cui、M. D. Plumbley 等, 《亚奈奎斯特频谱感知与学习挑战赛》，Frontiers of Computer Science，第 15 卷，第 4 期，第 1–5 页，2021年。[44] Y. She 和 K. Chen, 《稳健降秩回归》，Biometrika，第 104 卷，第 3 期，第 633–647 页，2017年。[45] S. Boyd 和 L. Vandenberghe, 《凸优化》。剑桥：剑桥大学出版社，2004年。