

基于逆回归方法的分布式特征筛选

张 妍, 张俊英*

太原理工大学数学学院, 山西 太原

收稿日期: 2024年12月26日; 录用日期: 2025年1月18日; 发布日期: 2025年1月29日

摘 要

本文我们提出了一个通过逆回归估计实现大数据设置的分布式筛选框架。本着分而治之的思想, 本文提出的框架可用分布估计条件方差的逆回归模型来表达相关关系。通过分量估计的聚合, 我们得到了一个最终的逆条件方差估计, 可以很容易地用于筛选特征。该框架支持分布式存储和并行计算, 因此在计算上具有吸引力。由于分量参数的无偏分布估计, 最终的聚合估计具有较高的精度, 且对数据段数量 m 不敏感。在一般条件下, 我们证明了聚合估计器在概率收敛界和均方误差率方面与集中估计器一样有效; 相应的筛选过程对广泛的相关度量具有一定的筛选特性。

关键词

超高维, Gini相关系数, 变量筛选, 特征排序

Distributed Feature Screening via Inverse Regression

Yan Zhang, Junying Zhang*

College of Mathematics, Taiyuan University of Technology, Taiyuan Shanxi

Received: Dec. 26th, 2024; accepted: Jan. 18th, 2025; published: Jan. 29th, 2025

Abstract

In this paper, we propose a distributed screening framework for big data setup via inverse regression estimator. In the spirit of divide-and-conquer, the proposed framework expresses the dependent relative by inverse regression model in which can be distributively estimated inverse conditional variance. With the component estimates aggregated, we obtain a final inverse conditional variance estimator that can be readily used for screening features. This framework enables distributed storage and parallel computing and thus is computationally attractive. Due to the unbiased

*通讯作者。

distributive estimation of the component parameters, the final aggregated estimate achieves a high accuracy that is insensitive to the number of data segments m . Under mild conditions, we show that the aggregated estimator is as efficient as the centralized estimator in terms of the probability convergence bound and the mean squared error rate; the corresponding screening procedure enjoys sure screening property for a wide range of correlation measures.

Keywords

Ultrahigh Dimension, Gini Correlation Coefficient, Variable Screening, Feature Ranking

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着数据生成和采集技术的快速发展,在许多科学领域经常会遇到具有大量特征的海量数据。对于经典的统计方法,高维性同时对计算成本、统计准确性和算法稳定性提出了挑战[1][2]。为了简化计算过程,一种自然的策略是在详细分析之前筛选出最不相关的特征。这个过程被称为特征筛选。随着维数从高降至低,分析难度大大降低。在文献中,这方面已经做了大量的工作;其中,基于相关性的筛选方法引起了学者们的广泛关注。这些方法基于特征与响应之间的某种关联度量进行筛选,弱相关性的特征被视为不相关的特征,将被删除。这种类型的方法可以在没有强模型假设(甚至无模型)下方便地实现。因此,它们通常用于分析具有复杂结构的高维数据,例如, Fan 和 Lv (2008)提出了基于皮尔逊相关的确定独立性筛选(SIS) [3],随后他们又把 SIS 方法进一步推广到广义线性模型和非参数可加模型中[4][5]。Zhu 等(2011)提出了一种基于效用度量的确定独立排序和筛选(SIRS),该度量与给定预测因子的响应的整个条件分布有关[6]。Li 等(2012)提出了基于 Kendall 秩相关的鲁棒秩相关筛选(RRCS) [7]。Li 等(2012)开发了一种基于距离相关的无模型确定独立筛选程序(DC-SIS) [8]。Wu 和 Yin (2015)提出了一种分布函数确定独立性筛选(DF-SIS)方法,该方法使用一种度量来检验两个变量之间的独立性[9]。Zhou 等(2019)提出了一种鲁棒相关度量来筛选包含极值的特征[10]。

特征筛选[11]在许多应用中已被证明是一个有吸引力的策略。现有的方法大多是在特征数 p 较大,但样本量 N 为中等的情况下开发的。然而,在现代科学研究中,数据分析师不得不处理大数据集(其中 p 和 N 都是巨大的)的情况越来越普遍,例如,在现代全基因组基因研究[12]中,对数十万名参与者进行了数以百万计的 SNP 基因分型。在互联网研究中,一个杀毒软件每分钟可以扫描数百万个 URL 中的数万个关键字。当面对大 p 大 N 的数据时,由于存储瓶颈和算法可行性,直接采用经典筛选方法在数值上效率低下,例如,对于 $N = p = 10,000$ 的数据集,众所周知的 DC-SIS 需要大约 60 小时才能在具有 3.2 GHz CPU 和 32 GB 内存的计算机上进行全面筛选。因此,在实践中开发方便计算的大数据筛选方法[13]-[15]是可取的。

当数据集太大而无法在一台计算机上处理时,考虑使用分而治之策略是很自然的。在该策略中,先将一个大问题分解为较小的可管理子问题,然后将相应的子输出组合得到最终输出。本着这种思想,许多机器学习和统计方法都已经被重建用于处理大数据(例如, Chen 和 Xie, 2014 [16]; Xu *et al.*, 2016 [17]; Jordan 等人, 2019 [18]; Banerjee *et al.*, 2019 [19])。这些鼓舞人心的工作激励我们探索利用这种有前景的策略进行大数据特征筛选的可行性。

本文提出了一种基于聚合相关测度的分布式特征筛选框架,并将其称为聚合相关筛选(ACS)。在 ACS 中,我们将相关度量表示为几个组成参数的函数,每个组成参数都可以使用来自数据段的自然 U 统计量

进行分布估计。将无偏分量估计结合在一起, 我们得到了一个聚合的相关估计, 可以很容易地用于特征筛选。在本文提出的 ACS 框架中, 将海量数据集分成 m 个可管理的数据段进行处理, 这些数据段可存储在台计算机中, 并通过并行计算完成相应的局部估计。因此, 它为大 p 大 N 数据的特征筛选提供了一条计算上有吸引力的途径。此框架也适用于数据自然存储在不同位置的设置(例如, 医院级别的医疗数据)。分量参数的 U 统计量估计作为一种有效、方便的降阶技术, 保证了聚合相关估计量的高精度和相应筛选过程的可靠性。在一般条件下, 我们证明了在概率收敛界和均方误差(MSE)率方面, 聚合相关估计器与经典的集中估计器一样有效。这种全效率对 m 的选择不敏感, m 的选择可以由问题本身指定, 也可以由用户决定。对于广泛的相关度量, 我们进一步证明了 ACS 在不需要指定参数模型(无模型)的情况下具有确定的筛选特性。

本文所提出的 ACS 植根于分量估计。在文献中, 该思想已被用于分布式恢复由数据段中可分离的光滑估计方程定义的集中估计量(例如, Chen 等人, 2008 [20]; Lin and Xi, 2011 [21])。不幸的是, 这些工作不能直接适用于基于相关性的筛选, 因为许多常用的相关度量的集中估计量通常不是由估计方程定义的, 并且在数据段(例如, SIRS 和 DC)中不可分离。本文所提出的 ACS 来自于集中相关估计量的自然组合; 它并不寻求完全恢复集中式估计器, 而是提供一种有效且计算负担得起的替代方案。我们的研究结果作为使用这种自然策略进行分布式特征筛选提供了理论支持。

2. 核估计中的筛选方法

2.1. 大数据特征筛选

设 $D = \{(Y_i, \mathbf{X}_i)\}_{i=1}^N$ 是 $\{Y, \mathbf{X}\}$ 的 N 个独立同分布的样本, 其中 Y 是一个支持 ϕ_y 的响应变量, $\mathbf{X} = (X_1, \dots, X_p)^T$ 是一个 p 维协变量向量。我们感兴趣的是 p 和 N 都很大的情况。当一个数据集是大量的且高维时, 通常可以合理地假设只有少数的协变量(特征)与响应相关。设 $F(x_j|Y)$ 是 Y 给定条件下 X_j 的条件分布函数。如果 $F(x_j|Y)$ 依赖于 Y , 则认为这个特征 X_j 是相关的。我们使用 M 来表示相关特征的索引集, 并定义 $M^c = \{1, \dots, p\} \setminus M$ 。特征筛选的目标是在详细分析之前, 用 $j \in M^c$ 去除大多数不相关的特征 X_j 。

具体来说, 设 $w_j \geq 0$ 是 Y 和 X_j 之间相关性强度的一个度量。设 $\hat{w}_j$ 是基于 D 的 w_j 的集中估计。预先设定阈值 $r_n > 0$, 则可以保留在

$$\hat{M} = \{j : \hat{w}_j > r_n, j = 1, 2, \dots, p\}.$$

中的特征并把其他的去除。当样本量 N 适中时, 这种经典方法是有效的。然而, 当 p 和 N 都很大时, 基于完整数据集 D 计算 $\{\hat{w}_j\}_{j=1}^p$ 在数值上可能是昂贵的。

2.2. 逆回归聚合筛选法

受近期分布式学习工作的启发, 我们考虑采用分而治之的思想[22]来解决大数据特征筛选问题。不失一般性, 假设初始完整数据集 D 被相等地划分为 m 个可管理的片段 $\{D_l\}_{l=1}^m$, 每个片段包含 $n = N/m$ 个观测值。根据计算环境的不同, 这些数据段可以分布地存储在台计算机上并进行处理, 也可以由单台计算机按顺序进行处理。

我们考虑通过逆回归(Li, 1991) [23]基于核矩阵 $\text{cov}(E(\mathbf{X}|Y))$ 进行特征筛选, 不失一般性, 假定 $E(\mathbf{X}) = \mathbf{0}$, $E(Y) = 0$, 令 $G_j(Y) = E(X_j|Y)$, $R = \text{cov}(E(\mathbf{X}|Y)) = (r_{i,j})_{p \times p}$, $r_{i,j} = \text{cov}[E(X_i|Y), E(X_j|Y)]$ 。设 $w_j^l = r_{j,j}^l = \text{var}(E(X_j|Y \in D_l))$, $w_j = \frac{1}{m} \sum_{l=1}^m w_j^l$, 对于 $j = 1, 2, \dots, p$, $w_j^l > 0$, 若 X_j 与 Y 独立, 则有 $w_j^l = 0$ ($j = 1, 2, \dots, p, l = 1, 2, \dots, m$)。因此, 重写活动集 $M = \{j : w_j > 0, j = 1, 2, \dots, p\}$ 。

现在我们通过核方法估计 w_j 。令 $g(y)$ 表示 Y 的密度函数, $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ 为 $\mathbf{X}$ 的副本, 定义

$$G(Y) \triangleq (G_1(Y), \dots, G_p(Y))^T = (E(X_1|Y), \dots, E(X_p|Y))^T \triangleq E(\mathbf{X}|Y)^T,$$

$$h(Y) = (h_1(Y), \dots, h_p(Y))^T = (G_1(Y)g(Y), \dots, G_p(Y)g(Y))^T.$$

用核平滑方法[1]估计 w_j 如下:

$$\hat{h}_j^l(Y) = \frac{m}{hN} \sum_{\{X_{ij}, Y_i\} \in D_l} X_{ij} K((Y - Y_i)/h),$$

$$\hat{g}^l(Y) = \frac{m}{Nh} \sum_{Y_i \in D_l} K((Y - Y_i)/h).$$

$$\hat{G}_j^l(Y) = \hat{h}_j^l(Y) / \hat{g}^l(Y), \hat{R}^l = \frac{m}{N} \sum_{Y_h \in D_l} \hat{G}_j^l(Y_h) \hat{G}_j^l(Y_h)^T,$$

$$\hat{w}_j^l \triangleq \hat{R}_j^l = \frac{m}{N} \sum_{Y_h \in D_l} (\hat{G}_j^l(Y_h))^2$$

且

$$\hat{w}_j = \frac{1}{m} \sum_{l=1}^m \hat{w}_j^l, \quad (1)$$

这里 h 表示带宽, $K(\cdot)$ 为核函数, 对于带宽的选择按照 plug-in 方法(Ruppert *et al.*, 1996 [24])。

本文所提出的分布式筛选采用了逆回归方法和核估计器。我们首先进行组件初始化 $\hat{w}_j^l$, 然后将它们组合在一起, 得到一个很容易用于特征筛选的聚合估计。我们将该方法命名为 AIK (聚合逆回归方法和核估计)。

2.3. 理论分析

现在, 我们提供一些使用 AIK 方法的理论证明。显然, AIK 的筛选性能依赖于聚合相关估计 $\hat{w}_j$ 的准确性。我们表明, $\hat{w}_j$ 是一种有效的估计 w_j 的工具; 这是 AIK 的理论基础。我们的理论研究基于以下技术条件:

- **C1.** 密度函数 $g(y)$ 在 $D_l, l=1, 2, \dots, m$ 上有连续的二阶导数, 且 $f'(y), f''(y)$ 具有有限上界, 即对于某正数 M_1, M_2 , $f'(y) \leq M_1$, $f''(y) \leq M_2$ [25] [26]。
- **C2.** 核函数 $K(\cdot)$ 一致有界, 且 $\int t^2 K(t) dt < \infty$ 。
- **C3.** 对于正的常数 $\kappa, c_3, \gamma > \kappa$, $\max_{\{Y \in \mathcal{D}_l\}} [E(X_j|Y)]^2 < c_3 N^{-\kappa}$ 。

基于上述条件, 我们推导了所提方法 AIK 的统一筛选框架。

定理 2.1 在条件 C1~C3 下, 假定带宽 $h = O(N^{-\gamma})$, γ 在条件 C3 中定义, 对于一些正常数 C_2, c_5 ,

$$P\left(\max_{1 \leq j \leq p} |\hat{w}_j - w_j| \geq c_3 N^{-\kappa}\right) \leq C_2 \exp(-c_5 N^{\gamma-\kappa}),$$

且若取 $r_n = c_3 N^{-\kappa}$, 则

$$P(M \subset \hat{M}) \geq 1 - C_2 s_n N \exp(-c_5 N^{\gamma-\kappa}),$$

这里 c_5 为正常数。

3. 定理证明

引理 1. 对任意随机变量 X , 下列两个语句等价:

- **A.** 存在 $H > 0$ 使得对于所有的 $|t| < H$, 有 $Ee^{tX} < \infty$ 。
- **B.** 存在 $r > 0$ 使得 $Ee^{s(X-EX)} < e^{rs^2}$ 。

引理 2. 假定 X 是一个随机变量, 对于某个 $a > 0$ 有 $E(e^{a|X|}) < \infty$ 。则对于任意 $M > 0$, 存在正常数 b 和 c 使得

$$P(|X| \geq M) \leq be^{-cM}.$$

引理 3. (霍丁夫不等式[27] [28]) 设独立随机变量 $X_i, i=1, \dots, n$ 对于一些 $a_i, b_i, i=1, \dots, n$ 满足 $P(X_i \in [a_i, b_i]) = 1$ 。则对任意 $\varepsilon > 0$, 有

$$P(|X - E(X)| \geq \varepsilon) \leq 2 \exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \quad (2)$$

这里 $X = (X_1 + \dots + X_n)/n$ 。

引理 4. 设 $a(u)$ 和 $b(u)$ 是 u 的两个一致有界函数, 即存在 $M_5 > 0, M_6 > 0$, 使得 $\sup_{u \in U} |a(u)| \leq M_5$, $\sup_{u \in U} |b(u)| \leq M_6$ 。对于给定的 $u \in U$, $\hat{A}(u)$ 和 $\hat{B}(u)$ 是基于大小为 n 的样本的 $a(u)$ 和 $b(u)$ 的估计。假定对任意小的 $\varepsilon \in (0, 1)$, 存在正常数 c_1, c_2 和 s , 使得[29]

$$\sup_{u \in U} P(|\hat{A}(u) - a(u)| \geq \varepsilon) \leq c_1 \left(1 - \frac{\varepsilon s}{c_1}\right)^n, \quad (3)$$

$$\sup_{u \in U} P(|\hat{B}(u) - b(u)| \geq \varepsilon) \leq c_2 \left(1 - \frac{\varepsilon s}{c_2}\right)^n, \quad (4)$$

则

$$\sup_{u \in U} P\left(\left|(\hat{A}(u))^2 - (a(u))^2\right| \geq \varepsilon\right) \leq C_7 \exp\left(-\frac{\varepsilon}{C_7 h}\right), \quad (5)$$

$$\sup_{u \in U} P\left(\left|\hat{A}(u)/\hat{B}(u) - a(u)/b(u)\right| \geq \varepsilon\right) \leq C_9 \exp\left(-\frac{\varepsilon}{C_9 h}\right), \quad (6)$$

定理 2.1 的证明. 证明步骤如下:

步骤 1: 证明对任意的 $\varepsilon \in (0, 1)$, $1 \leq j \leq p$, 可以找到一个正常数 C 使得

$$P\left(\left|\hat{h}_j'(Y) - hg'(Y)m_j'(Y)\right| \geq \varepsilon\right) \leq C \exp\left(-\frac{\varepsilon}{h}\right).$$

注意到

$$\begin{aligned} & P\left(\left|\hat{h}_j'(Y) - hg'(Y)m_j'(Y)\right| \geq \varepsilon\right) \\ &= P\left(\left|\hat{h}_j'(Y) - E[\hat{h}_j'(Y)] + E[\hat{h}_j'(Y)] - hg'(Y)m_j'(Y)\right| \geq \varepsilon\right) \\ &\leq P\left(\left|\hat{h}_j'(Y) - E[\hat{h}_j'(Y)]\right| + \left|E[\hat{h}_j'(Y)] - hg'(Y)m_j'(Y)\right| \geq \varepsilon\right) \\ & E[\hat{h}_j'(Y)] - hg'(Y)m_j'(Y) = E\left(X_{ij}K\left(\frac{Y-Y_i}{h}\right)\right) - hg'(Y)m_j'(Y) \\ &= E\left(m_j'(Y_1)K\left(\frac{Y-Y_1}{h}\right)\right) - hg'(Y)m_j'(Y) \triangleq \Delta(Y, h). \end{aligned} \quad (7)$$

上述第一个等式是表示 $E(X_{ij}|Y_i)$ 为 $m_j^l(Y_i)$ 。

又有

$$h^{-1}\Delta(Y, h) = \int \{m_j^l(Y-th)g^l(Y-th) - m_j(Y)g^l(Y)\}K(t)dt.$$

因为 $K(t)$ 是对称的, $\int K(t)dt = 0$ 。此外,

$$\begin{aligned} \lim_{h \rightarrow 0} h^{-2} \left\{ m_j^l(Y-th)g^l(Y-th) - m_j^l(Y)g(Y) - [m_j^{l'}(Y)g(Y) + m_j^l(Y)g^{l''}(Y)]th \right\} \\ \rightarrow \{m_j^{l''}(Y)g(Y) + 2m_j^{l'}(Y)g'(Y) + m_j(Y)g^{l'''}(Y)\}t^2. \end{aligned} \quad [26] \quad (8)$$

因此通过控制收敛定理以及条件 C1~C2, $m_j^{l''}(Y)g(Y) + 2m_j^{l'}(Y)g'(Y) + m_j(Y)g^{l'''}(Y)$ 一致有界, 对于 $Y \in \phi_Y$ 及一些常数 C_0 , $h^{-3}\Delta(Y, h)$ 一致有界, 也就是说,

$$\text{当 } h \rightarrow 0 \text{ 时, } E[\hat{h}_j^l(Y)] - hg^l(Y)m_j^l(Y) \leq C_0 h^3$$

即取 $C_0 h^3 = \varepsilon/2$ 时, $E[\hat{h}_j^l(Y)] - hg^l(Y)m_j^l(Y) \leq \varepsilon/2$ 。因此, (7)式变为

$$\begin{aligned} P\left(\left|\hat{h}_j^l(Y) - hg^l(Y)m_j^l(Y)\right| \geq \varepsilon\right) &\leq P\left(\left|\hat{h}_j^l(Y) - E[\hat{h}_j^l(Y)]\right| \geq \varepsilon/2\right) \\ &\leq P\left(\left|\hat{h}_j^l(Y) - E[\hat{h}_j^l(Y)]\right| \geq \varepsilon/2, \max\{|X_{ij}|\} \leq T\right) + P\left(\left|\hat{h}_j^l(Y) - E[\hat{h}_j^l(Y)]\right| \geq \varepsilon/2, \max\{|X_{ij}|\} \geq T\right) \\ &\leq P\left(\left|\hat{h}_j^l(Y) - E[\hat{h}_j^l(Y)]\right| \geq \varepsilon/2\right) + P\left(\max\{|X_{ij}|\} \geq T\right) = S_1 + S_2, \end{aligned} \quad (9)$$

其中 $S_1 = P\left(\left|\hat{h}_j^l(Y) - E[\hat{h}_j^l(Y)]\right| \geq \varepsilon/2\right)$, $S_2 = P\left(\max\{|X_{ij}|\} \geq T\right)$ 。

根据引理 3 可知 $S_1 \leq 2\exp\left(\frac{-2\varepsilon^2 N}{mT^2}\right)$, 再由引理 2 可知 $S_2 \leq C_1 N/m \exp(-c_1 T)$ 。则

$$\begin{aligned} P\left(\left|\hat{h}_j^l(Y) - hg^l(Y)m_j^l(Y)\right| \geq \varepsilon\right) &\leq 2\exp\left(\frac{-2\varepsilon^2 N}{mT^2}\right) + C_1 N/m \exp(-c_1 T) \\ &= 2\exp\left(-\frac{\varepsilon}{h} \cdot \frac{2\varepsilon h N}{mT^2}\right) + C_1 \exp\left(-\frac{\varepsilon}{h} \cdot c_1 h \log(N/m)T/\varepsilon\right). \end{aligned} \quad (10)$$

又因为 $h = O\left(\left(\frac{N}{m}\right)^{-\gamma}\right)$, $T = O\left((N/m)^\tau\right)$, 这里 $\gamma < \tau < (1-\gamma)/2$, 对于大的 N ,

$$\frac{2\varepsilon h N}{mT^2} = C \cdot \left(\frac{N}{m}\right)^{1-\gamma-2\tau} > 1, c_1 h \log(N/m)T/\varepsilon = C \cdot \left(\frac{N}{m}\right)^{\tau-\gamma} \log(N/m) > 1$$

因此, (10)式变为

$$P\left(\left|\hat{h}_j^l(Y) - hg^l(Y)m_j^l(Y)\right| \geq \varepsilon\right) \leq C \exp\left(-\frac{\varepsilon}{h}\right).$$

相似地, 我们也可以证得对于正常数 C_2 ,

$$P\left(\left|\left(\hat{G}_j^l(Y_i)\right)^2 - \left(m_j^l(Y_i)\right)^2\right| \geq \varepsilon\right) \leq C_2 \exp\left(-\frac{\varepsilon}{h}\right).$$

步骤 2: 证明对任意的 $\varepsilon \in (0, 1)$, 推导出 $P(\max_{1 \leq j \leq p} |\hat{w}_j - w_j| > \varepsilon)$ 的上界。

注意到

$$\begin{aligned}
P\left(\left|\hat{w}_j' - w_j'\right| \geq \varepsilon\right) &= P\left(\left|\frac{m}{N} \sum_{Y_i \in D_l} \left(G_j^l(Y_i)\right)^2 - \frac{m}{N} \sum_{Y_i \in D_l} \left(m_j^l(Y_i)\right)^2\right| \geq \varepsilon\right) \\
&= P\left(\left|\sum_{Y_i \in D_l} \left(G_j^l(Y_i)\right)^2 - \frac{m}{N} \sum_{Y_i \in D_l} \left(m_j^l(Y_i)\right)^2\right| \geq \frac{N}{m} \varepsilon\right) \\
&\leq \sum_{Y_i \in D_l} P\left(\left|\left(G_j^l(Y_i)\right)^2 - \left(m_j^l(Y_i)\right)^2\right| \geq \frac{N}{m} \varepsilon\right) \\
&\leq \frac{N}{m} \max_{Y_i \in D_l} P\left(\left|\left(G_j^l(Y_i)\right)^2 - \left(m_j^l(Y_i)\right)^2\right| \geq \frac{N}{m} \varepsilon\right) \\
&\leq \frac{N}{m} C_2 \exp\left(-\frac{\varepsilon}{h}\right).
\end{aligned} \tag{11}$$

利用同样的思想, 我们可以证明 $P\left(\left|\hat{w}_j - w_j\right| \geq \varepsilon\right) \leq C_2 N \exp\left(-\frac{\varepsilon}{h}\right)$ 。又

$$\begin{aligned}
P\left(\max_{1 \leq j \leq p} \left|\hat{w}_j - w_j\right| \geq c_3 N^{-\kappa}\right) &\leq \sum_{j=1}^p P\left(\left|\hat{w}_j - w_j\right| \geq c_3 N^{-\kappa}\right) \\
&\leq C_2 N \exp\left(-c_4 N^{-\kappa+\gamma}\right) \\
&\leq C_2 \exp\left(-c_4 N^{-\kappa+\gamma} + 2 \log p\right)
\end{aligned}$$

$h = O\left(N^{-\gamma}\right)$, c_4 是一个正常数。取 $\log p = O\left(N^{\gamma-\kappa}\right)$, 即对正常数 ξ , $c_4 - 2\xi > 0$, 有 $\log p = \xi N^{\gamma-\kappa}$ 。则上式变为

$$\begin{aligned}
P\left(\max_{1 \leq j \leq p} \left|\hat{w}_j - w_j\right| \geq c_3 N^{-\kappa}\right) &\leq C_2 \exp\left(-\left(c_4 - 2\xi\right) N^{-\kappa+\gamma}\right) \\
&= C_2 \exp\left(-c_5 N^{-\kappa+\gamma}\right), c_5 = c_4 - 2\xi.
\end{aligned}$$

步骤 3: 取 $r_n = 2c_3 N^{-\kappa}$, 证明 $P\left(\mathcal{M} \subset \hat{\mathcal{M}}\right) \geq 1 - C_2 s_n N \exp\left(-c_5 N^{\gamma-\kappa}\right)$ 。

$$\begin{aligned}
P\left(\mathcal{M} \subset \hat{\mathcal{M}}\right) &= P\left(\min_{j \in \mathcal{M}} \hat{w}_j \geq c_3 N^{-\kappa}\right) = P\left(\min_{j \in \mathcal{M}} w_j - \min_{j \in \mathcal{M}} \hat{w}_j \leq \min_{j \in \mathcal{M}} w_j - c_3 N^{-\kappa}\right) \\
&\geq P\left(\min_{j \in \mathcal{M}} w_j - \min_{j \in \mathcal{M}} \hat{w}_j \leq 2c_3 N^{-\kappa} - c_3 N^{-\kappa}\right) \geq P\left(\max_{j \in \mathcal{M}} \left|w_j - \hat{w}_j\right| \leq c_3 N^{-\kappa}\right) \\
&\geq 1 - P\left(\max_{j \in \mathcal{M}} \left|w_j - \hat{w}_j\right| \geq c_3 N^{-\kappa}\right) \geq 1 - s_n P\left(\left|w_j - \hat{w}_j\right| \geq c_3 N^{-\kappa}\right) \\
&\geq 1 - C_2 s_n N \exp\left(-c_5 N^{\gamma-\kappa}\right).
\end{aligned}$$

以上就完成了定理 2.1 的证明。

参考文献

- [1] 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>
- [2] 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>
- [3] 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>
- [4] Fan, J. and Song, R. (2010) Sure Independence Screening in Generalized Linear Models with Np-Dimensionality. *The Annals of Statistics*, **38**, 3567-3604. <https://doi.org/10.1214/10-aos798>
- [5] Fan, J., Feng, Y. and Song, R. (2011) Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models. *Journal of the American Statistical Association*, **106**, 544-557. <https://doi.org/10.1198/jasa.2011.tm09779>

- [6] Zhu, L., Li, L., Li, R. and Zhu, L. (2011) Model-Free Feature Screening for Ultrahigh-Dimensional Data. *Journal of the American Statistical Association*, **106**, 1464-1475. <https://doi.org/10.1198/jasa.2011.tm10563>
- [7] Li, R., Zhong, W. and Zhu, L. (2012) Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, **107**, 1129-1139. <https://doi.org/10.1080/01621459.2012.695654>
- [8] Li, G., Peng, H., Zhang, J. and Zhu, L. (2012) Robust Rank Correlation Based Screening. *The Annals of Statistics*, **40**, 1846-1877. <https://doi.org/10.1214/12-aos1024>
- [9] Wu, Y. and Yin, G. (2015) Conditional Quantile Screening in Ultrahigh-Dimensional Heterogeneous Data. *Biometrika*, **102**, 65-76. <https://doi.org/10.1093/biomet/asu068>
- [10] Zhou, Y., Liu, J., Hao, Z. and Zhu, L. (2019) Model-Free Conditional Feature Screening with Exposure Variables. *Statistics and Its Interface*, **12**, 239-251. <https://doi.org/10.4310/sii.2019.v12.n2.a5>
- [11] Wang, H. and Xia, Y. (2009) Shrinkage Estimation of the Varying Coefficient Model. *Journal of the American Statistical Association*, **104**, 747-757. <https://doi.org/10.1198/jasa.2009.0138>
- [12] Fan, J. and Ren, Y. (2006) Statistical Analysis of DNA Microarray Data in Cancer Research. *Clinical Cancer Research*, **12**, 4469-4473. <https://doi.org/10.1158/1078-0432.ccr-06-1033>
- [13] Fan, J., Ma, Y. and Dai, W. (2014) Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Varying Coefficient Models. *Journal of the American Statistical Association*, **109**, 1270-1284. <https://doi.org/10.1080/01621459.2013.879828>
- [14] Hall, P. and Miller, H. (2009) Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems. *Journal of Computational and Graphical Statistics*, **18**, 533-550. <https://doi.org/10.1198/jcgs.2009.08041>
- [15] Luo, S. and Chen, Z. (2014) Sequential Lasso Cum EBIC for Feature Selection with Ultra-High Dimensional Feature Space. *Journal of the American Statistical Association*, **109**, 1229-1240. <https://doi.org/10.1080/01621459.2013.877275>
- [16] Chen, X. and Xie, M. (2014) A Split-and-Conquer Approach for Analysis of. *Statistica Sinica*, **24**, 1655-1684. <https://doi.org/10.5705/ss.2013.088>
- [17] Xu, C., Zhang, Y., Li, R. and Wu, X. (2016) On the Feasibility of Distributed Kernel Regression for Big Data. *IEEE Transactions on Knowledge and Data Engineering*, **28**, 3041-3052. <https://doi.org/10.1109/tkde.2016.2594060>
- [18] Jordan, M.I., Lee, J.D. and Yang, Y. (2018) Communication-Efficient Distributed Statistical Inference. *Journal of the American Statistical Association*, **114**, 668-681. <https://doi.org/10.1080/01621459.2018.1429274>
- [19] Gonçalves, A.R., Liu, X. and Banerjee, A. (2019) Two-Block vs. Multi-Block ADMM: An Empirical Evaluation of Convergence. arXiv: 1907.04524.
- [20] Chen, J. and Chen, Z. (2008) Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, **95**, 759-771. <https://doi.org/10.1093/biomet/asn034>
- [21] Lin, N. and Xi, R. (2011) Aggregated Estimating Equation Estimation. *Statistics and Its Interface*, **4**, 73-83. <https://doi.org/10.4310/sii.2011.v4.n1.a8>
- [22] Wang, H. (2012) Factor Profiled Sure Independence Screening. *Biometrika*, **99**, 15-28. <https://doi.org/10.1093/biomet/asr074>
- [23] Li, K. (1991) Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, **86**, 316-327. <https://doi.org/10.2307/2290563>
- [24] Ruppert, D., Sheather, S.J. and Wand, M.P. (1995) An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*, **90**, 1257-1270. <https://doi.org/10.1080/01621459.1995.10476630>
- [25] Zhang, J., Zhang, R. and Lu, Z. (2015) Quantile-Adaptive Variable Screening in Ultra-High Dimensional Varying Coefficient Models. *Journal of Applied Statistics*, **43**, 643-654. <https://doi.org/10.1080/02664763.2015.1072141>
- [26] Zhang, J., Zhang, R. and Zhang, J. (2017) Feature Screening for Nonparametric and Semiparametric Models with Ultrahigh-Dimensional Covariates. *Journal of Systems Science and Complexity*, **31**, 1350-1361. <https://doi.org/10.1007/s11424-017-6310-6>
- [27] Hoeffding, W. (1948) A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, **19**, 293-325. <https://doi.org/10.1214/aoms/1177730196>
- [28] Wu, X. and Zhang, J. (2017) Researches on Rademacher Complexities in Statistical Learning Theory: A Survey. *Acta Automatica Sinica*, **43**, 20-39.
- [29] Schechtman, E. and Yitzhaki, S. (1999) On the Proper Bounds of the Gini Correlation. *Economics Letters*, **63**, 133-138. [https://doi.org/10.1016/s0165-1765\(99\)00033-6](https://doi.org/10.1016/s0165-1765(99)00033-6)