关于累积差异的再次讨论

周亭攸¹, 朱利平^{2*}

1. 浙江财经大学数据科学学院, 杭州 310018;

2. 中国人民大学统计与大数据研究院, 北京 100872

E-mail: zhoutingyou@163.com, zhu.liping@ruc.edu.cn

收稿日期: 2019-07-24; 接受日期: 2020-01-21; 网络出版日期: 2021-02-01; * 通信作者

国家自然科学基金 (批准号: 11801501, 11731011 和 11931014) 和北京市自然科学基金 (批准号: Z190002) 资助项目

摘要 本文再次讨论累积差异——一种关于条件均值独立的稳健度量. 首先证明, 若 $E(Y | X)$ 不依赖于 X , 则相应的累积差异的样本估计是 n 相合的; 否则是根号 n 相合的. 本文使用自助法近似其渐近零分布, 并证明该自助法的相合性. 其次, 本文使用累积差异作为边际效用的度量, 对超高维均值回归中的有用协变量进行筛选. 本文证明该方法的一致排序性和确定筛选性, 并通过大量的数值模拟展示其稳健的有限样本性质.

关键词 自助法 均值独立 累积差异 确定筛选性 超高维

MSC (2020) 主题分类 62G10

1 引言

记 $X \in \mathbb{R}^1$ 和 $Y \in \mathbb{R}^1$ 为两个随机变量, 度量和检验 X 与 Y 之间的相互依赖关系是统计研究中的一个基本问题. Pearson 相关系数可能是应用最广泛的关于 X 与 Y 之间的线性相关性的度量之一. 随后发展出了 Kendall- τ 相关系数 (参见文献 [1]) 及其拓展版本 (参见文献 [2])、Spearman- ρ 相关系数 (参见文献 [3])、Hoeffding 指数 (参见文献 [4]) 及其拓展版本 (参见文献 [5, 6]), 以及距离相关系数 (参见文献 [7, 8]). 其中 Kendall- τ 和 Spearman- ρ 是度量 X 与 Y 之间的序相关性, 其余的则是用来度量 X 与 Y 之间的统计独立性.

Shao 和 Zhang^[9] 引入了一个重要概念, 即均值独立性. 具体来说, 如果 $E(Y | X)$ 不随 X 变化, 则 Y 是均值独立于 X 的, 即 $E(Y | X) = E(Y)$. 假设 $(\tilde{X}, \tilde{Y})$ 是 (X, Y) 的一个独立副本, $d\text{var}(X)$ 是 X 的距离方差, $I(\cdot)$ 是示性函数, $\|\cdot\|$ 代表 Euclid 范数. 他们定义鞅差相关系数 (martingale difference correlation, MDC) 如下:

$$\text{MDC}(Y | X) \stackrel{\text{def}}{=} \frac{[-E\{Y - E(Y)\}\{\tilde{Y} - E(Y)\}\|X - \tilde{X}\|]^{1/2}}{\{\text{var}(Y)d\text{var}(X)\}^{1/2}}, \quad (1.1)$$

英文引用格式: Zhou T Y, Zhu L P. Cumulative divergence revisited (in Chinese). Sci Sin Math, 2021, 51: 2049–2064, doi: 10.1360/SCM-2019-0450

随后, Zhou 等^[10] 提出使用累积差异

$$\text{CD}(Y | X) \stackrel{\text{def}}{=} \frac{\text{CCov}(Y | X)}{\text{var}(Y)} \quad (1.2)$$

来检验 $E(Y | X) = E(Y)$, 其中, $\text{CCov}(Y | X)$ 表示 Y 在给定 X 条件下的累积协方差, 其具体形式定义如下:

$$\text{CCov}(Y | X) \stackrel{\text{def}}{=} E[\text{cov}^2\{Y, \mathbf{1}(X < \tilde{X}) | \tilde{X}\}]. \quad (1.3)$$

如果 (X, Y) 为联合正态分布, 则秩差相关系数和累积差异都等价于 Pearson 相关系数的平方. 但由于累积差异在计算过程中使用的是 X 的观测值的秩而不是其观测值本身, 因此对于 X 的单调变换具有不变性, 并在 X 的观测值中存在异常值或极端值时具有稳健性. 所以建议在异常值或极端值存在的情形下使用累积差异而非秩差相关系数.

给定一个随机样本 $\{(X_i, Y_i), i = 1, \dots, n\}$, 累积差异 $\text{CD}(Y | X)$ 的样本估计如下:

$$\widehat{\text{CD}}(Y | X) \stackrel{\text{def}}{=} \frac{n^{-3} \sum_{j=1}^n [\sum_{i=1}^n (Y_i - \bar{Y}) \{I(X_i < X_j) - F_n(X_j)\}]^2}{\widehat{\text{var}}(Y)}, \quad (1.4)$$

其中 $\bar{Y} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n Y_i$, $F_n(X_j) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n I(X_i < X_j)$, $\widehat{\text{var}}(Y) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. 本文有如下两个主要结果:

(1) 我们证明了, 当 $E(Y | X) = E(Y)$ 时, $\widehat{\text{CD}}(Y | X)$ 是 n 相合的; 否则是根号 n 相合的. 因为其渐近零分布是无穷多个独立的 $\chi^2(1)$ 随机变量的加权和, 且其权重取决于 (X, Y) 的联合分布, Zhou 等^[10] 提出使用自助法来近似该渐近零分布, 但却没有提供理论支撑. 本文在理论上证明了该自助法的相合性.

(2) 我们建议使用累积差异代替 Pearson 相关系数, 进行 Fan 和 Lv^[11] 提出的独立筛选并尽可能多地过滤掉不重要的协变量. 这是因为在超高维均值回归中, 由于数据集太大, 协变量中极有可能存在异常值或极端值. 在一定正则条件下, 我们证明了该方法的一致排序性和确定筛选性, 并通过大量的数值模拟展示了其稳健的有限样本性质.

需要强调的是, 一般来说, $\text{CD}(Y | X) \neq \text{CD}(X | Y)$, 即累积差异是不对称的. 当且仅当 $E(Y | X) = E(Y)$ 时, $\text{CD}(Y | X) = 0$, 而 $E(Y | X) = E(Y)$ 并不能推出 $\text{CD}(X | Y) = 0$. 例如, $Y = \exp(X)\varepsilon$, 其中 X 和 ε 独立地来自于正态分布. 显然, $E(Y | X) = E(Y) = 0$, 但是 $E(X | Y) \neq E(X) = 0$. 特别地, 如果 X 和 Y 相互独立, 那么 $\text{CD}(Y | X) = \text{CD}(X | Y) = 0$. Zhu 等^[12] 建议使用 $\text{CD}(X | Y)$ 度量 X 和 Y 之间的独立性. 本文主张使用 $\text{CD}(Y | X)$, 而不是 $\text{CD}(X | Y)$, 来度量 X 和 Y 之间的均值独立性. 这是因为 $\text{CD}(Y | X)$ 能够专门并彻底地刻画 $E(Y | X) = E(Y)$ 的偏离程度.

本文余下内容的组织如下. 第 2 节简要回顾累积差异的概念, 并彻底研究其理论性质, 证明自助法在检验条件均值独立中的相合性. 第 3 节提出一种基于累积差异的独立筛选方法, 并严格证明其一致排序性和确定筛选性. 第 4 节对本文进行总结. 所有技术证明均在附录中.

2 均值独立性检验

如无特殊说明, 后文中均假定 X 是连续的随机变量且 $\text{var}(X) > 0$, $0 < \text{var}(Y) < \infty$. 根据定义,

累积差异为 $\text{CD}(Y | X) \stackrel{\text{def}}{=} \text{CCov}(Y | X) / \text{var}(Y)$. 给定样本观察 $\{(X_i, Y_i), i = 1, \dots, n\}$, 我们使用矩方法估计 $\text{CD}(Y | X)$ 并将相应的估计量记为 $\widehat{\text{CD}}(Y | X)$, 其具体形式参见 (1.4). 我们用 “ $\xrightarrow{d}$ ” 表示 “依分布收敛”. 定理 2.1 是 $\widehat{\text{CD}}(Y | X)$ 的依分布收敛结果.

定理 2.1 (1) 若 $\text{E}(Y | X) = \text{E}(Y)$, 则 $\text{CD}(Y | X) = 0$ 且

$$n\kappa^{-1}\{\widehat{\text{CD}}(Y | X)\} \xrightarrow{d} Q \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1),$$

其中 $\kappa \stackrel{\text{def}}{=} \text{E}\{[Y - \text{E}(Y)]^2\}[\text{E}\{I(X < \tilde{X}) - F(\tilde{X})\}^2] / \text{var}(Y)$, $\chi_j^2(1)$ 是独立的自由度为 1 的卡方随机变量, λ_j 是依赖于 (X, Y) 的联合分布的非负常数且 $\text{E}(Q) = 1$. 进一步地, 若假定 $\text{E}(Y^2 | X) = \text{E}(Y^2)$, 则 κ 可简化为 $\kappa = \text{E}\{F(X) - F^2(X)\}$.

(2) 若 $\text{E}(Y | X) \neq \text{E}(Y)$, 则 $\text{CD}(Y | X) > 0$ 且

$$n^{1/2}\{\widehat{\text{CD}}(Y | X) - \text{CD}(Y | X)\} \xrightarrow{d} \mathcal{N}\left\{0, \frac{\varsigma_0^2}{\text{var}^2(Y)}\right\},$$

其中 ς_0^2 的定义参见 (A.3).

定理 2.1 表明, 当且仅当 Y 均值独立于 X 时, 给定 X 的条件下 Y 的累积差异为 0. 该性质保证了基于累积差异的检验的相合性. 于是, 当 $\widehat{\text{CD}}(Y | X)$ 足够大时, 即 $n \widehat{\text{CD}}(Y | X)$ 大于某个临界值 c_α 时, 我们拒绝 $H_0: \text{E}(Y | X) = \text{E}(Y)$. 根据定理 2.1(2), 可计算基于累积差异的检验的功效:

$$\begin{aligned} \beta(\varphi) &\stackrel{\text{def}}{=} \Pr(n \widehat{\text{CD}}(Y | X) > c_\alpha | \text{CD}(Y | X) = \varphi > 0) \\ &= \Pr\left[\frac{n^{1/2}\{\widehat{\text{CD}}(Y | X) - \varphi\}}{\sigma} > \frac{c_\alpha - n\varphi}{n^{1/2}\sigma} \mid \text{CD}(Y | X) = \varphi > 0\right] \\ &\approx 1 - \Phi\left(\frac{c_\alpha - n\varphi}{n^{1/2}\sigma}\right) \rightarrow 1, \quad \text{当 } n \rightarrow \infty \text{ 时,} \end{aligned}$$

其中 $\sigma = \varsigma_0 / \text{var}(Y)$, $\Phi(\cdot)$ 是标准正态 $\mathcal{N}(0, 1)$ 的累积分布函数.

由于上述检验的渐近零分布是无穷多个独立卡方随机变量的加权和, 且权重 λ_j 没有显式表达, 所以很难确定检验的临界值. 为了近似该渐近零分布, Zhou 等^[10] 提出了以下自助法:

(1) 基于样本观测 $\{(X_i, Y_i), i = 1, \dots, n\}$ 估计 $\text{CD}(Y | X)$, 并将相应的估计量记作 $\widehat{\text{CD}}(Y | X)$.

(2) 定义 $Y_i^* \stackrel{\text{def}}{=} \bar{Y} + a_i \varepsilon_i$, 其中 $\varepsilon_i \stackrel{\text{def}}{=} Y_i - \bar{Y}$, 且 a_i 为独立权重并满足 $\Pr(a_i = 1) = \Pr(a_i = -1) = 1/2$, $i = 1, \dots, n$. 基于自助法产生的样本 $\{(X_i, Y_i^*), i = 1, \dots, n\}$ 重新估计 $\text{CD}(Y^* | X)$, 并将相应的估计量记作 $\widehat{\text{CD}}(Y^* | X)$.

(3) 将步骤 2 重复 m 次得到 $\widehat{\text{CD}}^{(k)}(Y^* | X)$, 其中 $k = 1, \dots, m$. 最后用 $\widehat{\text{CD}}^{(k)}(Y^* | X)$ ($k = 1, \dots, m$) 的经验分布函数近似 $\widehat{\text{CD}}(Y | X)$ 的渐近零分布.

上述检验的 p 值则定义为

$$p \text{ 值} \stackrel{\text{def}}{=} m^{-1} \sum_{k=1}^m I\{\widehat{\text{CD}}^{(k)}(Y^* | X) \geq \widehat{\text{CD}}(Y | X)\}.$$

自始至终, 固定 $m = 1,000$. 当检验的 p 值小于事先给定的显著性水平 α 时, 拒绝 $H_0: \text{E}(Y | X) =$

$E(Y)$.

在上述自助法中, a_i 与 (ε_i, X_i) 独立且满足 $E(a_i) = 0$. 因此, $E(a_i \varepsilon_i | X_i) = 0$. 再加上 Y_i^* 和 $\bar{Y}$ 定义, 能够保证

$$E(Y_i^* | X_i) = n^{-1} \sum_{i=1}^n E(Y_i | X_i).$$

根据大数定律可知, 上式右边依概率收敛到 $E(Y)$. 于是我们有理由期待, 当 n 足够大时, $E(Y^* | X)$ 不随着 X 的变化而变化, 这意味着, 用自助法近似渐近零分布是合理的.

定理 2.2 严格说明了该自助法的相合性.

定理 2.2 假设 $\text{var}(X) > 0$, $0 < \text{var}(Y) < \infty$, 则

$$n\kappa^{-1}\{\widehat{\text{CD}}(Y^* | X)\} \xrightarrow{d} Q \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1),$$

其中 κ 和 λ_j 的定义参见定理 2.1.

下面进行数值模拟以展示基于累积差异的检验的有限样本性质.

例 2.1 设 $Y = c_0\{(X^2 + 1)^{-1} + 2 \sin X\} + \varepsilon$, 其中 X 来自于广义极值分布 $\text{GEV}(\mu, \sigma, \xi)$, 并且 μ 、 σ 和 ξ 分别表示位置、尺度和形状参数. 考虑以下两种情形:

(I) $\mu = 0$, $\sigma = 1$ 且 $\xi = 0$, 于是 X 的累积分布函数为 $F(x) = \exp\{-\exp(x)\}$, 进而有 $E(X) = \gamma$, $\text{var}(X) = \pi^2/6$, 其中 γ 为 Euler 常数.

(II) $\mu = 0$, $\sigma = 1$ 且 $\xi = 1$, 于是 X 的累积分布函数为 $F(x) = \exp\{-(1+x)^{-1}\}$, 进而有 $E(X) = \infty$.

在上述两种情形中, ε 都独立地来自于标准正态分布. 考察 $c_0 = 0, 0.2, 0.4, 0.6, 0.8, 1.0$. 显然, 当且仅当 $c_0 = 0$ 时, 原假设 $H_0: E(Y | X) = E(Y)$ 成立. 固定样本量 $n = 100$, 并将每个实验重复 500 次.

我们比较了以下 6 种检验的经验大小和功效: 基于累积差异 (cumulative divergence, CD) 的检验、基于 Pearson 相关系数的检验 (简称 ρ)、基于 Kendall- τ 序相关的检验 (简称 τ)、基于 MDC 的检验、基于距离相关系数 (distance correlation, DC) 的检验和 Hoeffding 独立性检验 (简称 H), 结果分别总结在表 1 和 2 中.

表 1 例 2.1 中不同检验的经验大小. 取 $c_0 = 0$ 并考察不同的显著性水平 $\alpha = 0.01, 0.02, 0.05, 0.10$

α	情形 (I)						情形 (II)					
	CD	ρ	τ	MDC	DC	H	CD	ρ	τ	MDC	DC	H
0.01	0.014	0.018	0.018	0.016	0.018	0.016	0.014	0.010	0.018	0.018	0.012	0.016
0.02	0.026	0.024	0.026	0.032	0.028	0.024	0.026	0.016	0.026	0.024	0.028	0.024
0.05	0.054	0.046	0.060	0.054	0.062	0.064	0.054	0.040	0.060	0.042	0.046	0.064
0.10	0.108	0.122	0.116	0.100	0.100	0.120	0.108	0.088	0.116	0.096	0.092	0.120

表 2 例 2.1 中不同检验的经验功效. 取显著性水平 $\alpha = 0.05$ 并考察 $c_0 = 0.2, 0.4, 0.6, 0.8, 1.0$

c_0	情形 (I)						情形 (II)					
	CD	ρ	τ	MDC	DC	H	CD	ρ	τ	MDC	DC	H
0.2	0.424	0.282	0.474	0.386	0.382	0.496	0.108	0.016	0.122	0.094	0.086	0.216
0.4	0.946	0.542	0.892	0.928	0.922	0.960	0.394	0.008	0.184	0.210	0.178	0.610
0.6	1.000	0.676	0.976	1.000	1.000	1.000	0.778	0.004	0.306	0.352	0.346	0.906
0.8	1.000	0.732	1.000	1.000	1.000	1.000	0.962	0.002	0.438	0.508	0.494	0.986
1.0	1.000	0.768	1.000	1.000	1.000	1.000	0.996	0.004	0.568	0.604	0.602	1.000

表 1 显示, 上述 6 种检验的经验大小均十分接近给定的显著性水平, 因此具有可比性. 从表 2 中看到基于 Pearson 相关系数 (ρ) 的检验的功效最小, 这说明 Pearson 相关系数对非线性依赖性不敏感. 而基于 CD 和 H 检验在两种情形中都展示了可比的优秀表现, 并且在情形 (II) 中显著地优于所有其他 4 种检验. 这也验证了基于 CD 的检验和 H 检验的稳健性. 但是, 基于 CD 的检验针对的是均值独立性的检验, 而 H 检验则针对的是统计独立性的检验. 所以, 当均值独立性是我们主要感兴趣的方面时, 我们建议使用基于 CD 的检验.

3 超高维均值回归中的特征筛选

考虑均值回归模型

$$E(Y | \mathbf{x}) = E(Y | \mathbf{x}_{\mathcal{A}}),$$

其中 $\mathbf{x} = (X_1, \dots, X_p)^T$ 是一个超高维的协变量, Y 是响应变量, $\mathbf{x}_{\mathcal{A}}$ 代表由下标 $\mathcal{A}$ 索引的协变量, 这里

$$\mathcal{A} \stackrel{\text{def}}{=} \{l: E(Y | \mathbf{x}) \text{ 随着 } X_l \text{ 的变化而变化, 其中 } l = 1, \dots, p\}.$$

本节提出了一种独立筛选方法, 以尽可能多地剔除上述均值回归中不重要的协变量.

独立筛选的最终目的是在协变量的维数 p 远大于可得到的样本量 n 时, 找到一个索引集的估计 $\hat{\mathcal{A}}$, 使得其满足 $\mathcal{A} \subseteq \hat{\mathcal{A}}$. 在此方向上, 文献中已有许多研究. 例如, Fan 和 Lv^[11] 建议通过假设 $E(Y | \mathbf{x})$ 是 $\mathbf{x}_{\mathcal{A}}$ 的线性函数, 进而使用 Pearson 相关系数进行独立筛选. Li 等^[13] 建议在观测值可能存在极端值的情形下, 使用 Kendall- τ 相关系数代替 Pearson 相关系数进行独立筛选. Fan 和 Song^[14] 放宽了线性模型的约束并假定 $E(Y | \mathbf{x})$ 来自一个广义线性模型, Fan 等^[15] 进一步放宽约束至假定 $E(Y | \mathbf{x})$ 是一个非参数加法模型. Shao 和 Zhang^[9] 提出了使用秩差相关系数进行独立筛选. Zhu 等^[12] 提出的独立排序和筛选方法以及 Li 等^[16] 提出的基于距离相关系数的独立筛选方法也可以用于超高维均值回归.

在实际中, 当协变量维数快速增长时, 其中存在极端观测值的可能性会增大. 我们通过例 3.1 说明这种现象.

例 3.1 协变量 $\mathbf{x} = (X_1, \dots, X_p)^T$ 来自于自由度为 3 的多元 T 分布, 且其均值为 0、协方差矩阵为 $\Sigma \stackrel{\text{def}}{=} (0.5^{|i-j|})_{p \times p}$. 于是 $E(X_l) = 0$ 且 $\text{var}(X_l) = 3, l = 1, \dots, p$. 固定样本量 $n = 50$, 并考察不同的维数 p 分别为 10、100、1,000 和 5,000. 当维数 p 给定时, 定义 $V_{\max} \stackrel{\text{def}}{=} \max\{X_{i,l}, i = 1, \dots, n, l = 1, \dots, p\}$, $V_{\min} \stackrel{\text{def}}{=} \min\{X_{i,l}, i = 1, \dots, n, l = 1, \dots, p\}$. 我们重复实验 500 次, 并将 $V_{\min}$ 和 $V_{\max}$ 的最小值, 5%、25%、50%、75% 和 95% 的分位数, 以及最大值列示在表 3 中.

表 3 例 3.1 中, 基于 500 次的重复实验所得到的 $V_{\min}$ 和 $V_{\max}$ 的最小值, 5%、25%、50%、75% 和 95% 的分位数, 以及最大值

p	$V_{\min}$							$V_{\max}$						
	最小值	5%	25%	50%	75%	95%	最大值	最小值	5%	25%	50%	75%	95%	最大值
10	-62.8	-20.8	-11.5	-8.6	-6.7	-5.3	-4.0	3.8	5.0	6.9	8.7	11.3	19.7	42.9
100	-256.3	-42.8	-24.3	-18.5	-14.7	-11.3	-8.7	8.6	11.5	14.5	18.6	25.8	41.1	167.8
1000	-293.9	-88.2	-49.3	-38.9	-31.6	-24.5	-16.8	19.7	24.9	31.2	39.1	54.6	97.1	295.9
5000	-596.9	-152.4	-91.4	-69.5	-55.0	-42.7	-31.9	34.7	41.9	53.6	70.2	93.7	182.5	1607.9

表 3 显示, 随着协变量维数的增加, 观测到的协变量的值变得越来越极端. 具体而言, 当 $p = 10$ 时, $V_{\min}$ 的最小值为 -62.8 ; 当 p 增加至 100 时, 最小值降至 -256.3 . 更极端的值 -596.9 在协变量维数 p 急剧增加到 5,000 时出现. 可以想象当 p 更大时, 这种现象可能会更加严重. 此外, 当 $p = 5,000$ 时, 我们发现 $V_{\max}$ 的最大值 (1607.9) 远远大于其 95% 的分位数 (182.5), 这从另一方面表明, 即使真实的分布函数并非厚尾, 维数的激增也会导致出现极端值的可能大大增加.

虽然累积差异和秩差相关系数都是用来度量 Y 的条件均值依赖于 $\mathbf{x}$ 的程度, 并且都不需要假定 $E(Y | \mathbf{x})$ 的参数结构, 但由于累积差异在处理极端值时更加稳健, 因此建议使用累积差异 $CD(Y | X_l)$ ($l = 1, \dots, p$) 进行独立筛选, 用以过滤掉那些不重要的协变量. 我们建议的独立筛选方法如下:

- (1) 基于样本 $\{(\mathbf{x}_i, Y_i), i = 1, \dots, n\}$ 估计 $CD(Y | X_l)$, 相应的样本估计记作 $\widehat{CD}(Y | X_l)$;
- (2) 保留 $\mathbf{x}_{\hat{\mathcal{A}}}$ 作为有用的协变量, 其中 $\hat{\mathcal{A}} \triangleq \{l : \widehat{CD}(Y | X_l) \geq \varepsilon_n, l = 1, \dots, p\}$.

为了研究上述方法的筛选性质, 我们假定以下条件成立:

(C1) 存在一个常数 s_0 , 使得对所有的 $0 < s < s_0$, 都有 $E\{\exp(sY^2)\} < \infty$ 成立;

(C2) 存在某常数 $c > 0$ 和 $0 \leq \kappa < 1/2$, 使得 $\min_{l \in \mathcal{A}} CD(Y | X_l) - \max_{l \in \mathcal{A}^c} CD(Y | X_l) = cn^{-\kappa}$ 成立;

(C3) 存在某常数 $c > 0$ 和 $0 \leq \kappa < 1/2$, 使得 $\min_{l \in \mathcal{A}} CD(Y | X_l) \geq 2cn^{-\kappa}$ 成立.

这些条件不是最弱的条件, 假定它们只是为了方便技术推导. 当 Y 为正态分布时, 条件 (C1) 成立; 而条件 (C2) 要求重要的协变量和不重要的协变量是可分离的; 条件 (C3) 则要求重要协变量的信号强度足够大.

定理 3.1 总结了我们的筛选方法的理论性质.

定理 3.1 假定条件 (C1) 成立, 定义 $\alpha_n \triangleq \exp\{-c_1 n^{1-2(\gamma+\kappa)}\} + n(n-1)\exp(-c_2 n^\gamma)$, 其中 γ 是一个正的常数且满足 $\gamma + \kappa < 1/2$, c_1 和 c_2 是正的常数, 则

$$\Pr\left\{\max_{1 \leq l \leq p} |\widehat{CD}(Y | X_l) - CD(Y | X_l)| \geq cn^{-\kappa}\right\} \leq O(\alpha_n p).$$

(1) 一致排序性: 若更进一步地假定条件 (C2) 成立, 则

$$\Pr\left\{\min_{l \in \mathcal{A}} \widehat{CD}(Y | X_l) > \max_{l \in \mathcal{A}^c} \widehat{CD}(Y | X_l)\right\} > 1 - O(\alpha_n p);$$

(2) 确定筛选性: 若更进一步地假定条件 (C3) 成立, 则

$$\Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}) \geq 1 - O(\alpha_n |\mathcal{A}|).$$

定理 3.1 说明我们提出的方法在排序和筛选上都具有相合性, 这两者可以确保, 在渐近意义上, 所有真正有用的协变量都能被保留下来.

接下来, 将通过大量的数值模拟展示基于累积差异的独立筛选方法 (cumulative divergence based sure independent screening, CD-SIS) 的有限样本性质. 我们将 CD-SIS 与以下 6 种筛选方法进行比较: 由 Fan 和 Lv^[11] 提出的基于 Pearson 相关系数的筛选方法 (sure independent screening, SIS)、由 Fan 等^[15] 提出的针对非参数模型的筛选方法 (nonparametric independent screening, NIS)、由 Zhu 等^[12] 提出的不依赖于模型的确切排序与筛选方法 (sure independent ranking and screening, SIRS)、由 Li 等^[16] 提出的基于距离相关系数的筛选方法 (distance correlation based sure independent screening, DC-SIS)、

由 Li 等^[13]提出的基于 Kendall- τ 相关系数的筛选方法 (简称 τ -SIS) 以及 Shao 和 Zhang^[9]提出的基于鞅差相关系数的筛选方法 (martingale difference correlation based sure independent screening, MDC-SIS). 我们通过以下 3 个标准评价各个独立筛选方法的表现:

(1) $\mathcal{S}$: 为了确保有用的协变量都被筛选出来所需要的最小的模型大小. 该值越小说明筛选方法表现越好. 我们列示了基于 500 重复实验所得到的 $\mathcal{S}$ 的 5%、25%、50%、75% 和 95% 的分位数.

(2) $\mathcal{P}_{\text{ind}}$: 给定最终的模型大小 s , 各个有用的协变量被筛选出来的经验概率.

(3) $\mathcal{P}_{\text{all}}$: 给定最终的模型大小 s , 所有有用的协变量都被筛选出来的经验概率.

考察给定不同的模型大小的情形: $s_1 = [n/\log(n)]$, $s_2 = 2s_1$ 和 $s_3 = 3s_1$, 其中 $[x]$ 表示不超过 x 的最大整数.

例 3.2 考察 4 个不同的模型. 前 3 个模型的协变量 $\mathbf{x} = (X_1, \dots, X_p)^T$ 来自于均值向量为零、协方差矩阵为 $\Sigma = (0.5^{|i-j|})_{p \times p}$ 的多元正态分布, 独立的随机误差项 ε 则来自于标准正态分布. 在模型 (3.1) 中, 考虑简单的线性模型

$$Y = X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5 + \varepsilon. \quad (3.1)$$

在模型 (3.2) 中, 响应变量 Y 来自于 Bernoulli 分布且成功的概率为

$$\Pr(Y = 1 | \mathbf{x}) = \{1 + \exp(-X_1 - 0.8X_2 - 0.6X_3 - 0.4X_4 - 0.2X_5)\}^{-1}. \quad (3.2)$$

模型 (3.3) 是一个非线性异方差模型

$$Y = \frac{(X_1 + X_2)^2}{2} + \exp\left\{\frac{X_{21} + X_{22} + X_{23}}{3}\right\}\varepsilon. \quad (3.3)$$

在模型 (3.4) 中, 协变量 $\mathbf{x} = (X_1, \dots, X_p)^T$ 来自自由度为 1 的多元 T 分布, 独立的随机误差项 ε 来自于标准正态分布. 考察如下回归模型:

$$Y = \frac{3}{(1 + X_1^2)} + \frac{2}{1 + \exp(-X_8 - X_9 - X_{10})} + \text{sign}(X_{20}) + \varepsilon. \quad (3.4)$$

自始至终, 固定样本量 $n = 200$ 和协变量的维数 $p = 1,000$, 并将每个实验重复 500 次. 所得到关于 $\mathcal{P}_{\text{ind}}$ 、 $\mathcal{P}_{\text{all}}$ 和 $\mathcal{S}$ 的结果分别汇总在表 4 和 5 中.

从表 5 可以看出, 所有的筛选方法在线性模型 (3.1) 和广义线性模型 (3.2) 中都具有可比的优良表现. 这种现象并不奇怪, 因为当 $\mathbf{x}$ 和 Y 满足联合正态分布时 (在模型 (3.1) 中就是这种情形), 这些筛选方法所使用的边际效用实际上是等价的. 但在非线性模型 (3.3) 和 (3.4), 这些筛选方法则是完全不同的. 在模型 (3.3) 中, SIS、SIRS 和 τ -SIS 完全无法筛选出真正重要的协变量, 而 CD-SIS、NIS、DC-SIS 和 MDC-SIS 则表现相对良好. 但是从表 4 中, 我们发现 DC-SIS 会以较大的概率保留了协变量 X_{21} 、 X_{22} 和 X_{23} , 这些都是对响应变量的条件均值没有影响 (但对响应变量的条件分布有影响) 的协变量. 相比之下, CD-SIS 则仅保留对响应变量的条件均值有影响的协变量.

NIS 和 MDC-SIS 在模型 (3.3) 中表现良好. 但在 (3.4) 中, CD-SIS 是唯一的赢家: $\mathcal{S}$ 的中值小至 11, 这与真正重要的协变量的个数非常接近; 而 $\mathcal{S}$ 的四分位数仅为 21, 这表明 CD-SIS 的性能十分稳健. 由于该模型中协变量 $\mathbf{x}$ 来自于自由度为 1 的多元 T 分布, 所以其中存在的极端值会大大降低

表 4 例 3.2 中, 给定模型大小 s , 基于 500 次重复实验所得到的 $\mathcal{P}_{\text{ind}}$ 和 $\mathcal{P}_{\text{all}}$ 的经验概率

s	方法	(3.1)						(3.2)					
		X_1	X_2	X_3	X_4	X_5	$\mathcal{P}_{\text{all}}$	X_1	X_2	X_3	X_4	X_5	$\mathcal{P}_{\text{all}}$
s_1	CD-SIS	1.00	1.00	1.00	1.00	0.97	0.97	1.00	1.00	1.00	0.99	0.77	0.77
	SIS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	0.81	0.81
	NIS	1.00	1.00	1.00	1.00	0.94	0.94	1.00	1.00	1.00	0.98	0.64	0.63
	SIRS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	0.81	0.81
	DC-SIS	1.00	1.00	1.00	1.00	0.97	0.97	1.00	1.00	1.00	1.00	0.78	0.78
	τ -SIS	1.00	1.00	1.00	1.00	0.97	0.97	1.00	1.00	1.00	1.00	0.78	0.78
	MDC-SIS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	0.78	0.78
s_2	CD-SIS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	0.84	0.84
	SIS	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.87	0.87
	NIS	1.00	1.00	1.00	1.00	0.96	0.96	1.00	1.00	1.00	0.99	0.76	0.75
	SIRS	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.87	0.87
	DC-SIS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	0.85	0.85
	τ -SIS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	0.87	0.87
	MDC-SIS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	0.85	0.85
s_3	CD-SIS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	0.89	0.89
	SIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.91
	NIS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	0.99	0.81	0.80
	SIRS	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.91	0.91
	DC-SIS	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	0.89	0.89
	τ -SIS	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.90	0.90
	MDC-SIS	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.89	0.89
s	方法	(3.3)						(3.4)					
		X_1	X_2	X_{21}	X_{22}	X_{23}	$\mathcal{P}_{\text{all}}$	X_1	X_8	X_9	X_{10}	X_{20}	$\mathcal{P}_{\text{all}}$
s_1	CD-SIS	0.98	0.99	0.06	0.05	0.05	0.98	1.00	0.89	0.98	0.89	1.00	0.79
	SIS	0.21	0.21	0.10	0.10	0.09	0.12	0.04	0.27	0.41	0.29	0.48	0.00
	NIS	1.00	1.00	0.20	0.24	0.22	1.00	0.07	0.25	0.42	0.30	0.58	0.01
	SIRS	0.06	0.07	0.37	0.51	0.37	0.02	0.04	0.25	0.39	0.28	0.48	0.00
	DC-SIS	1.00	1.00	0.62	0.78	0.63	1.00	0.71	0.47	0.69	0.51	0.87	0.20
	τ -SIS	0.07	0.09	0.05	0.05	0.05	0.03	0.03	0.88	0.98	0.90	1.00	0.03
	MDC-SIS	1.00	1.00	0.08	0.07	0.08	1.00	0.73	0.50	0.70	0.52	0.88	0.22
s_2	CD-SIS	1.00	1.00	0.11	0.11	0.12	1.00	1.00	0.94	0.99	0.94	1.00	0.88
	SIS	0.27	0.28	0.15	0.16	0.16	0.15	0.07	0.36	0.50	0.39	0.57	0.00
	NIS	1.00	1.00	0.29	0.32	0.30	1.00	0.11	0.33	0.49	0.37	0.64	0.02
	SIRS	0.10	0.12	0.62	0.76	0.61	0.04	0.08	0.35	0.49	0.37	0.55	0.00
	DC-SIS	1.00	1.00	0.82	0.90	0.80	1.00	0.77	0.59	0.76	0.59	0.89	0.28
	τ -SIS	0.12	0.15	0.10	0.08	0.08	0.06	0.07	0.94	0.99	0.94	1.00	0.07
	MDC-SIS	1.00	1.00	0.14	0.14	0.14	1.00	0.79	0.60	0.77	0.60	0.90	0.31
s_3	CD-SIS	1.00	1.00	0.16	0.15	0.15	1.00	1.00	0.96	0.99	0.95	1.00	0.91
	SIS	0.33	0.35	0.22	0.22	0.21	0.22	0.11	0.42	0.55	0.45	0.60	0.01
	NIS	1.00	1.00	0.36	0.38	0.35	1.00	0.18	0.39	0.54	0.42	0.67	0.02
	SIRS	0.15	0.18	0.76	0.84	0.75	0.06	0.10	0.42	0.55	0.45	0.60	0.01
	DC-SIS	1.00	1.00	0.89	0.94	0.89	1.00	0.82	0.64	0.80	0.65	0.91	0.37
	τ -SIS	0.17	0.20	0.13	0.13	0.11	0.08	0.11	0.95	0.99	0.95	1.00	0.10
	MDC-SIS	1.00	1.00	0.18	0.18	0.19	1.00	0.83	0.66	0.81	0.66	0.92	0.38

SIS、NIS、SIRS、DC-SIS 和 MDC-SIS 的性能. 此外, 从表 4 还能发现, τ -SIS、SIS、NIS 和 SIRS 都不能有效地筛选出对 Y 具有对称影响的协变量 X_1 .

例 3.3 我们通过对一个白血病数据集的分析说明我们的方法在实际中的应用. Golub 等^[17] 以及 Fan 和 Lv^[11] 也曾经对该数据集进行过分析. 它包含 72 个观测个体, 其中 47 例来自急性淋巴细胞

表 5 例 3.2 中, 基于 500 次重复实验所得到的最小模型大小 S 的 5%、25%、50%、75%、95% 的分位数

方法	(3.1)					(3.2)				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
CD-SIS	5	5	5	6	20	5	6	11	34	265
SIS	5	5	5	5	12	5	5	9	24	190
NIS	5	5	5	8	49	5	7	17	72	368
SIRS	5	5	5	6	21	5	5	9	24	190
DC-SIS	5	5	5	6	19	5	5	10	32	247
τ -SIS	5	5	5	6	19	5	5	10	31	225
MDC-SIS	5	5	5	5	16	5	5	10	32	247

方法	(3.3)					(3.4)				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
CD-SIS	2	2	3	7	24	5	7	11	28	232
SIS	11	140	474	772	958	251	563	772	912	987
NIS	2	2	2	2	5	256	518	756	886	976
SIRS	90	397	628	859	985	276	595	793	913	990
DC-SIS	2	2	2	2	4	9	51	214	459	764
τ -SIS	57	339	607	818	961	62	259	493	740	947
MDC-SIS	2	2	2	2	7	10	49	196	448	744

白血病 (ALL) 类, 剩下 25 例来自急性骨髓性白血病 (AML) 类; 每个观察个体包含 7,129 个基因表达水平, 即样本量 $n = 72$ 而协变量维数 $p = 7,129$. 我们将 ALL 类的前 27 个观测个体和 AML 类中的前 11 个观测个体设置为训练样本, 其余的设置为测试样本.

首先将 7 种筛选方法, 即 CD-SIS、SIS、NIS、SIRS、DC-SIS、 τ -SIS 和 MDC-SIS, 应用于样本量为 $n_1 = 38$ 的训练集, 进而筛选出一部分协变量作为有用的协变量. 文献中有两类阈值规则来决定一个合适的筛选大小 (即筛选出的有用协变量的个数). Zhu 等^[12] 提出的软阈值规则通过生成一系列与响应变量独立的辅助变量模拟协变量, 进而估计出有用协变量的最小边际效用, 并以此作为标准筛选出有用的协变量. Fan 和 Lv^[11] 建议的硬阈值规则通过将每个协变量的重要性进行从大到小排序, 并选择某一指定数量的协变量作为有用的协变量. 根据文献 [18] 和 [19, 表 5] 可知, 基因表达数据中只有非常少的基因对疾病有影响. 考虑到样本量 $n_1 = 38$ 相对较小, 我们使用硬阈值规则并保留前 $s \stackrel{\text{def}}{=} c_0[n_1/\log(n_1)]$ 个边际效用最大的基因作为有用的协变量, 其中 c_0 是一个正常数. 在这里, 我们选择 $c_0 = 0.5$ 并保留 $s = 5$ 个协变量以避免过度拟合的问题. 筛选后保留的基因列于表 6 中. 基于这些筛选出来的基因, 我们在训练集和测试集上分别进行线性判别分析, 所得到的经验误判概率列在表 7 中.

表 6 所列示的筛选结果完全不同. 一个自然的问题出现了: 哪种筛选结果最可靠? 从表 7 中可以

表 6 白血病数据集中, 通过不同的筛选方法得到有用基因的序号

CD-SIS	SIS	NIS	SIRS	DC-SIS	τ -SIS	MDC-SIS
1882	1745	2020	1745	1834	1834	1834
2020	2020	2267	2020	2020	1882	2020
3320	3320	3320	3320	3320	3320	3320
4847	4847	4499	4847	4847	4847	4847
6218	5039	4847	5039	5039	6218	5039

表 7 白血病数据集中, 在训练集和测试集上进行线性判别分析所得到的经验误判概率

	CD-SIS	SIS	NIS	SIRS	DC-SIS	τ -SIS	MDC-SIS
训练集上的误判率	1/38	0/38	0/38	0/38	0/38	1/38	0/38
测试集上的误判率	2/34	6/34	8/34	6/34	6/34	2/34	6/34

图 1 白血病数据集中, 不同的筛选方法所选出的有用基因的箱线图. (a) CD-SIS; (b) SIS; (c) NIS; (d) SIRS; (e) DC-SIS; (f) τ -SIS; (g) MDC-SIS

看出, 所有的筛选方法在训练集上都具有可比的优良表现, 但在测试集上, 只有 CD-SIS 和 τ -SIS 的经验误判率最小, 约为 5.88%. 图 1 是不同的筛选方法所选出的有用基因的箱线图, 从中能够发现, 这些基因的表达中可能存在极端值或异常值. 这就解释了为什么稳健的筛选方法 (如 CD-SIS 和 τ -SIS) 的表现优于其他筛选方法.

4 总结

累积差异是一个关于条件均值独立性的稳健度量, 本文重新探讨了累积差异并得到如下研究成果. 首先, 在理论上证明了累积差异检验中自助法的相合性. 此外, 我们还注意到, 随着协变量维数的上升, 其观测值中出现极端值的可能性也会逐渐增加. 因此, 在超高维均值回归问题中, 我们建议使用累积差异作为各个协变量边际效用的度量, 并进行独立筛选以过滤掉那些不重要的协变量.

本文中关于累积差异只讨论了 X 和 Y 都是一元随机变量的情形. 实际上, 只需要一些投影技巧, 就可以将累积差异推广到两个多元随机向量的情形, 这也是我们后续的研究方向之一.

致谢 感谢审稿人对我们的文章提出的许多优秀的修改意见.

参考文献

- 1 Kendall M G. A new measure of rank correlation. *Biometrika*, 1938, 30: 81–93
- 2 Bergsma W, Dassios A. A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 2014, 20: 1006–1028
- 3 Spearman C. The proof and measurement of association between two things. *Amer J Phys*, 1904, 15: 72–101
- 4 Hoeffding W. A non-parametric test of independence. *Ann Math Statist*, 1948, 19: 546–557

- 5 Blum J R, Kiefer J, Rosenblatt M. Distribution free tests of independence based on the sample distribution function. *Ann Math Statist*, 1961, 32: 485–498
- 6 Zhou Y Q, Zhu L P. Model-free feature screening for ultrahigh dimensional data through a modified Blum-Kiefer-Rosenblatt correlation. *Statist Sinica*, 2018, 28: 1351–1370
- 7 Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. *Ann Statist*, 2007, 35: 2769–2794
- 8 Székely G J, Rizzo M L. Brownian distance covariance. *Ann Appl Stat*, 2009, 3: 1236–1265
- 9 Shao X F, Zhang J S. Martingale difference correlation and its use in high-dimensional variable screening. *J Amer Statist Assoc*, 2014, 109: 1302–1318
- 10 Zhou T Y, Zhu L P, Xu C, et al. Model-free forward screening via cumulative divergence. *J Amer Statist Assoc*, 2020, 115: 1393–1405
- 11 Fan J Q, Lv J C. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B Stat Methodol*, 2008, 70: 849–911
- 12 Zhu L P, Li L X, Li R Z, et al. Model-free feature screening for ultrahigh-dimensional data. *J Amer Statist Assoc*, 2011, 106: 1464–1475
- 13 Li G R, Peng H, Zhang J, et al. Robust rank correlation based screening. *Ann Statist*, 2012, 40: 1846–1877
- 14 Fan J Q, Song R. Sure independence screening in generalized linear models with NP-dimensionality. *Ann Statist*, 2010, 38: 3567–3604
- 15 Fan J Q, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J Amer Statist Assoc*, 2011, 106: 544–557
- 16 Li R Z, Zhong W, Zhu L P. Feature screening via distance correlation learning. *J Amer Statist Assoc*, 2012, 107: 1129–1139
- 17 Golub T R, Slonim D, Tamayo P, et al. Gene expression pattern recognition for cancer classification. *Blood*, 1999, 94: 656A–657A
- 18 Gupta S, Chaudhary K, Kumar R, et al. Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine. *Sci Rep*, 2016, 6: 23857
- 19 Dong Z L, Zhang N Q, Li C, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer*, 2015, 15: 489
- 20 Stute W. Nonparametric model checks for regression. *Ann Statist*, 1997, 25: 613–641
- 21 Kuo H H. Gaussian Measures in Banach Spaces. *Lecture Notes in Mathematics*, vol. 463. Berlin: Springer, 1975
- 22 Serfling R J. Approximation Theorems of Mathematical Statistics. New York: John Wiley & Sons, 1980

附录 A

定理 2.1 的证明 不失一般性, 假定 $E(Y) = 0$. 定义

$$\begin{aligned}\zeta_n(x) &\stackrel{\text{def}}{=} n^{-1/2} \sum_{i=1}^n (Y_i - \bar{Y}) \{I(X_i < x) - F_n(x)\} \\ &= n^{-1/2} \sum_{i=1}^n \{Y_i - E(Y)\} \{I(X_i < x) - F(x)\} + o_p(1).\end{aligned}$$

根据文献 [10, 定理 1], 可得 $CD(Y | X) = 0$ 当且仅当 $E(Y | X) = E(Y)$ 成立. 记 $\zeta(\cdot)$ 为一个零均值的 Gauss 过程且其协方差函数定义如下:

$$\text{cov}\{\zeta(x_1), \zeta(x_2)\} \stackrel{\text{def}}{=} E[Y^2 \{I(X < x_1 \wedge x_2) - F(x_1)I(X < x_2) - F(x_2)I(X < x_1) + F(x_1)F(x_2)\}],$$

其中 $x_1, x_2 \in \mathbb{R}^1$. 对 $i = 1, \dots, n$, 引入以下分位数表示 $X_i = F^{-1}(U_i)$, 其中 U_i 是 $[0, 1]$ 上的均匀分布.

于是,

$$\sup_{x \in \mathbb{R}^1} n^{-1/2} \left| \sum_{i=1}^n \{Y_i - E(Y)\} \{I(X_i < x) - F(x)\} \right|, \quad (\text{A.1})$$

$$\sup_{0 \leq u \leq 1} n^{-1/2} \left| \sum_{i=1}^n \{Y_i - E(Y)\} \{I(U_i < u) - F_U(u)\} \right| \quad (\text{A.2})$$

是同分布的, 其中 $F_U(u) \stackrel{\text{def}}{=} u$. 根据 U 的定义可知, $E(Y_i | U_i) = E(Y_i)$. 与文献 [20, 定理 1.1] 的证明类似, 我们能够证明 (A.2) 的渐近性质. 因此, $\zeta_n(x)$ 的极限过程是一个零均值且协方差函数为 $\text{cov}\{\zeta(x_1), \zeta(x_2)\}$ 的 Gauss 过程, 即当 $H_0: E(Y | X) = E(Y)$ 时, 有 $\zeta_n(\cdot) \xrightarrow{d} \zeta(\cdot)$, 并且

$$n \widehat{\text{CCov}}(Y | X) = \int_{-\infty}^{\infty} \zeta_n^2(x) dF_n(x) \xrightarrow{d} \int_{-\infty}^{\infty} \zeta^2(x) dF(x).$$

根据文献 [21] 中第一章第二节可得

$$n \left[E \left\{ \int_{-\infty}^{\infty} \zeta^2(x) dF(x) \right\} \right]^{-1} \widehat{\text{CCov}}(Y | X) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1),$$

其中 $\chi_j^2(1)$ 是独立的自由度为 1 的卡方随机变量, λ_j 是依赖于 (X, Y) 的联合分布的非负常数. 于是得到

$$n\kappa^{-1} \{\widehat{\text{CD}}(Y | X)\} \xrightarrow{d} Q \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1),$$

其中 $\kappa \stackrel{\text{def}}{=} E[Y^2 \{I(X < \tilde{X}) - F(\tilde{X})\}^2] / \text{var}(Y)$ 且 $E(Q) = 1$. 如果 $E(Y^2 | X) = E(Y^2)$ 成立, 则 κ 可化简为 $\kappa = E\{F(X) - F^2(X)\}$. 下面证明当 $E(Y | X) \neq E(Y)$ 时 $\widehat{\text{CCov}}(Y | X)$ 的渐近正态性. 定义

$$T_1 \stackrel{\text{def}}{=} \frac{(n-1)(n-2)}{n^2} \left\{ \sum_{i \neq j \neq k} \frac{\omega_0(i, j, k) + \omega_0(j, i, k) + \omega_0(j, k, i)}{3n(n-1)(n-2)} \right\},$$

其中 $\omega_0(i, j, k) \stackrel{\text{def}}{=} Y_i Y_k \{I(X_i < X_j) - F(X_j)\} \{I(X_k < X_j) - F(X_j)\}$. 定义

$$T_2 \stackrel{\text{def}}{=} \frac{n-1}{n^2} \left\{ \sum_{i \neq j} \frac{Y_i^2 \{I(X_i < X_j) - F(X_j)\}^2 + Y_j^2 \{I(X_j < X_i) - F(X_i)\}^2}{2n(n-1)} \right\},$$

$$T_3 \stackrel{\text{def}}{=} \frac{1}{n^2} \left\{ \frac{1}{n} \sum_{i=1}^n Y_i^2 F^2(X_i) \right\}.$$

于是可得 $\widehat{\text{CCov}}(Y | X) = (T_1 + T_2 + T_3) \{1 + o_p(1)\}$. 可以发现, T_1 、 T_2 和 T_3 中括号里面都是 U -统计量. 根据文献 [22, 定理 5.4.A], 得到 $T_2 = O_p(n^{-1})$, $T_3 = O_p(n^{-2})$, 即 $T_2 = o_p(n^{-1/2})$ 且 $T_3 = o_p(n^{-1/2})$. 再由文献 [22, 定理 5.5.1], 可得 $n^{1/2} \{T_1 - \text{CCov}(Y | X)\} \xrightarrow{d} \mathcal{N}(0, \varsigma_0^2)$, 其中

$$\varsigma_0^2 \stackrel{\text{def}}{=} \text{var}[E\{\omega_0(i, j, k) + \omega_0(j, i, k) + \omega_0(j, k, i) | X_i, Y_i\}]. \quad (\text{A.3})$$

经过简单的计算知, $E\{\omega_0(i, j, k) + \omega_0(j, i, k) + \omega_0(j, k, i) \mid X_i, Y_i\}$ 是随着 (X_i, Y_i) 的变化而变化的, 也意味着 (A.3) 不会退化到 0. 再加上 $\widehat{\text{var}}(Y)$ 的弱收敛性和 Slutsky 定理, 我们能证得 $\widehat{\text{CD}}(Y \mid X)$ 的渐近正态性. $\square$

定理 2.2 的证明 不失一般性, 假定 $E(Y) = 0$. 记 $\zeta^*(\cdot)$ 为一个零均值的 Gauss 过程且其协方差函数定义如下:

$$\text{cov}\{\zeta^*(x_1), \zeta^*(x_2)\} \stackrel{\text{def}}{=} E[Y^{*2}\{I(X < x_1 \wedge x_2) - F(x_1)I(X < x_2) - F(x_2)I(X < x_1) + F(x_1)F(x_2)\}],$$

其中 $x_1, x_2 \in \mathbb{R}^1$. 定义

$$\begin{aligned}\zeta_n^*(x) &\stackrel{\text{def}}{=} n^{-1/2} \sum_{i=1}^n (Y_i^* - \bar{Y}^*) \{I(X_i < x) - F_n(x)\} \\ &= n^{-1/2} \sum_{i=1}^n \{Y_i^* - E(Y^*)\} \{I(X_i < x) - F(x)\} + o_p(1).\end{aligned}$$

根据 Y_i^* 的定义, 容易验证 $E(Y_i^* \mid X) = E(\bar{Y} + a_i \varepsilon_i \mid X_i) = 0$, 这说明当 n 足够大时, Y^* 的条件均值独立于 X . 与定理 2.1 中的讨论类似, 有

$$\begin{aligned}\sup_{x \in \mathbb{R}^1} n^{-1/2} \left| \sum_{i=1}^n \{Y_i^* - E(Y^*)\} \{I(X_i < x) - F(x)\} \right|, \\ \sup_{0 \leq u \leq 1} n^{-1/2} \left| \sum_{i=1}^n \{Y_i^* - E(Y^*)\} \{I(U_i < u) - F_U(u)\} \right|\end{aligned}$$

是同分布的. 于是可得 $\zeta_n^*(x)$ 的极限分布是一个零均值且协方差函数为 $\text{cov}\{\zeta^*(x_1), \zeta^*(x_2)\}$ 的 Gauss 过程. 进而有 $\zeta_n^*(\cdot) \xrightarrow{d} \zeta^*(\cdot)$ 且

$$n \widehat{\text{CCov}}(Y^* \mid X) = \int_{-\infty}^{\infty} \{\zeta_n^*(x)\}^2 dF_n(x) \xrightarrow{d} \int_{-\infty}^{\infty} \{\zeta^*(x)\}^2 dF(x).$$

下面证明 $\text{cov}\{\zeta^*(x_1), \zeta^*(x_2)\} = \text{cov}\{\zeta(x_1), \zeta(x_2)\}$ 成立, 其中等号右边已在定理 2.1 中定义. 根据重期望公式, 只要 $E(Y^{*2} \mid X) = E(Y^2 \mid X)$, 上式就成立. 根据定义可知, $Y^* = \bar{Y} + a\varepsilon$, 且 a_i 为独立权重且满足 $\Pr(a_i = \pm 1) = 1/2$. 因此,

$$E(Y^{*2} \mid X) = E(\bar{Y}^2 + 2\bar{Y}a\varepsilon + a^2\varepsilon^2 \mid X) = E(\bar{Y}^2 \mid X) + E(\varepsilon^2 \mid X) = E(Y^2 \mid X).$$

这说明 $\text{cov}\{\zeta^*(x_1), \zeta^*(x_2)\} = \text{cov}\{\zeta(x_1), \zeta(x_2)\}$, 进而有

$$n \left[E \left\{ \int_{-\infty}^{\infty} \zeta^2(x) dF(x) \right\} \right]^{-1} \widehat{\text{CCov}}(Y^* \mid X) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1),$$

再加上 $\text{var}(Y^*) = \text{var}(Y)$, 于是得到

$$n\kappa^{-1} \{\widehat{\text{CD}}(Y^* \mid X)\} \xrightarrow{d} Q,$$

其中 κ 和 Q 的定义参见定理 2.1. 这说明 $\widehat{\text{CD}}(Y^* | X)$ 会收敛到 $\widehat{\text{CD}}(Y | X)$ 的渐近零分布. $\square$

定理 3.1 的证明 根据定理 2.1 中的讨论, 有 $\widehat{\text{CCov}}(Y | X) = T_1\{1 + o_p(1)\}$. 为了建立 $\widehat{\text{CD}}(Y | X_l)$ ($l = 1, \dots, p$) 的一致收敛性, 只需证明 $T_{1,l}^*$ 的一致收敛性即可. 这里 $T_{1,l}^*$ 是 $T_{1,l}$ 的一个缩放, 其定义为

$$T_{1,l}^* \stackrel{\text{def}}{=} \frac{1}{n(n-1)(n-2)} \sum_{i \neq j, j \neq k, i \neq k} h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k),$$

其中

$$\begin{aligned} h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k) &\stackrel{\text{def}}{=} 3^{-1} [Y_i Y_k \{I(X_{il} < X_{jl}) - F(X_{jl})\} \{I(X_{kl} < X_{jl}) - F(X_{jl})\} \\ &\quad + Y_j Y_k \{I(X_{jl} < X_{il}) - F(X_{il})\} \{I(X_{kl} < X_{il}) - F(X_{il})\} \\ &\quad + Y_j Y_i \{I(X_{jl} < X_{kl}) - F(X_{kl})\} \{I(X_{il} < X_{kl}) - F(X_{kl})\}] \end{aligned}$$

是上述 U- 统计量 $T_{1,l}^*$ 的核函数.

我们将核函数 $h(\cdot)$ 分解成两部分, 即

$$h(\cdot) = h(\cdot)I\{|h(\cdot)| \leq M\} + h(\cdot)I\{|h(\cdot)| > M\},$$

其中 M 将在后文中指定. 类似地, 记 $T_{1,l}^* = T_{11,l}^* + T_{12,l}^*$, 其中

$$\begin{aligned} T_{11,l}^* &= \sum_{i \neq j, j \neq k, i \neq k} \frac{h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)I\{|h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)| \leq M\}}{n(n-1)(n-2)}, \\ T_{12,l}^* &= \sum_{i \neq j, j \neq k, i \neq k} \frac{h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)I\{|h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)| > M\}}{n(n-1)(n-2)}. \end{aligned}$$

在总体水平上, 我们也将 $S_l = E\{h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)\}$ 同样地分解为两部分 $S_{1,l} + S_{2,l}$, 其中

$$\begin{aligned} S_{1,l} &\stackrel{\text{def}}{=} E[h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)I\{|h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)| \leq M\}], \\ S_{2,l} &\stackrel{\text{def}}{=} E[h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)I\{|h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)| > M\}]. \end{aligned}$$

显然, $T_{11,l}^*$ 和 $T_{12,l}^*$ 分别是 $S_{1,l}$ 和 $S_{2,l}$ 的无偏估计.

首先研究 $T_{11,l}^*$. 根据 Markov 不等式, 对任意的 $0 < t < s_0 m^*$ (其中 $m^* = [n/3]$), 有

$$\Pr(T_{11,l}^* - S_{1,l} \geq \varepsilon) \leq \exp(-t\varepsilon) \exp(-tS_{1,l}) E\{\exp(tT_{11,l}^*)\}.$$

根据文献 [22, 定理 5.1.6], 可得 $T_{11,l}^* = (n!)^{-1} \sum_{n!} \Omega(X_{1l}, Y_1, \dots, X_{nl}, Y_n)$, 其中求和符号是对 $(1, \dots, n)$ 所有可能的排列进行求和, 且每一个 $\Omega(X_{1l}, Y_1, \dots, X_{nl}, Y_n)$ 都是 $m^* = [n/3]$ 个独立同分布的随机变量的平均. 根据 Jensen 不等式, 有

$$\begin{aligned} E\{\exp(tT_{11,l}^*)\} &\leq (n!)^{-1} \sum_{n!} E[\exp\{t\Omega(X_{1l}, Y_1, \dots, X_{nl}, Y_n)\}] \\ &= E^{m^*}(\exp[m^{*-1}th(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)I\{|h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)| \leq M\}]), \end{aligned}$$

再加上文献 [22, 引理 5.6.1.A], 于是得到

$$\Pr(T_{11,l}^* - S_{1,l} \geq \varepsilon) \leq \exp \left\{ \frac{-t\varepsilon + M^2 t^2}{2m^*} \right\}.$$

取 $t = \varepsilon m^*/M^2$, 有 $\Pr(T_{11,l}^* - S_{1,l} \geq \varepsilon) \leq \exp\{-\varepsilon^2 m^*/(2M^2)\}$. 再结合 U- 统计量的对称性, 可得 $\Pr(|T_{11,l}^* - S_{1,l}| \geq \varepsilon) \leq 2 \exp\{-\varepsilon^2 m^*/(2M^2)\}$.

下面研究 $T_{12,l}^*$. 由 Markov 不等式, 对任意 $s > 0$, 有

$$\begin{aligned} S_{2,l}^2 &\leq E\{h^2(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)\} \Pr\{|h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)| > M\} \\ &\leq \frac{E\{h^2(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)\} E[\exp\{s|h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)|\}]}{\exp(sM)}. \end{aligned}$$

假定条件 (C1) 能够保证 $E\{h^2(\cdot)\}$ 和 $E[\exp\{s|h(\cdot)|\}]$ 是有界的. 取 $M = cn^\gamma$, 其中 $0 < \gamma < 1/2 - \kappa$. 当 n 足够大时, 有 $S_{2,l} \leq \varepsilon/2$, 进而可得 $\Pr(|T_{12,l}^* - S_{2,l}| \geq \varepsilon) \leq \Pr(|T_{12,l}^*| > \varepsilon/2)$. 此外, 我们还发现

$$\left\{ |T_{12,l}^*| > \frac{\varepsilon}{2} \right\} \subseteq \{|Y_i Y_j| > M \text{ 对某些 } 1 \leq i < j \leq n \text{ 成立}\}.$$

为了说明这一点, 假定 $|Y_i Y_j| \leq M$ 对所有的 $1 \leq i < j \leq n$ 成立, 则 $|h(X_{il}, Y_i; X_{jl}, Y_j; X_{kl}, Y_k)| \leq M$ 且 $T_{12,l}^* = 0$, 这与事件 $|T_{12,l}^*| > \varepsilon/2$ 相矛盾. 根据假定条件 (C1) 和 Markov 不等式可知, 存在正的常数 C 使得

$$\Pr(|Y_i Y_j| > M) \leq \Pr(|Y_i| > M^{1/2}) + \Pr(|Y_j| > M^{1/2}) \leq 2C \exp(-sM),$$

其中 $0 < s < s_0$. 进而推出 $\Pr(|T_{12,l}^* - S_{2,l}| \geq \varepsilon) \leq 2n(n-1)C \exp(-sM)$. 由于取 $M = cn^\gamma$, 故

$$\Pr(|T_{1,l}^* - S_l| \geq 2\varepsilon) \leq 2 \exp\{-\varepsilon^2 n^{1-2\gamma}\} + 2n(n-1)C \exp(-sn^\gamma).$$

令 $\varepsilon = cn^{-\kappa}$, 其中 $c > 0$ 且 $0 < \kappa < 1/2 - \gamma$. 于是有

$$\begin{aligned} &\Pr\left\{ \max_{1 \leq l \leq p} |\widehat{\text{CD}}(Y | X_l) - \text{CD}(Y | X_l)| \geq cn^{-\kappa} \right\} \\ &\leq p \max_{1 \leq l \leq p} \Pr\{|\widehat{\text{CD}}(Y | X_l) - \text{CD}(Y | X_l)| \geq cn^{-\kappa}\} \leq O(p\alpha_n), \end{aligned}$$

其中 α_n 的定义参见定理 3.1. 至此完成了定理第一部分的证明.

下面证明一致排序性. 根据假定条件 (C2), 定义

$$\delta \stackrel{\text{def}}{=} \min_{l \in \mathcal{A}} \text{CD}(Y | X_l) - \max_{l \in \mathcal{A}^c} \text{CD}(Y | X_l) = cn^{-\kappa}.$$

因此,

$$\begin{aligned} &\Pr\left\{ \min_{l \in \mathcal{A}} \widehat{\text{CD}}(Y | X_l) > \max_{l \in \mathcal{A}^c} \widehat{\text{CD}}(Y | X_l) \right\} \\ &= 1 - \Pr\left\{ \min_{l \in \mathcal{A}} \widehat{\text{CD}}(Y | X_l) - \min_{l \in \mathcal{A}} \text{CD}(Y | X_l) + \delta \leq \max_{l \in \mathcal{A}^c} \widehat{\text{CD}}(Y | X_l) - \max_{l \in \mathcal{A}^c} \text{CD}(Y | X_l) \right\}. \end{aligned}$$

简单地计算可得

$$\begin{aligned} & \Pr\left\{\min_{l \in \mathcal{A}} \widehat{\text{CD}}(Y | X_l) - \min_{l \in \mathcal{A}} \text{CD}(Y | X_l) + \delta \leq \max_{l \in \mathcal{A}^c} \widehat{\text{CD}}(Y | X_l) - \max_{l \in \mathcal{A}^c} \text{CD}(Y | X_l)\right\} \\ & \leq 2\Pr\left\{\max_{1 \leq l \leq p} |\widehat{\text{CD}}(Y | X_l) - \text{CD}(Y | X_l)| \geq \frac{\delta}{2}\right\}. \end{aligned}$$

于是一致排序性得证.

接着证明确定筛选性. 定义事件 $E_1 \stackrel{\text{def}}{=} \{\mathcal{A} \not\subseteq \widehat{\mathcal{A}}\}$ 和 $E_2 \stackrel{\text{def}}{=} \{\text{存在某个 } l \in \mathcal{A} \text{ 满足 } |\widehat{\text{CD}}(Y | X_l) - \text{CD}(Y | X_l)| > cn^\kappa\}$. 如果 $\mathcal{A} \not\subseteq \widehat{\mathcal{A}}$, 则存在某个 $l \in \mathcal{A}$ 满足 $\widehat{\text{CD}}(Y | X_l) < cn^\kappa$. 根据假定条件 (C3) 可得 $|\widehat{\text{CD}}(Y | X_l) - \text{CD}(Y | X_l)| > cn^\kappa$. 因此, $E_1 \subseteq E_2$ 成立, 且 $\Pr(E_1) \leq \Pr(E_2)$. 于是有

$$\begin{aligned} \Pr(\mathcal{A} \subseteq \widehat{\mathcal{A}}) &= 1 - \Pr(E_1) \geq 1 - \Pr(E_2) \\ &\geq 1 - |\mathcal{A}| \max_{l \in \mathcal{A}} \Pr\{|\widehat{\text{CD}}(Y | X_l) - \text{CD}(Y | X_l)| > cn^\kappa\} \\ &\geq 1 - O(|\mathcal{A}|\alpha_n). \end{aligned}$$

证毕. □

Cumulative divergence revisited

Tingyou Zhou & Liping Zhu

Abstract We revisit cumulative divergence, a robust metric to measure the conditional mean independence between two random variables. First, we prove that if $E(Y | X)$ does not varies with X , then the corresponding sample estimator of cumulative divergence is n consistent; otherwise, it is root- n consistent. We propose a wild bootstrap procedure to approximate the asymptotic null distribution and prove that this bootstrap method is consistent. Second, we propose an independent screening procedure using cumulative divergence to filter out unimportant covariates in ultrahigh dimensional mean regressions. We establish the ranking consistency property and the sure screening property for our proposal and demonstrate its robust performance through comprehensive numerical studies.

Keywords bootstrap, mean independence, cumulative divergence, sure screening property, ultrahigh dimensional

MSC(2020) 62G10

doi: 10.1360/SCM-2019-0450