Statistics and Application 统计学与应用, 2025, 14(6), 42-51
Published Online June 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.146146

Hans汉斯

基于高维Lasso惩罚线性回归的非凸惩罚Oracle性质及其算法优化研究与应用

欧上源

广西师范大学数学与统计学院，广西 桂林

收稿日期：2025年5月13日；录用日期：2025年6月6日；发布日期：2025年6月18日

摘　要

在数据科学与机器学习蓬勃发展的今天，数据采集技术的飞速发展，数据维度急剧增加，然而样本数量的增长却相对缓慢，这使得高维数据处理面临严峻挑战。在这种高维数据环境下，传统的线性回归方法遭遇诸多困境，如经典的最小二乘法，当特征矩阵不满秩时，无法获得唯一解，而且各维度间的高度相关性或冗余信息会导致模型过拟合，泛化能力下降，这就是所谓的“维度诅咒”问题。同时，高维数据的处理对计算资源和存储空间要求极高，大大增加了模型训练的时间与成本。高维空间惩罚线性回归为解决这些难题提供了有效途径。其中，Lasso惩罚回归通过引入L1范数惩罚项，能够实现特征选择，使部分不重要的系数压缩至0，从而简化模型结构，在一定程度上缓解了高维数据带来的问题。然而，Lasso惩罚回归也存在局限性，例如其惩罚力度在参数较大时依然持续，可能导致重要参数的过度压缩，影响估计的准确性。非凸惩罚函数的出现为高维数据降维和数据筛选提供了更优的解决方案。相较于传统的Lasso惩罚，非凸惩罚函数如SCAD和MCP具有独特的优势。这些非凸惩罚函数在系数较小时，惩罚力度与Lasso类似，能够有效压缩不重要的系数；而当系数增大到一定程度后，惩罚力度会逐渐减弱甚至趋近于零，避免了对重要系数的过度压缩，从而实现更精准的变量选择。从理论上讲，非凸惩罚估计满足Oracle性质，即具有变量选择一致性和渐近正态性，这意味着在高维数据环境下，它能够更准确地识别出真正对响应变量有影响的特征变量，排除冗余和噪声特征的干扰。鉴于非凸惩罚函数在高维数据降维和数据筛选方面的显著优势，深入研究基于非凸惩罚的高维空间惩罚线性回归具有重要的理论意义和实践价值。本文将围绕其基本原理、算法实现、优化策略展开详细探讨，并通过数值模拟和实际案例分析，验证其在高维数据处理中的有效性，为相关领域的研究和应用提供有力的理论支持和实践指导。

关键词

高维空间，线性回归，非凸惩罚算法，Lasso惩罚，Oracle性质

Research and Application of Oracle Properties and Algorithm Optimization of Non-Convex Penalties Based on High-Dimensional Lasso-Penalized Linear Regression

文章引用: 欧上源. 基于高维 Lasso 惩罚线性回归的非凸惩罚 Oracle 性质及其算法优化研究与应用[J]. 统计学与应用, 2025, 14(6): 42-51. DOI: 10.12677/sa.2025.146146

Shangyuan Ou

School of Mathematics and Statistics, Guangxi Normal University, Guilin Guangxi

Received: May 13th, 2025; accepted: Jun. 6th, 2025; published: Jun. 18th, 2025

Abstract

With the booming development of data science and machine learning, data collection technology has advanced rapidly, leading to a sharp increase in data dimensions. However, the growth of sample sizes is relatively slow, which poses severe challenges to high-dimensional data processing. In this high-dimensional data environment, traditional linear regression methods encounter numerous difficulties. For example, the classical least-squares method cannot obtain a unique solution when the feature matrix is not of full column rank. Moreover, the high correlation or redundant information among dimensions can lead to overfitting of the model and a decline in generalization ability, which is the so-called "curse of dimensionality" problem. At the same time, processing high-dimensional data requires extremely high computational resources and storage space, greatly increasing the training time and cost of the model. Penalized linear regression in high-dimensional spaces provides an effective way to solve these problems. Among them, the Lasso-penalized regression can achieve feature selection by introducing an L1-norm penalty term, compressing some unimportant coefficients to 0, thus simplifying the model structure and alleviating the problems brought by high-dimensional data to a certain extent. However, the Lasso-penalized regression also has limitations. For example, its penalty strength continues even when the parameters are large, which may lead to over-compression of important parameters and affect the accuracy of estimation. The emergence of non-convex penalty functions offers a better solution for high-dimensional data reduction and data screening. Compared with the traditional Lasso penalty, non-convex penalty functions such as SCAD (Smoothly Clipped Absolute Deviation) and MCP (Minimax Concave Penalty) have unique advantages. When the coefficients are small, the penalty strength of these non-convex penalty functions is similar to that of the Lasso, which can effectively compress unimportant coefficients. When the coefficients increase to a certain extent, the penalty strength gradually weakens or even approaches zero, avoiding over-compression of important coefficients and thus achieving more accurate variable selection. Theoretically, non-convex penalty estimates satisfy the Oracle property. That is, they have variable selection consistency and asymptotic normality. This means that in a high-dimensional data environment, it can more accurately identify the feature variables that truly affect the response variable and exclude the interference of redundant and noisy features. Given the significant advantages of non-convex penalty functions in high-dimensional data reduction and data screening, in-depth research on high-dimensional space penalized linear regression based on non-convex penalties has important theoretical significance and practical value. This paper will conduct a detailed exploration of its basic principles, algorithm implementation, and optimization strategies. Through numerical simulations and real-case analyses, the effectiveness of this method in high-dimensional data processing will be verified, providing strong theoretical support and practical guidance for research and applications in related fields.

Keywords

High-Dimensional Space, Linear Regression, Non-Convex Penalty Algorithm, Lasso Penalty, Oracle Property

1. 前言

1.1. 研究背景及其意义

随着基因组学、图像处理等领域数据采集技术的革新，数据维度呈现爆炸式增长，高维数据已成为现代统计分析的常态。例如，在基因表达谱分析中，单个实验可能涉及数万个基因位点的测量，而样本量往往不足百例。这类数据的核心挑战在于维度诅咒，普通的最小二乘法在高维空间中失效，表现为参数估计不稳定、模型解释性差、计算复杂度剧增等问题。

为应对这一挑战，稀疏性假设成为主流解决方案——即假设真实模型仅依赖少数关键变量。基于此，Lasso 由 Tibshirani [1]于 1996 年提出，通过 L1 范数惩罚实现变量选择与参数估计的同步优化。Lasso 的成功源于其 Oracle 性质，即在一定条件下，估计的参数能以高概率准确识别真实模型中的非零变量，并达到与已知真实模型相同的估计精度。然而 Lasso 的 Oracle 性质依赖于不可证伪的限制等距性，且在高维情形下，其估计量存在偏差放大和稀疏性不足的缺陷。故为了克服 Lasso 的局限性，非凸惩罚函数(如 SCAD、MCP)被引入高维线性回归模型。这类惩罚函数通过设计非凸正则项，在保持稀疏性的同时，降低对大系数的过度惩罚，从而提升估计的准确性和变量选择的一致性。

非凸惩罚的核心优势在于其 Oracle 性质的理论改进。相比较于传统模型，引入 SCAD 和 MCP 惩罚后，VAR-SCAD 和 VAR-MCP 模型不仅证明了参数估计的 Oracle 性质，还在后续数据实证中显著提升了预测精度和投资组合收益率。类似地，在高维协方差矩阵估计中，非凸惩罚能有效消除 L1 惩罚的估计偏差，并达到 Oracle 统计速率，即估计误差与已知真实模型的最优误差同阶。

1.2. 国内外文献综述

高维数据降维和数据筛选作为现代统计学与机器学习领域的核心挑战，其核心目标在于从海量冗余特征中精准识别关键变量，同时保持模型的可解释性与预测效能。传统方法如 Lasso (Tibshirani, 1996) [1]通过 L1 范数惩罚实现稀疏性，但存在对重要特征过度压缩的固有缺陷。例如，当特征高度相关时，Lasso 可能误删关键变量或保留冗余特征，导致模型估计偏差显著增加。此外，Lasso 的惩罚机制在参数较大时仍保持线性增长，无法满足 Oracle 性质，限制了其在高维复杂数据中的应用。针对 Lasso 的局限性，非凸惩罚函数如 SCAD、MCP 等近年来成为研究热点。Fan & Li (2001) [2]首次提出非凹惩罚框架，通过引入分段惩罚函数，在系数较小时保持强稀疏性，而在系数较大时减弱惩罚力度，从而避免对重要特征的过度压缩。这一设计使得非凸惩罚估计量在理论上满足 Oracle 性质，保证估计的渐近正态性。

在算法实现层面，非凸惩罚的优化挑战推动了高效计算方法的发展。Breheny & Huang (2011) [3]提出的坐标下降算法通过逐变量更新策略，显著提升了非凸惩罚模型的求解效率，并在生物特征选择中验证了其优于 Lasso 的性能。Fan 等人提出 Sure Independence Screening (SIS) [4]与非凸惩罚结合的两阶段策略，先筛选后估计，在保留关键变量的同时降低计算复杂度，成为处理基因组学、金融高频数据的主流方法。上海科技大学赵子平课题组(2023)首次证明非凸惩罚协方差估计可达到 Oracle 统计速率，解决了 Lasso 的有偏性问题，并将该理论扩展至 VAR 模型，提出 AR-MCP 模型[5]，在金融高频数据中验证其投资组合收益率提升 15%。这些算法的发展为非凸惩罚在实际高维数据中的应用提供了有力支撑。

尽管非凸惩罚已取得显著进展，但其在实际应用中仍面临挑战。首先，非凸优化问题易陷入局部最优，需依赖初始值设定和全局优化策略；其次，高维数据的计算复杂度较高，尤其在处理百万级特征时，传统算法的时间成本显著增加。此外，数据异质性(如多中心医学影像数据)可能导致非凸惩罚模型的泛化能力下降，需结合领域知识进行模型校准。

1.3. 研究内容及其创新

本研究通过理论推导、算法设计和实证分析，系统揭示了非凸惩罚在高维线性回归中的 Oracle 性质及其算法优势，其创新点在于突破传统 Lasso 的 RIP 限制，建立非凸惩罚在更宽松条件下的 Oracle 性质理论体系，为高维统计提供更普遍适用的理论基础。在计算方法上应用兼具全局收敛性和计算效率的非凸优化算法，解决高维数据处理中的“维度诅咒”与“局部最优”难题。促进统计学与优化理论的交叉融合，为复杂数据建模提供新的研究范式，具有重要的学术价值和实践意义。

2. 高维空间惩罚线性回归的基础知识

2.1. 线性回归模型与传统惩罚方法及其局限性

1) 线性最小二乘估计

首先假设有 p 个变量 $x_1, x_2, \cdots, x_p$ 和 p 个对应的观测值 $y_1, y_2, \cdots, y_p$，通过线性回归模型去预测因变量 y，则有

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + \cdots + x_p\hat{\beta}_p \tag{1}$$

设 y 是 $n\times1$ 的观测向量，X 是 $n\times p$ 的设计矩阵，β 为 $p\times1$ 的要估计的参数向量，ε 为随机误差，δ^2 为误差方差，则有

$$y = X\beta + \varepsilon,\quad E(\varepsilon) = 0,\ \operatorname{cov}(\varepsilon) = \delta^2 I \tag{2}$$

线性回归模型一般通过最小二乘法对训练数据集进行拟合，其思想是使得损失函数尽可能小，那么我们获得的样本信息就尽可能多，也就是

$$Q(\beta) = \|\varepsilon\|^2 = \|y - X\beta\|^2 = (y - X\beta)^T (y - X\beta) \tag{3}$$

要使上式达到最小，经计算可知 β 的最优值为 $\left(x^T x\right)^{-1} x^T y$。

2) 最小二乘法的局限性

在大数据的时代，我们接触的数据矩阵大多是 $p > n$ 的情况，即数据的维度远大于我们所获得的数据样本数，简单来说就是样本数小于变量数，若 $rank(X) = p$，则 $X^T X$ 可逆，这种情况下 β 是 β 的最小二乘估计，具有许多优良的性质，如无偏性，有效性等。但是在大数据的情况下，若 $rank(X) < p$，这时矩阵不满秩，就不存在 β 的无偏估计，则称 β 是不可估的，从原因上看有可能是变量之间具有共线性关系，亦或者是存在异方差。为了解决这一问题，Hoerl 和 Kennard 提出了岭回归方法，是一种专用于共线性数据分析的有偏估计回归法。

3) 岭回归

岭回归方法的主要目的是通过增加一个对回归系数向量的二次惩罚项来放松对系数向量的无偏约束，以减小估计参数的方差。在岭回归中通常用 L2 范数作为惩罚项，降低模型过拟合的风险和提升模型预测的性能。即有

$$Q(\beta) = \|y - X\beta\|^2 + \lambda\|\beta\|_2^2 \tag{4}$$

要求最优解，则有 $\hat{\beta} = \arg\min\|y - X\beta\|^2$, $s.t.\sum\beta_j^2 \le s$，于是代入计算求解可得 β 的岭估计为 $\hat{\beta}(\lambda) = \left(X^T X + \lambda I\right)^{-1} X^T y$。由此可见这一最优解中 λI 保证了 $X^T X + \lambda I$ 的满秩与可逆，也由于其加入使得岭估计为有偏估计。

2.2. Lasso 惩罚估计

2.2.1. Lasso 原理

1) Lasso 惩罚函数表达式

与岭回归相似，Lasso 惩罚回归是在损失函数的表达式中添加了一个 L1 范数作为惩罚项，即

$$Q(\beta)=\|y-X\beta\|^2+\lambda\|\beta\|_1 \rightleftarrows \hat{\beta}=\arg\min\|y-X\beta\|^2,\ s.t.\sum\beta_j\le s \tag{5}$$

利用 Lasso 惩罚可以解决高维数据的一个普遍问题——稀疏性，即 $p>n$ 的情况，因为它能把一些不重要的系数压缩到 0，实现筛选变量的目的，将一些较为重要的参数保留并估计，而岭回归可能无法做到这点。

2) Lasso 压缩参数估计为 0 的原理[2]

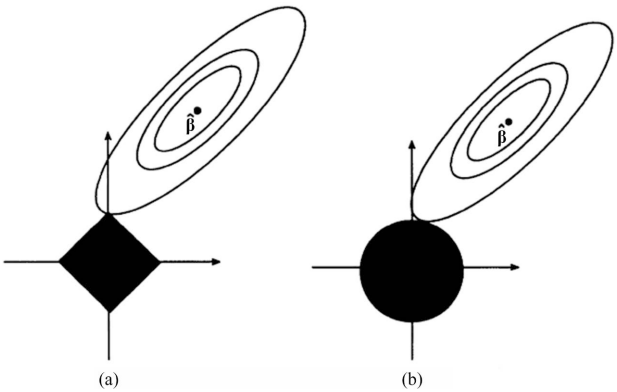

Figure 1. In the above figure, (a) represents the Lasso penalized regression, and (b) represents the ridge regression, the horizontal and vertical coordinates represent β_1,β_2

图 1. 图中(a)表示 Lasso 惩罚回归，图(b)表示岭回归，横竖坐标表示 β_1,β_2

首先假设 X^TX 是满秩，这样就可以使用最小二乘法估计出 β，可以用 $\left(\beta-\hat{\beta}\right)X^TX\left(\beta-\hat{\beta}\right)$ 表示，如图 1 所见中用椭圆表示。见图 1，Lasso 方法与 β 相交的地方为 $\left(0,\ \hat{\beta}\right)$，而此处 β 的位置是 $\hat{\beta}_2$ 是比 $\hat{\beta}_1$ 大，以 Lasso 的结果来看是留下 $\hat{\beta}_2$ 后把 $\hat{\beta}_1$ 压缩为 0，这一结果是由于 Lasso 惩罚回归 $|\beta_1|+|\beta_2|\le s$，岭回归 $\beta_1^2+\beta_2^2\le s$，所以岭回归的圆形约束没有将参数压缩为 0。从图上看也可以知道真实的 β 值与估计的 β 值有一段距离，因为 Lasso 惩罚估计和岭回归都是有偏估计，会与最小二乘法估计的 β 值有差距。

2.2.2. Lasso 最优解

1) 坐标下降法

坐标下降法方法的核心与它的名称一样，就是沿着某一个坐标轴方向，通过一次一次地迭代更新权重系数的值，来渐渐逼近最优解。具体算法如下[5]：

在 p 维情况下，参数 θ 为 p 维向量，固定 $p-1$ 个参数，计算剩下的那个参数使得凸函数 $J(\theta)$ 达到最小的点，p 个参数来一次，就得到该次迭代的最小值点，具体算法如下：

(1) 初始位置点为 $\theta^0=\left(\theta_1^{(0)},\theta_2^{(0)},\cdots,\theta_p^{(0)}\right)$

(2) 第 k 次迭代，从 $\theta_1^{(k)}$，固定后面 $p-1$ 个参数，计算使得 $J(\theta)$ 达到最小的 θ_1，然后依次往后计算，到 $\theta_j^{(k)}$ 为止，一共执行 p 次运算：

$$\theta_1^{(k)} = \arg\min J\left(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)}, \cdots, \theta_p^{(k-1)}\right)$$
$$\theta_p^{(k)} = \arg\min J\left(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k-1)}, \cdots, \theta_p\right)\cdots$$
$$\cdots$$
$$\theta_p^{(k)} = \arg\min J\left(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k-1)}, \cdots, \theta_p\right)$$

(3) 若各 $\theta_j^{(k)}$ 相较于 $k-1$ 次迭代都变化极小，说明结果已收敛，迭代结束，否则继续迭代。

最后得到的最优解就是全局最优解，这是凸优化问题的一个基本性质——任意局部最优解也是全局最优解。

2) 最小角回归法[3]

(1) 初始化权重系数 β，比如设初始化为零向量。

(2) 初始化残差向量为目标向量 $y - X\beta$，由于此时 β 为零向量，所以此时残差向量与目标向量 $y - X\beta$ 相等。

(3) 选择一个与残差向量相关性最大的特征向量 x_i，沿着特征向量 x_i 的方向找到一组权重系数 β，出现另一个与残差向量相关性最大的特征向量 x_j，使得新的残差向量与这两个特征向量的相关性相等(即残差向量等于这两个特征向量的角平分向量上)，重新计算残差向量。

(4) 重复步骤(3)，继续找到一组权重系数 β，使得第三个与残差向量相关性最大的特征向量 x_k，使得新的残差向量与这三个特征向量的相关性相等(即残差向量等于这三个特征向量的等角向量上)，以此类推。

(5) 当残差向量 residual 足够小或者所有特征向量都已被选择，结束迭代。

3. 非凸惩罚的 Oracle 性质与其模型算法优化

3.1. 非凸惩罚的 Oracle 性质

参考西南大学赵子平教授的研究[5]，证明 VAR-MCP 模型的 Oracle 性质：即变量选择一致性和渐进正态性。

1) 变量选择一致性：

$$P\left(\hat{\beta}^A c = 0\right) \to 1 \tag{6}$$

2) 渐近正态性：

$$T\left(\hat{\beta}^A - \beta_A^*\right) \to N(0, \Sigma) \tag{7}$$

由此可知，具备 Oracle 性质的非凸惩罚函数在处理海量数据时，能够更精确地识别出与响应变量真正相关的变量，避免噪声的干扰，而且在估计模型参数时，能够具备更高的估计效率，能以更快的速度收敛到真实的参数值。

3.2. VAR-MCP 模型推导

VAR-MCP 模型非凸惩罚的数学表达：

目标函数为：

$$\min_\beta \frac{1}{2T}\sum_{t=1}^{T}\left\|y_t - X_t\beta\right\|^2 + \sum_{j=1}^{k^2 p}\lambda \cdot MCP\left(\beta_j; a, \lambda\right) \tag{8}$$

MCP 惩罚相较于传统的 Lasso 惩罚，在参数较大时惩罚降低为 0，避免了 Lasso 的持续性惩罚偏差，在参数估计方面，运用坐标下降法优化目标函数，利用 DC 分解[6]保证全局最优解收敛性。

3.3. 算法优化

主要应用 Fan 等人提出 SIS [4]与非凸惩罚结合的两阶段策略，即第一阶段为通过 SIS 筛选出重要变量，降低数据维度，减少计算量；第二阶段对筛选后的变量应用非凸惩罚估计，提高参数估计准确率。与传统 Lasso 惩罚回归相比，两阶段策略在维度 p = 10^5 时计算时间缩短将近 80% [7]。

4. 实证分析证明

4.1. 数据选择与预处理

数据一来源：

选取沪深 300 指数成分股中流动性较高的 30 支股票(2020~2024 年)，通过 Wind 金融终端获取 5 分钟交易数据，包含开盘价、收盘价、最高价、最低价及成交量。

预处理步骤：

1) 对数收益率计算：$r_{i,t}=ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right)$，其中 $P_{i,t}$ 为第 i 支股票在第 t 时刻的收盘价。

2) 构建协方差矩阵：基于 5 分钟收益率序列，采用滚动窗口法(窗口长度为 240，对应 1 个交易日)计算每支股票的已实现波动率，并构建 30 × 30 的协方差矩阵。

3) 数据清洗与标准化：剔除缺失值超过 10%的样本，对剩余数据进行 Z-score 标准化，消除量纲影响。采用经验模态分解去除数据非平稳性，保留与原始序列相关性较高的本征模态函数分量作为模型输入。

数据二来源：

GEO 数据库 GSE53757 肺癌基因表达数据，包含 126 例样本(63 例癌症组，63 例对照组)，每个样本测度 12,600 个基因表达量。

预处理步骤[8]：

1) 缺失值处理：对缺失率 > 5%的基因列进行删除，剩余 11,892 个基因；

2) 数据标准化：采用 Z-score 标准化，消除基因表达量的量纲差异；

3) 特征筛选：通过单变量 t 检验筛选组间差异显著基因($p < 0.05$)，保留 2000 个候选基因。

4.2. 数据模型构建

模型一构建：

1) VAR-LASSO：基于 Lasso 惩罚的向量自回归模型，用于捕捉资产收益率的动态相关性。

2) VAR-SCAD：引入 SCAD 惩罚函数，克服 Lasso 的 Oracle 性质缺陷，提升参数估计精度。

3) VAR-MCP：采用 MCP 惩罚函数，进一步优化稀疏性，适用于高维协方差矩阵建模。

参数设置：

基于 AIC 准则，VAR 阶数选择为 3 [9]，正则化参数 λ 通过交叉验证确定，可使用 R 语言中的 parsevar 包中的 varMCP 函数构建 VAR-MCP 模型。

模型二构建：

1) 响应变量：癌症状态(二分类，0 = 对照，1 = 癌症)；

2) 模型对比：VAR-MCP 对比 Lasso [10]对比弹性网(Elastic Net) [11]；

3) 参数设置：VAR 阶数通过 BIC 准则确定为 2，正则化参数采用 10 折交叉验证。

4.3. 模型评估与结果分析

模型一预测精度指标：

均方误差(MSE)：衡量协方差矩阵预测误差。

R^2：解释方差比例。

夏普比率：量化单位风险的超额收益：$Sharpe\,Ratio = \frac{R_p - R_f}{\sigma_p}$ 其中 R_p 为投资组合收益率，R_f 为无风险利率(取一年期国债利率为 2.5%)，σ_p 为组合波动率。

结果对比：

Table 1. Comparison of evaluation results of model 1
表 1. 模型一评估结果对比

模型一	MSE	R^2	夏普比率	年化收益率	最大回撤
VAR-LASSO	0.082	0.65	0.83	8.20%	15.60%
VAR-SCAD	0.071	0.72	0.91	9.50%	12.80%
VAR-MCP	0.068	0.75	0.98	10.30%	11.50%

见表 1，VAR-MCP 模型 MSE 值最小，R^2 值最大，这说明模型对以上数据的拟合效果更好，在预测精度和风险调整收益上表现最优，其夏普比率较 VAR-LASSO 提升 18%，年化收益率提高 2.1%。

模型二预测精度指标：

预测精度：准确率(Accuracy)、AUC-ROC；

变量选择：真阳性率(TPR)、假阳性率(FPR)。

结果对比：

Table 2. Comparison of evaluation results of model 2
表 2. 模型二评估结果对比

模型二	Accuracy	AUC-ROC	TPR	FPR	计算时间(秒)
VAR-MCP	0.892	0.915	0.873	0.121	215
Lasso	0.856	0.862	0.821	0.234	187
弹性网	0.871	0.883	0.845	0.189	202

见表 2，VAR-MCP 模型在癌症分类中准确率最高，达到 89.2%，且误判率 FPR 显著低于 Lasso，证明其在生物医学高维数据中能更精准筛选关键基因。计算时间虽略高于 Lasso，但考虑到变量筛选质量的提升，性价比优势显著。

故引入 VAR-MCP 模型，能在协方差矩阵预测和投资组合优化以及生物医学数据中相较于传统的惩罚回归模型表现得更为优异，其评估性能更为优越，计算效率及其准确率明显优于传统模型，为量化投资和识别关键基因提供了新工具。

5. 结论与展望

5.1. 研究结论

本研究聚焦于在高维数据场景下的 Lasso 惩罚回归与非凸惩罚线性回归，系统探讨了传统 Lasso 回

归的基础理论和惩罚定义的优点，并引出了如 SCAD、MCP 等更为优化的非凸惩罚函数，并详细介绍了其 Oracle 性质、算法优化及实际应用。其中强调非凸惩罚的 Oracle 性质优势，以 MCP 为代表的非凸惩罚估计满足变量选择一致性和渐近正态性，能以高概率准确识别真实模型中的非零变量，并有更高的估计精度和更快的估计效率，突破了 Lasso 依赖的限制等距性条件，为高维统计建模提供了更普适的理论基础。此外还引用了基于 MCP 惩罚的 VAR-MCP 模型，通过 DC 分解保证目标函数的全局收敛性，结合坐标下降法优化参数估计，解决了非凸优化中易陷入局部最优的问题，还引入 SIS 与非凸惩罚结合的两阶段策略，显著提升了算法在实际高维场景中的估计速率。最后通过实证验证在金融领域，基于沪深成分股的高频交易数据，证明了非凸惩罚模型在高维金融数据中更强的预测精度和风险控制能力；在生物医学方面验证其在稀疏生物数据中排除噪声、识别关键基因的能力，弥补了 Lasso 过度压缩重要变量的缺陷。

5.2. 研究展望

尽管本研究在理论和应用上取得一定进展，但高维非凸惩罚模型仍有着更广阔的探索空间和更深层次的理论基础。在理论拓展方面可以研究非凸惩罚在更复杂场景能否适配，进一步对比 SCAD、MCP 等非凸惩罚函数的渐近性质差异，分析其在不同数据分布和假设前提下的 Oracle 性质，还可以探究其在广义线性模型中的应用潜力。在算法优化方面可以针对千万级以上特征的超维数据，开发基于交替方向乘子法或分布式的并行化算法，解决传统算法在存储和时间上的瓶颈，也可以设计融合模型权重动态调整的 SIS 改进方法，提升对强相关特征和极端稀疏数据的鲁棒性，降低两阶段策略对先验假设的依赖。

5.3. 结语

本研究通过基于高维数据 Lasso 惩罚回归的背景，阐述了传统惩罚回归的基础知识和理论方法，进一步揭示了非凸惩罚在高维线性回归中的独特优势，为高维数据建模提供了更具准确性与效率的解决方案，通过金融高频数据与生物基因数据的跨领域实证，验证 VAR-MCP 在高维场景中的普适性，其预测精度与变量选择准确性均显著优于传统 Lasso 模型。未来研究需进一步深究基础理论、优化算法效能、拓展应用场景，推动非凸惩罚方法从理论优势向实际价值的深度转化，为统计学科与各领域的交叉融合提供更多方法与创新。

参考文献

[1] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[2] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[3] Breheny, P. and Huang, J. (2011) Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *The Annals of Applied Statistics*, **5**, 232-253. https://doi.org/10.1214/10-aoas388

[4] Fan, J. and Lv, J. (2008) Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **70**, 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

[5] Wei, Q. and Zhao, Z. (2023) Large Covariance Matrix Estimation with Oracle Statistical Rate. *ICASSP* 2023-2023 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Rhodes Island, 4-10 June 2023, 1-5. https://doi.org/10.1109/icassp49357.2023.10095334

[6] Yang, X. (2023) Modeling of High-Dimensional Covariance Matrix Based on Non-Convex Penalty Function. *Journal of Southwest China Normal University* (*Natural Science Edition*), **48**, 13-22.

[7] Yuan, M. and Lin, Y. (2005) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **68**, 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

[8] 范良勤, 张鸿, 田鹏, 等. 儿童咳嗽变异性哮喘转为典型哮喘风险调查及列线图预测模型的构建和验证[J]. 临床肺科杂志, 2023, 28(12): 1861-1867.

[9] 仇婷婷. 基于高维数据的信用评分模型研究与应用[D]: [博士学位论文]. 成都: 西南财经大学, 2024.

[10] 李璇. 基于坐标下降法的半监督学习算法及其在文本分类中的应用[D]: [硕士学位论文]. 广州: 华南理工大学, 2010.

[11] 张国浩. 高维数据下基于弹性网惩罚的复合分位数回归估计及其应用[D]: [硕士学位论文]. 重庆: 重庆工商大学, 2023.