

# 摘 要

随着信息技术的不断发展, 获取海量及高维数据的方式更加方便快捷. 如何从海量及高维数据中提取有价值的信息越来越受到人们的关注. 同时, 大数据极易被异常值破坏, 或数据中包含具有重尾分布的变量, 这使得经典的统计分析方法不再适用. 因此, 对于高维数据下的统计模型进行研究, 具有重要的理论意义和应用价值.

针对高维数据中可能含有的结构信息及异常值问题, 基于具有平滑特性的fused lasso惩罚函数, 本文提出带有fused lasso惩罚的自适应Huber回归模型, 并证明该模型的非渐近收敛性. 在算法方面, 利用模型的特殊结构, 设计可用于求解非光滑凸优化问题的交替方向乘子法, 并分析算法的收敛性. 然后通过低维和高维的数值模拟验证该方法的有效性. 最后将模型运用于三个基因型真实数据, 分析该模型的可解释性.

针对高维数据所面临的维数过高问题, 本文基于半参数Copula估计方法, 设计一个不依赖于模型假设的超高维特征筛选方法, 在一些假设下, 证明该方法的确定性筛选性质. 同时在线性和重尾数据下验证该方法的有效性. 并进一步结合惩罚估计方法和坐标下降算法, 对实际数据进行分析, 阐明该筛选方法的可解释性.

**关键词:** Huber函数, Fused lasso, 交替方向乘子法, Copula估计, 确定性筛选性质

# 本文符号说明

|                                          |                            |
|------------------------------------------|----------------------------|
| $\mathbb{R}$ :                           | 全体实数组成的集合                  |
| $\mathbb{R}^n$ :                         | 全体 $n$ 维实向量组成的集合           |
| $\mathbf{0}_p$ :                         | $p$ 维零向量                   |
| $\text{dom} f$ :                         | 函数 $f$ 的有效域                |
| $\text{Ran} f$ :                         | 函数 $f$ 的值域                 |
| $\partial f(x)$ :                        | $f(x)$ 在 $x$ 点的次微分         |
| $f^*(y)$ :                               | $f(x)$ 在 $y$ 点的共轭函数        |
| $\Pi_{\mathcal{B}_\infty^{(\rho)}}(x)$ : | $x$ 在半径为 $\rho$ 的无穷范数球上的投影 |
| $\mathbf{A}^\top$ :                      | 矩阵 $A$ 的共轭转置               |
| $\mathbf{I}$ :                           | $p \times p$ 维单位矩阵         |
| $\forall$ :                              | 任意的                        |
| $\langle \cdot, \cdot \rangle$ :         | 向量或矩阵的内积                   |
| $\odot$ :                                | Hadamard乘积                 |
| $\ \cdot\ _1$ :                          | 向量的1范数                     |
| $\ \cdot\ _2$ :                          | 向量的欧氏范数或2范数                |
| $\ \cdot\ _\infty$ :                     | 向量的 $\infty$ 范数            |
| $a \lesssim b$ :                         | $a \leq Cb, C > 0$ .       |

# 目 录

|                                       |     |
|---------------------------------------|-----|
| 摘要.....                               | I   |
| ABSTRACT.....                         | III |
| 本文符号说明 .....                          | V   |
| 第一章 绪论.....                           | 1   |
| 1.1 研究背景及意义.....                      | 1   |
| 1.2 研究现状.....                         | 1   |
| 1.3 论文创新点 .....                       | 4   |
| 1.4 主要内容及结构.....                      | 4   |
| 第二章 带有fused lasso惩罚的自适应Huber回归模型..... | 7   |
| 2.1 模型构造.....                         | 7   |
| 2.2 非渐近统计性质与主要结果.....                 | 8   |
| 2.3 理论证明 .....                        | 12  |
| 2.4 算法设计 .....                        | 17  |
| 2.5 数值模拟 .....                        | 23  |
| 2.6 案例分析 .....                        | 27  |
| 2.7 本章小结 .....                        | 29  |
| 第三章 基于半参数Copula估计的超高维特征筛选方法.....      | 31  |
| 3.1 基础知识 .....                        | 31  |
| 3.1 方法设计 .....                        | 33  |
| 3.2 确定性筛选性质与理论证明.....                 | 34  |
| 3.3 数值模拟 .....                        | 38  |
| 3.4 案例分析 .....                        | 45  |
| 3.5 本章小结 .....                        | 48  |
| 第四章 结论及展望.....                        | 49  |

|              |    |
|--------------|----|
| 参考文献 .....   | 51 |
| 致    谢 ..... | 57 |
| 附    录 ..... | 59 |

# 第一章 绪论

## §1.1 研究背景及意义

随着计算机技术的高速发展,数据搜集变得越来越容易,这意味着在实际问题中,需要同时考虑的变量维数正在飞速增加.然而,在便捷获取高维数据的同时,数据中可能含有重尾分布的噪声,这会使得数据结构遭到不同程度的破坏.

回归分析作为统计学的重要研究方法,包括了一大类通过建模来分析数据的方法,并且这些方法被广泛应用于生物医学、金融工程、生产生活等各个方面.作为一类最常见于线性回归模型的参数估计方法,最小二乘估计方法将二次函数作为损失项,来衡量估计误差.由于二次损失函数是一个光滑可导的凸函数,易于寻找极值,因此该方法在实际中应用广泛.然而,对于具有重尾分布的数据,使用最小二乘估计方法可能会遇到理论和实际上的困难.因此,建立稳健的回归模型来处理重尾数据,是一个值得深入研究的问题.另外,在算法方面,虽然选取适当的惩罚函数能同时实现参数估计和变量选择,但由于数据维数大,惩罚函数非光滑等特征,传统算法对回归系数的估计具有一定困难.因此,设计行之有效的优化算法对模型进行高效求解,也具有重要的研究意义.

进一步地,随着计算机自动化处理数据逐渐取代反复低效的人工数据处理方式,高维甚至超高维数据几乎随处可见.然而当数据维数呈指数级攀升,即在超高维数据背景下时,构造有效的超高维特征筛选方法是对数据进行初步降维的必要手段.

因此,结合统计理论分析和快速稳定的优化算法,探讨约束变量选择模型以及超高维特征筛选方法,对于高维统计问题的研究具有重要的实际意义.

## §1.2 研究现状

在实际生活中,数据往往包含潜在的结构信息.其中,具有序列结构的数据随处可见,例如在高维生物医学研究<sup>[21]</sup>、信号处理<sup>[18]</sup>、以及空气质量分析<sup>[8]</sup>等领域.为了处理此类带有序列特征的数据,Tibshirani等<sup>[29]</sup>首先提出了带有fused lasso惩罚的最小二乘回归模型.此后,关于fused lasso的惩罚估计吸引了众多学者的研究.例如,Mao等<sup>[25]</sup>将时间先

验信息运用于缺失交通数据, 然后使用fused lasso正则项来拟合交通数据的时间相关性; Petersen等<sup>[27]</sup>提出了fused lasso加性模型; Corsaro等<sup>[6]</sup>将fused lasso应用于利润加权, 研究了多周期利润选择问题; Cui等<sup>[7]</sup>使用fused lasso框架, 根据特征相关性进行特征重排; Degras等<sup>[8]</sup>将fused lasso应用于多元时间序列分析.

另外在算法方面, 结合fused lasso惩罚项的特点, 学者们提出了许多不同的估计方法. 例如, Tibshirani等<sup>[29]</sup>考虑了当正则参数固定时, fused lasso惩罚最小二乘模型对应于二次规划问题, 因此设计了SQOPT两阶段动态算法, 但当变量维数较大时, 求解过程相对复杂, 计算时间也较长. Li等<sup>[20]</sup>运用线性化交替方向乘子法研究了fused lasso模型. Wang等<sup>[31]</sup>提出了一种新的估计方法, 可将所涉及的约束条件数量大幅减少, 但同时会导致计算复杂度增加. Li等<sup>[21]</sup>提出了一种高效的基于非精确半光滑牛顿的增广拉格朗日算法来求解大规模fused lasso问题, 但是牛顿类算法通常对函数的性质要求更加严格.

虽然可以通过一些经典的统计分析方法研究数据中的结构信息, 但是由实际中使用的数据集可以看出, 大数据易被异常值污染, 或数据中含有重尾分布的噪声, 这使得许多传统估计方法不再适用. 为解决这一挑战, 有学者提出了Huber函数来进行稳健估计和推理. 许多学者对该函数展开了广泛研究. 例如, Sun等<sup>[28]</sup>提出了一种具有自适应调节参数的Huber回归模型, 并阐明该参数与随机噪声的样本大小、维数、以及矩相关. Huang和Wu<sup>[14]</sup>将分段Huber损失函数应用于噪声只满足一个弱矩条件的情况, 并建立了一个描述超额泛化误差与预测误差之间差距的比较定理. Liu等<sup>[23]</sup>将Huber损失函数与广义lasso惩罚项相结合, 实现了针对重尾数据的参数估计和变量选择, 且主要关注信息准则中的自由度公式.

另外在算法方面, Sun等<sup>[28]</sup>利用Fan等<sup>[10]</sup>提出的局部迭代自适应最小化算法解决了带有lasso惩罚的Huber回归问题, 该方法在拟合高维模型时可以控制精度和统计误差, 但没有考虑数据中的复杂结构信息. Chen等<sup>[4]</sup>设计了加速近端梯度算法来求解带有弹性网惩罚的多元Huber矩阵回归模型, 该模型可以减少但不能很好地抵抗离群值对估计量的负面影响. Luo等<sup>[24]</sup>提出了一种鲁棒分布算法, 在数据具有有限二阶矩的重尾或非对称误差时拟合线性回归模型, 该方法分别在低维和高维数据中结合了Barzilai-Borwein梯度下降算法和局部自适应majorize最小化方法. Ghosh等<sup>[12]</sup>基于有向Huber-Markov惩罚项, 提出了基于最大似然估计的超分辨率算法.

然而, 当特征维数随样本容量呈指数阶增长时, 传统的统计方法可能会失效, 因此迫切要求针对超高维问题的特征筛选方法的出现. Fan和Li<sup>[40]</sup>首先分析了超高维数据在各种

统计问题中带来的困难, 并提出特征筛选方法SIS(Sure Independent Screening), 即利用响应变量和单个协变量之间的皮尔逊相关系数来度量各个协变量的重要程度, 从而使协变量的维数下降到适当程度, 使传统惩罚估计方法可以有效使用. 虽然边际筛选方法可以实现超高维数据的快速降维, 但该方法忽略了变量间可能存在的强相关性, 从而可能导致选入过多的冗余变量; 同时逐个协变量筛选的方式可能漏选那些边际不相关但联合相关的协变量. 因此, Fan和Lv<sup>[41]</sup>提出了适用于线性回归模型的迭代特征筛选方法ISIS(Iteratively Sure Independent Screening). 之后, Liu等<sup>[42]</sup>对超高维特征筛选方法做出较为系统的总结, 并将这些方法分为两类: 基于模型的特征筛选和无模型特征筛选.

基于模型的特征筛选方法可以从参数模型, 非参数模型, 以及半参数模型三个角度进行分类. 在参数模型假设下, Fan等<sup>[43]</sup>将ISIS方法扩展到一个伪似然框架, 并对广义线性模型的情形进行了研究. Wang<sup>[44]</sup>提出了前向回归筛选方法, 并证明了该方法具有确定性筛选性质. Fan和Song<sup>[45]</sup>提出了一般化筛选框架, 即在广义线性模型中通过最大边际似然估计进行排序. Zhao和Li<sup>[46]</sup>研究了针对Cox模型的超高维特征筛选方法. Xu和Chen<sup>[47]</sup>提出了稀疏限制极大似然估计量以获得计算效率, 且该方法在超高维广义线性模型的背景下具有良好的理论保证. Liu等<sup>[48]</sup>基于条件相关学习, 提出了变系数模型的确定性独立筛选方法. 在非参数模型和半参数模型假设下, Fan等<sup>[49]</sup>提出了一种基于超高维加性模型的非参数独立性筛选方法. Fan等<sup>[50]</sup>提出了一种边际非参数筛选方法, 并设计了两种数据驱动的迭代NIS方法来选择阈值参数和变量. Liang等<sup>[51]</sup>提出了PFR(Profiled Forward Regression)算法, 用于半参数部分线性模型的超高维变量筛选. He等<sup>[52]</sup>提出了一种基于分位数回归的自适应非参数筛选方法, 并用于异构数据分析.

上述特征筛选方法是基于模型的假设提出的. 然而在没有先验信息的情况下, 指定模型结构往往会造成估计偏差. 因此, Zhu等<sup>[53]</sup>提出了活跃预测变量和非活跃预测变量的概念, 并在无模型假设下确定独立的排序筛选程序. Li等<sup>[54]</sup>提出了一种基于距离相关的确定独立筛选方法. Shao和Zhang<sup>[55]</sup>提出了一种秩差相关性来衡量变量之间条件均值独立性. Mai和Zou<sup>[56]</sup>提出了fused Kolmogorov筛选方法, 该方法更适用于分类数据. Cui等<sup>[57]</sup>基于均值方差指数进行超高维判别分析. Ma等<sup>[58]</sup>基于分位数偏相关进行特征筛选, 并应用扩展的贝叶斯信息准则进行最佳子集筛选. Zhu等<sup>[59]</sup>提出使用投影相关性来表征两个随机向量之间的依赖关系. Pan等<sup>[60]</sup>提出球协方差来作为随机变量间相关性的度量. 综上可知, 特征筛选可以从模型框架下分离出来, 而且无模型筛选方法更加稳健.

从以往学者们提出的超高维特征筛选方法可以看出, 其中一类筛选方法是通过变量

之间的相关性进行研究. 而作为描述随机变量间相关性的重要指标, Copula函数在金融、计量经济学、以及系统神经科学等领域应用广泛. 从函数角度来看, Copula是一种多元分布函数. 该函数可以作为研究无标度相关性的方法, 也可以作为构建二元分布族的起点<sup>[61]</sup>, 这是统计学家对其感兴趣的两个主要原因. 例如, Deheuvels<sup>[63]</sup>使用“经验依赖函数”估计Copula函数, 并构造了非参数独立性检验. Schweizer和Wolff<sup>[64]</sup>讨论了随机变量在严格单调变换下Copula的基本不变性, 并提出了“*Schweizer and Wolff's  $\sigma$* ”依赖测度. Genest等<sup>[65]</sup>研究了与Copula函数相关的多元分布族中半参数估计方法的理论性质. Liu<sup>[66]</sup>提出了一个Copula 多标签学习方法来刻画标签和特征间的相关性.

### §1.3 论文创新点

本文的主要工作有以下两个:

一、针对高维数据中可能含有的结构信息及异常值问题, 本文提出带有fused lasso惩罚的自适应Huber回归模型, 证明该模型的非渐近统计性质. 利用模型的特殊结构, 设计交替方向乘子法进行求解, 并分析算法的收敛性. 然后通过数值模拟验证该方法的有效性, 最后将模型运用于三个基因型真实数据, 分析该模型的可解释性.

二、针对高维数据所面临的维数过高问题, 本文基于半参数Copula估计方法, 设计不依赖于模型假设的超高维特征筛选方法, 在一些假设下, 证明该方法的确定性筛选性质. 同时在线性和重尾数据下验证该方法的有效性, 最后对实际数据进行分析, 阐明该筛选方法的可解释性.

### §1.4 主要内容及结构

本文的主要内容及结构如下:

第一章主要介绍本文所要研究问题的背景和意义, 研究现状及分析, 以及论文创新点, 提出本文所要研究的主要问题.

第二章主要研究高维数据中的结构信息与异常值问题, 通过构造带有fused lasso惩罚的自适应Huber回归模型, 证明估计参数满足非渐近统计性质; 采用交替方向乘子法求解问题, 并保证算法的收敛性; 通过低维和高维数据模拟, 验证算法的有效性; 最后通过三个真实基因数据集, 分析模型的可解释性.

第三章主要研究高维数据中的维数过高问题, 基于半参数Copula估计, 构造超高维特征筛选方法; 证明其确定性筛选性质; 通过非线性和重尾模型进行数值模拟, 验证筛选方法的有效性; 最后通过真实生存数据集, 分析筛选方法的可解释性.

第四章是本文的总结与展望, 总结本文所研究问题的主要方法以及得到的结论, 也提出今后需要努力的研究方向.



以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/848055105105007005>