引用格式:刘漫雨,黄彬,刘佳乐. 超高维异方差数据下基于边际经验似然的分位数特征筛选[J]. 北京化工大学学报(自然科学版),2023,50(2):112 - 118.

LIU ManYu, HUANG Bin, LIU JiaLe. Quantile screening for ultrahigh-dimensional heterogeneous data by marginal empirical likelihood[J]. Journal of Beijing University of Chemical Technology (Natural Science), 2023,50(2):112 - 118.

超高维异方差数据下基于边际经验似然的分位数特征筛选

刘漫雨 黄 彬* 刘佳乐

(北京化工大学 数理学院, 北京 100029)

摘 要: 针对超高维异方差数据,基于边际经验似然提出一种分位数特征筛选方法,该方法不依赖于模型假定,且计算简单快捷,无须进行复杂的参数估计和迭代计算。同时,沿袭经验似然方法的优点,该方法对分布的假设较宽松。在一定的正则条件下,理论上证明了所提方法满足确定筛选性质。此外,为了筛选出对响应变量有影响的所有协变量,将上述方法进行推广,得到一种基于边际经验似然的分布函数特征筛选方法。最后,通过数值模拟和实例分析验证了所提出的两种方法具有良好的有限样本性质。

关键词: 超高维数据; 异方差; 边际经验似然; 分位数筛选; 确定筛选性质

中图分类号: O212 **DOI**: 10.13543/j.bhxbzr.2023.02.014

引 言

随着科技的发展,超高维数据越来越多地出现在诸如基因表达、信号处理、金融分析等领域中。在这类数据中,协变量的维数远大于样本量,且随着样本量的增加呈指数级增长,然而只有少量的协变量对响应变量有影响,呈现稀疏性的特点。现有的基于惩罚的变量选择方法都面临着计算成本、统计精度及算法稳定性等挑战[1-3],不能很好地处理超高维数据的降维问题。为了解决这一问题,近年来许多学者提出了各种超高维数据的特征筛选方法。Fan 等[4]针对线性模型提出一种基于 Pearson 相关系数的确定独立性筛选(SIS)方法,随后他们将 SIS 方法进一步推广到广义线性模型和非参数可加模型[5-6]。在无模型假定的条件下,Zhu 等[7]基于响应变量的条件分布和协变量的边际相关性提出了确定独立性排序筛选(SIRS)方法;Li 等[8]基于距离相关系数提出了距离相关性确定独立性筛选(DC-SIS)方法;Li 等[9]基于 Kendall τ 相关系数提出了稳健秩变量筛选(RRCS)方法。

为了解决超高维数据中异方差的问题,结合分位数回归的稳健性和全面性,He 等[10]通过样条函数逼近边际分位数回归的方式,提出了分位数自适应确定独立筛选(QaSIS)方法。Wu 等[11]提出了条件分位数特征筛选(Q-SIS)方法和条件分布函数特征筛选(DF-SIS)方法,该方法计算简单快捷,无须进行非参数估计,且对协变量没有有限矩条件的限制。Chang 等[12]创造性地将经验似然方法用于超高维数据,通过对零点处的边际似然比进行排序,提出了线性模型下基于边际经验似然的特征筛选(EL-SIS)方法,该方法只涉及单变量优化问题,便于计算,对分布的假设较宽松。随后,EL-SIS 方法又被进一步推广到半参数和非参数模型中[13-14]。

然而上述基于边际经验似然的筛选方法都是在一定的模型框架下,为了避免特定的模型假设,同时为了有效解决超高维数据中异方差的问题,本文将 EL-SIS 与条件分位数筛选方法相结合,提出了基于边际经验似然的分位数特征筛选(EL-QSIS) 和分布函数特征筛选 (EL-DFSIS)方法。沿袭经验似然方法的特征,所提方法具有计算简单快捷、无须参数

收稿日期: 2021-12-23
基金项目: 国家自然科学基金(12171024)
第一作者: 女,1995 年生,硕士生
*通信联系人
E-mail: huangbin@mail.buct.edu.cn

估计、对分布的假设较宽松、不依赖于模型假定等优点。通过理论证明、数值模拟和实例分析进一步验证了所提方法满足确定筛选性质且具有良好的有限样本性质。

1　变量筛选方法

1.1　基于边际经验似然的分位数特征筛选

令 Y 和 $\boldsymbol{X}=(X_1,X_2,\cdots,X_{p_n})^{\mathrm{T}}$ 分别表示响应变量和 p_n 维协变量，其中维数 p_n 随着 n 的增加呈指数级增长。不失一般性，假定 $E(X_k)=0$，$E(X_k^2)=1$，$k=1,2,\cdots,p_n$。假设 Y 与 $\boldsymbol{X}$ 之间满足稀疏性原则，即只有少部分协变量对响应变量有影响。对某给定的 $\tau\in(0,1)$，记条件 τ 分位数 $Q_\tau(Y\mid\boldsymbol{X})=\inf\{y:P(Y\leqslant y\mid\boldsymbol{X})\geqslant\tau\}$，为了筛选出其中的重要变量，定义活跃指标集为 $M_\tau=\{k:Q_\tau(Y\mid\boldsymbol{X})$ 依赖于 $X_k,k=1,2,\cdots,p_n\}$。记 $s_n=|M_\tau|$，其中 $|M_\tau|$ 表示 M_τ 中元素的个数，根据稀疏性原则，$s_n<n$。

由文献[11]，若 $Q_\tau(Y\mid X_k)=Q_\tau(Y)$，$k=1,2,\cdots,p_n$，则对任意 $t\in R$，$d_k(t)=0$，其中 $d_k(t)=E\{[\tau-I(Y<Q_\tau(Y))]I(X_k<t)\}$。因此可以用 $d_k(t)$ 衡量 X_k 和 Y 之间的边际相关性。记 $g_k=E\{d_k^2(X_k)\}$，则当 $Q_\tau(Y\mid X_k)=Q_\tau(Y)$ 时，$g_k=0$，反之 $g_k>0$。由此可见，g_k 越大，则越说明 X_k 是影响 Y 的条件 τ 分位数的重要变量。因此，可以通过度量边际效用 g_k 是否等于 0 来进行特征筛选。

设有独立同分布的样本 $\{(\boldsymbol{X}_i,Y_i)\}_{i=1}^n$，且样本协变量已经标准化，即

$$\frac{1}{n}\sum_{i=1}^n X_{ik}=0,\frac{1}{n}\sum_{i=1}^n X_{ik}^2=1,k=1,2,\cdots,p_n$$

令 $g_{jk}=d_k^2(X_{jk})$，$j=1,2,\cdots,n$，定义如下边际经验似然。

$$EL_k=\sup\left\{\prod_{j=1}^n\omega_j:\omega_j\geqslant 0,\ \sum_{j=1}^n\omega_j=1,\ \sum_{j=1}^n\omega_j g_{jk}=0\right\}\tag{1}$$

利用拉格朗日乘子法求解式(1)，得到边际经验似然比为

$$l_k=-2\ln\{EL_k\}-2n\ln n=2\sum_{j=1}^n\ln(1+\lambda g_{jk})$$

其中，拉格朗日乘子 λ 满足 $\sum_{j=1}^n\frac{g_{jk}}{1+\lambda g_{jk}}=0$。根据 Chang 等[12]的理论分析，当 $k\notin M_\tau$ 时，l_k 值不会太大，而当 $k\in M_\tau$ 时，l_k 将以很大的概率发散，故可将 l_k 作为度量指标进行排序从而筛选出重要变量。

为了给出 l_k 的估计，定义 $g_{jk}(j=1,2,\cdots,n)$ 的经验估计为

$$\hat{g}_{jk}=\left\{\frac{1}{n}\sum_{i=1}^n[\tau-I(Y_i<\hat{Q}_\tau(Y))]I(X_{ik}<X_{jk})\right\}^2$$

式中，$\hat{Q}_\tau(Y)$ 是基于 $Y_1,Y_2,\cdots,Y_n$ 的样本 τ 分位数。相应地，可估计 l_k 为

$$\hat{l}_k=2\sum_{j=1}^n\ln(1+\hat{\lambda}\hat{g}_{jk})$$

式中，$\hat{\lambda}$ 满足 $\sum_{j=1}^n\frac{\hat{g}_{jk}^\tau}{1+\hat{\lambda}\hat{g}_{jk}^\tau}=0$。从而 M_τ 可估计为

$$\hat{M}_\tau=\{k:\hat{l}_k\geqslant\gamma_n,k=1,2,\cdots,p_n\}$$

式中，γ_n 是预先设定的阈值。在实际应用中，常将 $\hat{l}_k,k=1,2,\cdots,p_n$ 按降序排列，此时可得 M_τ 的估计为

$$\hat{M}_\tau=\{k:\hat{l}_k\text{ 为前 }d_n\text{ 个最大的},k=1,2,\cdots,p_n\}$$

其中指定的模型大小 d_n 可仿照文献[4]取为 $[n/\ln n]$，这里 $[a]$ 表示取不大于 a 的整数。

1.2　理论性质

为了证明 EL-QSIS 方法满足确定筛选性质，假设下列正则化条件成立。

1)存在常数 $c>0$，使得

$\min_{k\in M_\tau}g_k\geqslant cn^{-\kappa}$，对某 $\kappa\in[0,1/2)$。

2)在 $Q_\tau(Y)$ 附近，$F(y)$ 二阶可微，对于 $f(y)$，存在正数 c_1、c_2，使得 $0<c_1<f(y)<c_2<\infty$ 一致成立，且 $f'(y)$ 一致有界，其中 $F(y)$ 和 $f(y)$ 分别为 Y 的分布函数和密度函数。

3)在给定 $\boldsymbol{X}_{M_\tau}$ 下，$\boldsymbol{X}_{M_\tau^c}$ 与 $I(Y<Q_\tau(Y))$ 条件独立，且 $\boldsymbol{X}_{M_\tau}$ 与 $\boldsymbol{X}_{M_\tau^c}$ 相互独立，其中 $\boldsymbol{X}_{M_\tau}=\{X_j:j\in M_\tau\}$，$\boldsymbol{X}_{M_\tau^c}=\{X_j:j\notin M_\tau\}$。

文献[11]有相同的条件 1)~3)。条件 1)要求重要变量对应的 g_k 中的最小值不能太小，这也意味着重要的协变量的信号不能太弱，这个条件被广泛应用于超高维数据的特征筛选中。条件 2)是分位数回归的常见条件。利用条件 3)可以把重要变量和非重要变量区分开，从而保证筛选排序的一致性。注意在文献[12]的条件 A.2 中要求协变量与响应变量的尾部满足指数衰减速率，而本文对协变量没有任何限制条件，因此，本文所提出的筛选方法对重尾分布更加稳健。

引理 1　在条件 1)、2)下，有

$$\max_j |\hat{g}_{jk} - g_{jk}| = O_p(n^{-\kappa})$$

由 $|\hat{Q}_\tau(Y) - Q_\tau(Y)| = O(n^{-1/2}(\ln n)^{1/2})$ *a. s.* 和文献[11]中定理 1 的证明步骤,容易得到引理 1 的结论。

定理 1 在条件 1)~3)下,存在常数 $C_1 > 0$,对任意 $\alpha \in (0, 1/2 - \kappa)$,有

$$\max_{k \in M_\tau} P\{\hat{l}_k < c^2 n^{2\alpha}\} \leqslant \exp\{-C_1 n^{1-2\kappa}\}$$

证明:对 $\forall k \in M_\tau$,由文献[15]得

$$\hat{l}_k = 2 \max_{\lambda \in \Lambda_n} \sum_{j=1}^{n} \ln\{1 + \lambda \hat{g}_{jk}\}$$

其中 $\Lambda_n = \{\lambda$:对任 $j = 1, 2, \cdots, n, 1 + \lambda \hat{g}_{jk} \geqslant n^{-1}\}$。注意 $0 \leqslant g_{jk} \leqslant 1$,且仿照文献[12]中定理 1、命题 2 的证明步骤,对某 $\varepsilon > 0$,取 $\hat{\lambda} = (n^{\varepsilon} \max_j \hat{g}_{jk})^{-1}$,则对某 $t > 0$,得

$$P\{\hat{l}_k < 2t\} \leqslant P\left\{\sum_{j=1}^{n} \frac{\hat{g}_{jk}}{n^{\varepsilon} \max_j \hat{g}_{jk}} < t + n^{1-2\varepsilon}\right\} =$$

$$P\left\{\sum_{j=1}^{n} \hat{g}_{jk} < (tn^{\varepsilon} + n^{1-\varepsilon}) \max_j \hat{g}_{jk}\right\} \leqslant P\left\{\frac{1}{\sqrt{n}\sigma_k} \sum_{j=1}^{n} (g_{jk} - g_k) < \frac{1}{\sigma_k}[(tn^{\varepsilon-1/2} + n^{1/2-\varepsilon}) \max_j \hat{g}_{jk} - \sqrt{n} g_k + \sqrt{n} \max_j |\hat{g}_{jk} - g_{jk}|]\right\} \leqslant P\left\{\frac{1}{\sqrt{n}\sigma_k} \sum_{j=1}^{n} (g_{jk} - g_k) < \frac{1}{\sigma_k}[(tn^{\varepsilon-1/2} + n^{1/2-\varepsilon}) \max_j g_{jk} - \sqrt{n} g_k + (tn^{\varepsilon-1/2} + n^{1/2} + n^{1/2-\varepsilon}) \cdot \max_j |\hat{g}_{jk} - g_{jk}|]\right\} \leqslant P\left\{\frac{1}{\sqrt{n}\sigma_k} \sum_{j=1}^{n} (g_{jk} - g_k) < \frac{1}{\sigma_k} \cdot [(tn^{\varepsilon-1/2} + n^{1/2-\varepsilon}) - \sqrt{n} g_k + (tn^{\varepsilon-1/2} + n^{1/2} + n^{1/2-\varepsilon}) \cdot \max_j |\hat{g}_{jk} - g_{jk}|]\right\}$$

式中,$g_k = E(g_{jk})$,$\sigma_k^2 = \mathrm{Var}(g_{jk})$。对于 $L \to \infty$,取 ε 使得 $n^{\varepsilon} = L/g_k$,且令 $2t = ng_k^2/L^2$,则有

$$\frac{tn^{\varepsilon}}{ng_k} = \frac{1}{2L}$$

$$\frac{n^{1-\varepsilon}}{ng_k} = \frac{1}{L}$$

$$\frac{(tn^{\varepsilon-1/2} + n^{1/2-\varepsilon}) - \sqrt{n} g_k}{\sigma_k} = \frac{\left(\frac{3}{2L} - 1\right)\sqrt{n} g_k}{\sigma_k}$$

由引理 1,有

$$\frac{(tn^{\varepsilon-1/2} + n^{1/2} + n^{1/2-\varepsilon}) \max_j |\hat{g}_{jk} - g_{jk}|}{\sigma_k} = O_p(n^{1/2} \cdot \max_j |\hat{g}_{jk} - g_{jk}|) = O_p(n^{1/2-\kappa})$$

这一项可以被忽略或被 $O_p(n^{1/2} g_k)$ 代替,这是因为在条件 1)下,对 $\forall k \in M_\tau$,$n^{1/2} g_k \geqslant cn^{1/2-\kappa}$,从而

$$P\left\{\hat{l}_k < \frac{c^2 n^{1-2\kappa}}{L^2}\right\} \leqslant P\left\{\hat{l}_k < \frac{ng_k^2}{L^2}\right\} \leqslant P\left\{\frac{1}{\sqrt{n}\sigma_k} \sum_{j=1}^{n} (g_{jk} - g_k) < \frac{(3/(2L) - 1)\sqrt{n} g_k}{\sigma_k}\right\}$$

进一步由文献[12]的引理 1、命题 2 可知,存在常数 $C_1 > 0$,使得

$$P\left\{\hat{l}_k < \frac{c^2 n^{1-2\kappa}}{L^2}\right\} \leqslant \exp\{-C_1 n^{1-2\kappa}\}$$

最后,对某 $\alpha \in (0, 1/2 - \kappa)$,取 $L = n^{1/2-\kappa-\alpha}$,则定理 1 成立。

定理 2(确定筛选性质) 在条件 1)~3)下,存在常数 $C_1 > 0$,对任 $\alpha \in (0, 1/2 - \kappa)$ 和 $\gamma_n = c^2 n^{2\alpha}$,有

$$P\{M_\tau \subset \hat{M}_\tau\} \geqslant 1 - s_n \exp\{-C_1 n^{1-2\kappa}\}$$

证明:由定理 1 及 $P\{M_\tau \not\subset \hat{M}_\tau\} = P\{$存在 $k \in M_\tau$,使得 $\hat{l}_k < c^2 n^{2\alpha}\} \leqslant s_n \max_{k \in M_\tau} P\{\hat{l}_k < c^2 n^{2\alpha}\}$,定理 2 显然成立。

从定理 2 可知,协变量维数 p_n 随样本量 n 的增加呈指数级增长,且满足

$$\ln p_n = O(n^{1-2\kappa})$$

则当 $n \to \infty$ 时,有 $P\{M_\tau \subset \hat{M}_\tau\} \to 1$,说明估计的重要变量指标集 $\hat{M}_\tau$ 以概率 1 包含真实的重要变量指标集 M_τ,即所提出的 EL-QSIS 方法满足确定筛选性质。

1.3 基于边际经验似然的分布函数特征筛选

若关注的活跃指标集为

$$M = \{k: F(y|\boldsymbol{X}) 依赖于 X_k, k = 1, 2, \cdots, p_n\}$$

其中 $F(y|\boldsymbol{X}) = P(Y \leqslant y|\boldsymbol{X})$,则可将所提出的 EL-QSIS 方法进行推广,得到一种基于边际经验似然的分布函数特征筛选(EL-DFSIS)方法,且该方法不依赖于模型假定。这里令

$$h_k(y, t) = E\{[F(y) - I(Y \leqslant y)] I(X_k < t)\}$$

则可通过度量 $E\{h_k^2(Y, X_k)\}$ 是否等于 0 来进行特征筛选。类似地,令 $\tilde{g}_{jk} = \left\{\frac{1}{n} \sum_{i=1}^{n} [\hat{F}(Y_j) - I(Y_i \leqslant Y_j)] I(X_{ik} < X_{jk})\right\}^2$,$j = 1, 2, \cdots, n$,$\hat{F}(y) = \frac{1}{n} \cdot \sum_{l=1}^{n} I(Y_l \leqslant y)$,边际经验似然比 $\check{l}_k = 2 \sum_{j=1}^{n} \ln\{1 + \tilde{\lambda} \tilde{g}_{jk}\}$,其中 $\tilde{\lambda}$ 满足 $\sum_{j=1}^{n} \frac{\tilde{g}_{jk}}{1 + \tilde{\lambda} \tilde{g}_{jk}} = 0$。基于 $\check{l}_k$,M 可估计为

$$\hat{M} = \{k: \check{l}_k \geqslant \tilde{\gamma}_n, k = 1, 2, \cdots, p_n\}$$

其中 $\tilde{\gamma}_n$ 是预先设定的阈值。

类似地，在一定的正则条件下，仿照文献[12]和上述定理 1、2 的证明步骤，可以证明 EL－DFSIS 方法也满足确定筛选性质。

2　数值模拟与实例分析

本节通过数值模拟和实例分析来验证所提出的 EL－QSIS、EL－DFSIS 筛选方法的有限样本性质，并且分别将它们与 QaSIS[10]、Q－SIS[11] 和 SIRS[7]、DF－SIS[11]、EL－SIS[12] 等方法进行比较。

在数值模拟中考虑样本量 n 为 150 或 300，协变量维数 $p_n = 3\,000$，筛选出的变量个数 $d_n = n/\ln n$，对每种情形重复 300 次试验。评价指标包括：p_0—真实的模型大小；P_{ALL}—在给定模型尺寸 d_n 下，300 次重复试验中所有重要预测变量被选中的比例；*Median*—300 次重复试验中包含所有重要预测变量的最小模型尺寸的中位数；*IQR*—300 次重复试验中包含所有重要预测变量的最小模型尺寸的四分位差。

例 1　考虑异方差线性模型

$$Y = X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5 + \sigma(\boldsymbol{X})\varepsilon$$

式中，$\boldsymbol{X} = (X_1, X_2, \cdots, X_{p_n})^{\mathrm{T}} \sim N_{p_n}(\boldsymbol{0}, \boldsymbol{\Sigma})$，$\boldsymbol{\Sigma} = (0.8^{|i-j|})(i, j = 1, 2, \cdots, p_n)$，$\sigma(\boldsymbol{X}) = X_{20} + X_{21} + X_{22}$，且误差 $\varepsilon \sim N(0,1)$ 或 $t(4)$。考虑分位点 $\tau = 0.5$ 或 $\tau = 0.75$，此时真实的重要预测变量的个数分别为 5 和 8。模拟结果见表 1。

表 1　例 1 的模拟结果

Table 1　Simulation results for Example 1

误差	分位数	方法	p_0	$n = 150$			$n = 300$		
				P_{ALL}	*Median*	*IQR*	P_{ALL}	*Median*	*IQR*
$N(0,1)$		SIRS	8	0.470	32	42	0.957	26	44
		DF－SIS	8	0.983	9	20	1	9	24
		EL－SIS	8	0.960	10	4	1	10	8
		EL－DFSIS	8	0.967	11	6	1	11	7
	0.5	QaSIS	5	1	5	1	1	5	0
	0.5	Q－SIS	5	1	5	0	1	5	0
	0.5	EL－QSIS	5	1	5	0	1	5	0
	0.75	QaSIS	8	0.88	11	13	1	10	6
	0.75	Q－SIS	8	0.887	10	10	1	9	7
	0.75	EL－QSIS	8	0.903	12	2	1	9	2
$t(4)$		SIRS	8	0.613	26	38	1	27	41
		DF－SIS	8	1	11	17	1	9	26
		EL－SIS	8	0.980	10	7	1	10	9
		EL－DFSIS	8	0.980	12	4	1	11	5
	0.5	QaSIS	5	1	5	2	1	5	0
	0.5	Q－SIS	5	1	5	1	1	5	0
	0.5	EL－QSIS	5	1	5	1	1	5	0
	0.75	QaSIS	8	0.860	12	16	1	10	13
	0.75	Q－SIS	8	0.883	13	10	1	9	10
	0.75	EL－QSIS	8	0.893	10	3	1	9	3

例 2　考虑异方差非线性模型

$$Y = X_1^2 \sin X_2 + X_3^3 + (\cos X_4)^3 + X_5 + \sigma(\boldsymbol{X})\varepsilon$$

其他设置条件与异方差线性模型相同，模拟结果见表 2。特别地，在给定模型尺寸 d_n 下，表 3 给出了协变量 X_{20}、X_{21}、X_{22} 在 300 次重复试验中被选中的比例（除去 $\tau = 0.5$ 的情形）P_{20}、P_{21}、P_{22}。

表 2 例 2 的模拟结果

Table 2 Simulation results for Example 2

误差	分位数	方法	p_0	$n=150$			$n=300$		
				P_{ALL}	*Median*	*IQR*	P_{ALL}	*Median*	*IQR*
$N(0,1)$		SIRS	8	0.373	38	44	0.957	27	40
		DF-SIS	8	0.853	12	63	0.987	9	41
		EL-SIS	8	0.867	24	26	0.997	19	20
		EL-DFSIS	8	0.897	21	19	1	17	15
	0.5	QaSIS	5	1	6	2	1	5	1
	0.5	Q-SIS	5	1	5	0	1	5	0
	0.5	EL-QSIS	5	1	5	1	1	5	0
	0.75	QaSIS	8	0.410	23	24	0.913	21	20
	0.75	Q-SIS	8	0.763	16	17	0.990	12	13
	0.75	EL-QSIS	8	0.823	19	3	1	17	2
$t(4)$		SIRS	8	0.313	39	52	0.980	24	51
		DF-SIS	8	0.613	16	72	1	12	39
		EL-SIS	8	0.703	25	24	1	16	18
		EL-DFSIS	8	0.737	24	17	1	20	14
	0.5	QaSIS	5	0.957	7	2	1	5	1
	0.5	Q-SIS	5	1	5	1	1	5	0
	0.5	EL-QSIS	5	1	5	1	1	5	0
	0.75	QaSIS	8	0.308	38	44	0.813	27	32
	0.75	Q-SIS	8	0.560	22	29	0.870	12	17
	0.75	EL-QSIS	8	0.823	23	5	0.993	15	4

表 3 例 2 中 X_{20}, X_{21}, X_{22} 被选中的比例

Table 3 Selection proportions of X_{20}, X_{21}, X_{22} for Example 2

误差	分位数	方法	$n=150$			$n=300$		
			P_{20}	P_{21}	P_{22}	P_{20}	P_{21}	P_{22}
$N(0,1)$		SIRS	0.540	0.627	0.513	0.977	0.990	0.977
		DF-SIS	0.993	0.960	0.903	0.997	0.997	0.990
		EL-SIS	0.963	0.933	0.977	0.997	1	1
		EL-DFSIS	0.910	0.900	0.977	1	1	1
	0.75	QaSIS	0.840	0.540	0.413	1	0.927	0.957
	0.75	Q-SIS	0.797	0.873	0.793	0.997	0.993	0.990
	0.75	EL-QSIS	0.840	0.877	0.793	1	1	1
$t(4)$		SIRS	0.603	0.727	0.497	0.993	1	0.983
		DF-SIS	0.787	0.830	0.787	1	1	1
		EL-SIS	0.777	0.837	0.887	1	1	1
		EL-DFSIS	1	0.903	0.997	1	1	1
	0.75	QaSIS	0.750	0.630	0.560	0.913	0.907	0.883
	0.75	Q-SIS	0.777	0.767	0.663	0.927	0.947	0.930
	0.75	EL-QSIS	0.907	0.930	0.990	1	0.997	0.993

从表 1 和表 2 的结果来看，EL－DFSIS 与 DF－SIS 方法表现相当，并且都要优于 SIRS 方法，表现为具有更小的最小模型尺寸和更高的重要变量覆盖率。与 DF－SIS 方法相比，EL－DFSIS 方法具有更小的最小模型尺寸 *IQR*；当模型为线性模型时，EL－DFSIS 方法与 EL－SIS 方法表现相当，但当模型为非线性模型时，EL－DFSIS 方法能以更高的比例筛选出所有重要预测变量，并且具有更小的最小模型尺寸。随着样本量 n 的增加，所有方法筛选出的包含所有重要预测变量的比例均显著增加。

当 $\tau=0.5$ 时，EL－QSIS、QaSIS 和 Q－SIS 方法均有良好的筛选表现，最小模型尺寸基本都为 5，且重要预测变量被选出的比例都接近于 1。但当 $\tau=0.75$ 时，由于存在异方差的影响，QaSIS 方法在所有结果中均表现最差，EL－QSIS 和 Q－SIS 方法筛选表现良好，且 EL－QSIS 方法比 Q－SIS 方法有更高的重要变量覆盖率和更小的最小模型尺寸 *IQR*，说明 EL－QSIS 方法对真实模型中重要预测变量的筛选更加准确有效。

表 3 展示了例 2 中对误差产生影响的重要变量被选中的比例。可以发现，与其他方法相比，EL－QSIS 和 EL－DFSIS 方法都能以较高的比例筛选出这 3 个重要变量。

例 3（实例分析）　将 EL－DFSIS 方法应用到实际数据分析中。考虑文献[16]中的白血病数据集，该数据集包含 47 例急性淋巴细胞白血病（ALL）患者和 25 例急性髓系白血病（AML）患者的 7 129 个基因表达水平数据。本文感兴趣的是识别可能影响白血病种类 ALL 和 AML 的基因。Golub 等[16]、Barut 等[17]发现 Zyxin、hSNF2b 和 TCRD（对应的基因为 X95735_at、D26156_s_at、U29175_at 和 M21624_at）为识别两类白血病的重要变量。在给定模型尺寸 $d=25$ 下，表 4 展示了 EL－DFSIS 与 SIRS、DF－SIS、EL－SIS 方法对这 4 个重要基因的筛选结果。

表 4　例 3 的筛选结果

Table 4　Screening results for Example 3

方法	M21624_at	X95735_at	D26156_s_at	U29175_at
SIRS		√	√	√
DF－SIS		√	√	√
EL－SIS	√	√	√	√
EL－DFSIS	√	√	√	√

√表示筛选出了重要基因，空白表示未筛选出重要基因。

从表中可见这 4 种方法均能筛选出 X95735_at、D26156_s_at 和 U29175_at 这 3 个重要基因，而 EL－DFSIS 和 EL－SIS 还筛选出了重要基因 M21624_at，说明基于经验似然的 EL－DFSIS 筛选方法在实际应用中也具有良好的变量筛选能力。

综上所述，与其他几种方法相比，本文提出的 EL－QSIS、EL－DFSIS 筛选方法具有良好的筛选降维效果，表现在可以更精确地筛选出重要变量，且能更稳定地用较小的模型尺寸识别出所有的重要变量。

3　结束语

针对超高维异方差数据，本文提出了基于边际经验似然的分位数特征筛选（EL－QSIS）和分布函数特征筛选（EL－DFSIS）方法，两种方法均不依赖于模型假定，无须进行复杂的参数估计和迭代计算，且对分布的假设较宽松。分析了所提方法的确定筛选性质，数值模拟与实例分析的结果表明与 QaSIS、Q－SIS 和 SIRS、DF－SIS、EL－SIS 等方法相比，EL－QSIS 和 EL－DFSIS 方法可以更精确地筛选出重要变量，且能更稳定地用较小的模型尺寸识别出所有的重要变量，具有良好的筛选降维效果。未来可考虑将该方法进一步推广到纵向数据、缺失数据的情况。

参考文献：

[1] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society, 1996, 58(1): 267－288.

[2] FAN J Q, LI R Z. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348－1360.

[3] ZOU H. The adaptive lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1418－1429.

[4] FAN J Q, LV J C. Sure independence screening for ultra-high dimensional feature space[J]. Journal of the Royal Statistical Society, 2008, 70(5): 849－911.

[5] FAN J Q, SONG R. Sure independence screening in generalized linear models with NP-dimensionality[J]. The Annals of Statistics, 2010, 38(6): 3567－3604.

[6] FAN J Q, FENG Y, SONG R. Nonparametric independence screening in sparse ultra-high-dimensional additive models[J]. Journal of the American Statistical Associa-

tion, 2011, 106(494): 544 – 557.
[7] ZHU L P, LI L X, LI R Z, et al. Model-free feature screening for ultrahigh-dimensional data[J]. Journal of the American Statistical Association, 2011, 106(496): 1464 – 1475.
[8] LI R Z, ZHONG W, ZHU L P. Feature screening via distance correlation learning[J]. Journal of the American Statistical Association, 2012, 107(499): 1129 – 1139.
[9] LI G R, PENG H, ZHANG J, et al. Robust rank correlation based screening[J]. The Annals of Statistics, 2012, 40(3): 1846 – 1877.
[10] HE X M, WANG L, HONG H G. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data[J]. The Annals of Statistics, 2013, 41(1): 342 – 369.
[11] WU Y S, YIN G S. Conditional quantile screening in ultrahigh-dimensional heterogeneous data[J]. Biometrika, 2015, 102(1): 65 – 76.
[12] CHANG J Y, TANG C Y, WU Y C. Marginal empirical likelihood and sure independence feature screening[J]. The Annals of Statistics, 2013, 41(4): 2123 – 2148.
[13] CHANG J Y, TANG C Y, WU Y C. Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood[J]. The Annals of Statistics, 2016, 44(2): 515 – 539.
[14] CHU Y, LIN L. Conditional SIRS for nonparametric and semiparametric models by marginal empirical likelihood[J]. Statistical Papers, 2020, 61: 1589 – 1606.
[15] OWEN A B. Empirical likelihood[M]. New York: Chapman & Hall/CRC, 2001.
[16] GOLUB T R, SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439): 531 – 537.
[17] BARUT E, FAN J Q, VERHASSELT A. Conditional sure independence screening[J]. Journal of the American Statistical Association, 2016, 111(515): 1266 – 1277.

Quantile screening for ultrahigh-dimensional heterogeneous data by marginal empirical likelihood

LIU ManYu HUANG Bin* LIU JiaLe
(College of Mathematics and Physics, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: We propose a quantile screening method based on marginal empirical likelihood for ultrahigh-dimensional heterogeneous data. The proposed model-free method is computationally simple because it can select active predictors without parameter estimation or an iterative algorithm, and inheriting the advantages of the empirical likelihood approach results in fewer restrictive distributional assumptions. The results reveal that the proposed procedure enjoys sure screening properties under certain technical conditions. Moreover, a distribution function screening method based on marginal empirical likelihood is suggested as a way to recover the whole active predictor set. Simulation results and real data analysis confirm that the proposed screening methods perform well when used with finite samples.

Key words: ultrahigh-dimensional data; heterogeneity; marginal empirical likelihood; quantile screening; sure screening property

(责任编辑:吴万玲)