Advances in Applied Mathematics 应用数学进展, 2021, 10(11), 3777-3782
Published Online November 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1011400

Hans汉斯

基于稳定相关系的超高维筛选研究

闫　习

南京信息工程大学数学与统计学院，江苏　南京

收稿日期：2021年10月9日；录用日期：2021年10月30日；发布日期：2021年11月13日

摘　要

特征筛选是超高维数据分析中非常重要的一环，筛选降维过程的准确性将影响到后续的建模分析。针对稳定特征筛选方法(SC-SIS)的不足之处进行改进，基于稳定相关系数提出了适用于超高维无模型假设下稳健特征筛选方法(RSCS)，相比SC-SIS，该方法对数据中存在异常点或协变量服从重尾分布更有稳健性，从理论上证明了RSCS方法具有确定性筛选性质，并通过蒙特卡洛数值模拟和小鼠基因组数据验证了RSCS方法的有限样本性质。

关键词

超高维数据，稳定相关系数，确定性筛选性质，稳健性

Feature Screening for Ultra-High Dimensional Data Based on Stable Correlation Coefficient

Xi Yan

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing Jiangsu

Received: Oct. 9th, 2021; accepted: Oct. 30th, 2021; published: Nov. 11th, 2021

Abstract

Feature screening is an important part of ultra-high-dimensional data analysis. The accuracy of the screening and dimensionality reduction process will affect the subsequent modeling analysis. Aiming at the shortcomings of the stable feature screening method (SC-SIS), based on the stable correlation coefficient, a robust feature screening method (RSCS) suitable for ultra-high-dimensional model-free assumptions is proposed. This paper proves theoretically that the proposed feature

文章引用: 闫习. 基于稳定相关系的超高维筛选研究[J]. 应用数学进展, 2021, 10(11): 3777-3782.
DOI: 10.12677/aam.2021.1011400

screening method satisfies the sure screening property. Numerical simulation and a real data application under the finite sample are conducted to evaluate the performance of the proposed method.

Keywords

Ultra-High-Dimensional Data, Stable Correlation Coefficient, Sure Screening Property, Robustness

1. 引言

在这个科技发展极其迅速的时代，超高维数据已经越发频繁地出现在生物医学、经济学和社会科学等众多领域之中。对超高维数据进行分析已经成为了推动社会发展的必要手段，然而，超高维数据中预测变量维数 p 远大于样本量 n，p 会随着 n 的增长呈现出指数增长趋势，且超高维数据一般服从稀疏性假设，即只有少数预测变量与响应变量相关。若不加以任何处理，直接对原始超高维数据进行分析不仅会耗费大量的时间精力，还会导致计算成本过高、预测精度降低等问题。在稀疏性假设的驱动下，考虑先对原始超高维数据进行降维，然后再利用降维后的数据集进行统计分析，特征筛选便是超高维数据降维的重要手段。

Fan 和 Lv [1]开创性地提出了基于线性模型的超高维特征筛选方法 SIS，通过对协变量与响应变量之间的边际皮尔逊相关系数进行排序来筛选重要变量。之后又将其扩展到广义线性模型[2] [3]，然而，皮尔逊相关系数只能检测到变量之间的线性关系，在检测非线性关系时可能会受到限制。为了解决这个问题，Li 和 Peng 等[4]通过可以识别两个变量之间单调关系的 Kendall τ 相关系数来建立稳健的秩相关指标。

上述方法都是基于特定模型的，当模型假设错误时相应的方法可能不再适用。基于此，Zhu 和 Li 等[5]首次提出了针对超高维数据的无模型筛选方法；Li 和 Zhu 等[6]通过可以同时识别线性和非线性关系的距离相关系数(DC)提出了无模型的筛选方法 DC-SIS；Shao 和 Zhang [7]提出并使用鞅差相关系数构建了筛选方法；Guo 和 Li 等[8]在距离相关系数的基础上，改进了其需要矩存在这一限制，提出了稳定相关系数(SC)，并基于此建立了无模型的特征筛选方法 SC-SIS。

在本文中，我们针对稳定筛选方法 SC-SIS 在数据中存在异常点或协变量服从重尾分布时无法达到降维效果这一情况，提出一种改进的筛选方法 RSCS，该方法自然地继承了 SC-SIS 无模型这一优点并且对数据分布更有稳健性。本文的安排如下，第一部分将详细介绍 RSCS 筛选方法，并表明其具有确定性筛选性质，第二部分运用蒙特卡洛模拟方法研究 RSCS 在有限样本下的表现，第三部分则通过实例检验 RSCS 的有效性。

2. 基于稳定相关系数的特征筛选方法

令 Y 为响应变量，支撑集为 ψ_y，$x=\left(X_1,X_2,\cdots,X_p\right)^{\mathrm{T}}$ 为 p 维协变量，其中 p 远远大于样本量 n，p 随着 n 的增长呈现指数增长趋势。由于超高维数据一般服从稀疏性假设，即只有少数自变量与响应变量有关，因此为了得到与响应变量有关的那些重要变量，定义如下重要变量集合和不重要变量集合：

$$\mathcal{A}=\left\{1\le j\le p: F\left(Y\mid x\right) \text{依赖于 } X_j\right\}.$$

$$\mathcal{I}=\left\{1\le j\le p: F\left(Y\mid x\right)\text{不依赖于 } X_j\right\}.$$

其中 $F\left(y\mid x\right)=P\left(Y\le y\mid x\right), x=\left(X_1,X_2,\cdots,X_p\right)^{\mathrm{T}}$，因此，$X_{\mathcal{A}}=\left\{X_j: j\in\mathcal{A}\right\}$ 可以表示重要变量，$X_{\mathcal{I}}=\left\{X_j: j\in\mathcal{I}\right\}$ 可以表示不重要变量，而筛选的目的就是识别所有重要变量。

根据 Guo 和 Li 等[8]提出的稳定相关系数，对于两个维度分别为 d_X 和 d_Y 的随机向量 X 和 Y，两者之间的稳定相关系数的定义为：

$$SC\left(X,Y\right)=\frac{S\operatorname{cov}\left(X,Y\right)}{\sqrt{S\operatorname{cov}\left(X,X\right)S\operatorname{cov}\left(Y,Y\right)}}.$$

其中 $S\operatorname{cov}\left(X,Y\right)$ 表示 X 和 Y 稳定协方差 $S\operatorname{cov}^2\left(X,Y\right)$ 的正平方根，$S\operatorname{cov}^2\left(X,Y\right)=:E_1+E_2-2E_3$，$E_1=E\left[\mathrm{e}^{-\left\|X-\tilde{X}\right\|^a-\left\|Y-\tilde{Y}\right\|^a}\right]$，$E_2=E\left[\mathrm{e}^{-\left\|X-\tilde{X}\right\|^a}\right]E\left[\mathrm{e}^{-\left\|Y-\tilde{Y}\right\|^a}\right]$，$E_3=E\left[\mathrm{e}^{-\left\|X-\tilde{X}\right\|^a-\left\|Y-\tilde{Y}\right\|^a}\right]$，$\left(\tilde{X},\tilde{Y}\right)$ 是与 $\left(X,Y\right)$ 独立同分布的向量，$\left\|\bullet\right\|^a$ 为欧氏距离的 a 次幂。

Guo 和 Li 等[8]证明稳定相关系数 SC 不仅可以在没有任何矩条件时度量变量间线性和非线性关系而且 $SC\left(X,Y\right)=0$ 与 X 和 Y 独立等价，并通过 SC 直接构建了筛选指标 $\omega_j=SC^2\left(X_j,Y\right)$，得到了筛选方法 SC-SIS，但是该方法在数据中存在异常点或者数据中协变量服从重尾分布时会失效，筛选得到的指标集远远不能达到降维的效果，为了改善这一缺点，提高筛选方法的稳健性和适用性，在 SC-SIS 筛选指标中用 $F_j\left(x\right)=P\left(X_j\le x\right)$ 代替 X_j，因为分布代表数据的整体趋势，具有较强的稳定性。故得到以下筛选指标

$$\omega_j^*=rsc^2\left(F_j\left(X_j\right),Y\right)=\frac{\operatorname{scov}^2\left(F_j\left(X_j\right),Y\right)}{\operatorname{scov}\left(F_j\left(X_j\right),Y\right)\operatorname{scov}\left(Y,Y\right)}.$$

为得到其估计，可以用经验分布函数 $F_{j,n}\left(x\right)=n^{-1}\sum_{i=1}^{n}I\left(X_{j,i}\le x\right)$ 来估计 X_j 的分布 $F_j\left(x\right)$，由此可得 $\widehat{\operatorname{scov}}\left(F_j\left(X_j\right),Y\right)=\hat{E}_{j,1}+\hat{E}_{j,2}-2\hat{E}_{j,3}$，其中

$$\hat{E}_{j,1}=n^{-1}\left(n-1\right)^{-1}\sum_{i=1}^{n}\sum_{l\ne i}^{n}\mathrm{e}^{-\left|F_{j,n}\left(X_{j,i}\right)-F_{j,n}\left(X_{j,l}\right)\right|^a-\left|Y_i-Y_l\right|^a},$$

$$\hat{E}_{j,2}=n^{-1}\left(n-1\right)^{-1}\sum_{i=1}^{n}\sum_{l\ne i}^{n}\mathrm{e}^{-\left|F_{j,n}\left(X_{j,i}\right)-F_{j,n}\left(X_{j,l}\right)\right|^a}n^{-1}\left(n-1\right)^{-1}\sum_{i=1}^{n}\sum_{l\ne i}^{n}\mathrm{e}^{-\left|Y_i-Y_l\right|^a},$$

$$\hat{E}_{j,3}=n^{-1}\left(n-1\right)^{-1}\left(n-2\right)^{-1}\sum_{i=1}^{n}\sum_{l\ne i}^{n}\sum_{k\ne i,l}^{n}\mathrm{e}^{-\left|F_{j,n}\left(X_{j,i}\right)-F_{j,n}\left(X_{j,l}\right)\right|^a-\left|Y_i-Y_k\right|^a}.$$

因此 ω_j^* 的估计为

$$\hat{\omega}_j^*=\widehat{rsc}^2\left(F_j\left(X_j\right),Y\right)=\frac{\widehat{\operatorname{scov}}^2\left(F_j\left(X_j\right),Y\right)}{\widehat{\operatorname{scov}}\left(F_j\left(X_j\right),Y\right)\widehat{\operatorname{scov}}\left(Y,Y\right)}.$$

稳健筛选指标 RSC 可以通过集合 $\hat{A}=\left\{1\le j\le p:\hat{\omega}_j^*\ge\gamma_n\right\}$ 来识别真正重要的协变量，其中 γ_n 是需要提前给定的阈值。事实上，我们可以找到一个预先给定的 d_n 来筛选相同的集合 $\hat{A}=\left\{1\le j\le p:\hat{\omega}_j^*\right.$ 从大到小排序的前 $\left.d_n\text{ 个}\right\}$，其中 d_n 参考值为 $d_n=k\left[n\log\left(n\right)\right]$，$k$ 为正数，d_n 参考值由 Fan 和 Lv [1]提出并被广泛使用。

接下来我们探讨所提出特征筛选方法的理论性质。为方便后续的证明，给出以下条件：

(C1)对 $c>0$ 且 $0\le\kappa<1/2$，RSC 中最小的真实变量集合满足 $\min_{j\in A}\omega_j^*\ge 2cn^{-\kappa}$。

条件 C1 要求所有重要变量的指标最小值有下界，对重要变量 X_j，其 ω_j 不能太接近 0。

定理 1 设阈值 $\gamma_n=cn^{-\kappa}$，其中 $c>0$ 且 $0\le\kappa<1/2$，则存在一个正常数 C 满足：

$$\Pr\left(\max_{1\le j\le p}\left|\hat{\omega}_j^*-\omega_j^*\right|\ge\gamma_n\right)\le O\left(p\exp\left\{-Cn^{1-2\kappa}\right\}\right).$$

在条件 C1 下，可以得到

$$\Pr\left(A\subset\hat{A}\right)\ge 1-O\left(q\exp\left(-Cn^{1-2\kappa}\right)\right).$$

其中 $q=|A|$ 表示 A 的势。

定理 1 保证了 RSCS 具有确定筛选性质，可以允许数据的维数 p 随着样本量 n 以指数的方式增长，并且能够在很大的概率下选出真正重要的变量。具体来说，当 n 趋于无穷时，$\log(p)=o\left(n^{1-2\kappa}\right)$，$\Pr\left(A\subset\hat{A}\right)\to 1$，定理的细节可参照 Guo 和 Li 等[8]中的定理 3.1。

3. 模拟研究

在这一部分，我们将通过蒙特卡洛模拟来研究 RSCS 的有限样本性质并与一些现有的方法做比较，如 SIS [1]、SIRS [5]、DC-SIS [6]、SC-SIS [8]。

为检验 RSCS 方法的稳健性，假设 X 服从混合分布 $(1-\alpha)X_n+\alpha X_t$，α 分别取 0、0.1 和 0.2，其中 X_t 为每个分量独立服从 $t(1)$ 分布的 p 维随机向量，X_n 服从为均值为 0，协方差矩阵为 Σ 的多元正态分布 $N(0,\Sigma)$，$\Sigma=(\sigma_{ij})_{p\times p}$ 且 $\sigma_{ij}=0.75^{|i-j|}$，$i,j=1,\cdots,p$。同时考虑误差 ε 服从标准正态 $N(0,1)$ 和 $t(1)$ 分布两种情况。令 MMS 表示包含所有重要变量的最小模型尺寸，MMS 值较小的筛选方法表明它在识别协变量与响应变量之间相关性上更有优势，我们将通过 500 次模拟实验中 MMS 的 25%，50%，75%和 95%分位数来比较不同方法的性能。

需要指出的是，稳定相关系数 SC 中的参数 a 并未指定，至于如何选取合适的 a 值，Guo 和 Li 等[8]指出，在理论研究中，a 的取值区间为 $(0,2]$，但包含所有重要变量的最小模型尺寸 MMS 会随着 a 的增加先减少再增大，所以，a 的合理取值在(0.3, 0.7)中，并在后续研究中，选取了 $a=0.5$，在本文中我们沿用了此设定。

例 1：我们考虑非线性模型：

$$Y=5X_1X_2+5I(X_3>0)+5\sin(2\pi X_4)+5X_5+\varepsilon.$$

其中 $I(\cdot)$ 表示示性函数，$\sin(\cdot)$ 为正弦函数并选取协变量维数和样本量分别为 $n=200, p=2000$。MMS 结果如下表 1 所示。

Table 1. The MMS results with different values of α in Example 1
表 1. 例 1 中 α 不同取值时 MMS 结果

	$\alpha=0$				$\alpha=0.1$				$\alpha=0.2$			
方法	25%	50%	75%	95%	25%	50%	75%	95%	25%	50%	75%	95%
	$\varepsilon\sim N(0,1)$											
SIS	7.0	46.0	472.5	1870.5	589.8	1231.0	1678.0	1898.1	697.2	1227.0	1927.5	2145.5
SIRS	37.0	220.0	220.0	1509.4	615.8	1193.5	1655.2	1887.4	862.0	862.0	862.0	862.0

Continued

DC-SIS	6.0	7.0	5.0	10.1	32.0	90.0	349.2	349.2	183.2	441.0	1031.2	1634.4
SC-SIS	6.0	6.0	6.5	7.0	6.0	7.0	21.0	46.0	6.0	17.0	44.0	157.0
RSCS	5.0	5.0	5.0	6.0	5.0	5.0	12.0	34.0	5.0	12.0	32.0	115.0
$\varepsilon \sim t(1)$												
SIS	292.5	994.5	1812.5	1995.0	774.8	1436.5	1688.8	1881.1	716.8	1193.0	1558.2	1843.2
SIRS	53.8	286.0	883.5	1594.1	685.2	1222.5	1646.0	1895.1	877.8	1363.0	1762.2	1926.2
DCSIS	6.0	6.0	25.0	105.9	64.0	170.0	473.2	1084.8	217.2	217.2	1062.5	1595.4
SCSIS	5.0	5.5	10.0	31.0	6.0	23.0	76.0	248.1	19.0	74.0	198.0	271.0
RSCS	5.0	5.5	8.0	24.0	5.0	17.0	41.0	126.1	15.0	53.0	122.0	169.0

由表 1 可以看出，对于 $\varepsilon \sim N(0,1)$ 和 $\alpha = 0$，SIRS 和 SIS 表现不佳。这表明这两个方法无法检测协变量与响应之间的非线性关系。相比之下，RSCS、SC-SIS 和 DC-SIS 可以有效检测非线性关系。当 $\alpha = 0.1$ 或 $\alpha = 0.2$ 时，DC-SIS 和 SC-SIS 的性能并不令人满意。这意味着这两种方法对异常值或重尾分布不稳健。而我们的方法 RSCS 在这种非线性情况下的所有设置中都具有最佳性能，这表明我们的方法不仅能够检测任何可能的相关关系，而且对异常值及重尾分布也不敏感。

4. 实例分析

我们通过心肌病转基因小鼠的微阵列数据进行实例分析[9]。该数据对 30 个小鼠进行了实验，目的是挑选出最影响小鼠中 G 蛋白偶联受体基因 Ro1 表达的基因，其中受体基因 Ro1 的表达水平会受到其他 6319 个基因的影响。因此，我们把 Ro1 的基因表达水平看作响应变量 Y，其他 6319 个基因看作协变量 x，维数为 6319，数据集的样本量 $n = 30$ 远远小于协变量的维数 $p = 6319$。

通过检查该数据中大部分协变量要么是重尾分布要么包含异常点，图 1 展示了前 100 个协变量标准化后的箱线图。这可能表明我们的方法 RSCS 比其他方法更加适用。

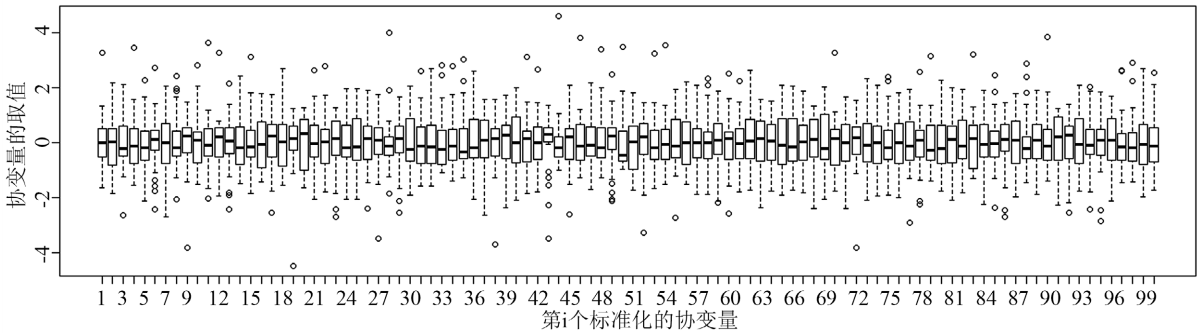

Figure 1. Boxplot of the first 100 covariates after normalization
图 1. 前 100 个协变量标准化后的箱线图

在应用筛选方法之后，可以得到各个基因的从大到小的排序，RSCS 表明 Msa.2134.0 和 Msa.1024.0 是最重要的前两个基因，与 SC-SIS 方法一致，说明 RSCS 在实际中是堪用的。

5. 总结

在本文中，我们提出了一种改进的基于稳定相关系数的筛选方法(RSCS)，通过将变量转变成分布函

数，结合稳定相关系数可以度量两个随机向量的相关性实现了这一想法，并建立了相应的确定性筛选性质，在我们的模拟研究中显示，这种方法(RSCS)对于协变量包含异常值或服从重尾分布的超高维数据非常有效。

参考文献

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

[2] 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

[3] Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh Dimensional Feature Selection: Beyond the Linear Mode. *The Journal of Machine Learning Research*, **10**, 2013-2038.

[4] Li, G., Peng, H., Zhang, J., *et al*. (2012) Robust Rank Correlation Based Screening. *The Annals of Statistics*, **40**, 1846-1877. https://doi.org/10.1214/12-AOS1024

[5] Zhu, L., Li, L., Li, R., *et al*. (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

[6] 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

[7] Shao, X. and Zhang, J. (2014) Martingale Difference Correlation and Its Use in High Dimensional Variable Screening. *Journal of the American Statistical Association*, **109**, 1302-1318. https://doi.org/10.1080/01621459.2014.887012

[8] Guo, X., Li, R., Liu, W., *et al*. (2021) Stable Correlation and Robust Feature Screening. *Science China Mathematics*, 1-16. https://doi.org/10.1007/s11425-019-1702-5

[9] Redfern, C., Coward, P., Degtyarev, M., *et al.* (1999) Conditional Expression and Signaling of a Specifically Designed GI-Coupled Receptor in Transgenic Mice. *Nature Biotechnology*, **17**, 165-169. https://doi.org/10.1038/6165