第 41 卷 第 1 期 应 用 数 学 学 报 Vol. 41 No. 1
2018 年 1 月 ACTA MATHEMATICAE APPLICATAE SINICA Jan., 2018

SEVIS 方法的局部线性估计及其在超高维数据下的应用

连亦 †

(中国科学技术大学统计与金融系，合肥 230026)

(E-mail: ymlian@mail.ustc.edu.cn)

陈 钊

(Department of Statistics, Pennsylvania State University, State College, USA, PA 16802)

(E-mail: chenzhao1985@gmail.com)

舒明良

(中国科学院数学与系统科学研究院，北京 100190)

(E-mail: shumingliang12@mails.ucas.ac.cn)

摘 要 在大数据时代的背景下，如何从超高维数据中筛选出真正重要的特征成为许多相关行业的研究者们广泛关注的一个问题. 特征筛选的核心思想就在于排除那些明显与因变量不相关的特征以达到这一目的. 基于核估计的 SEVIS (Sure Explained Variability and Independence Screening) 特征筛选方法在处理非对称，非线性数据下要在一定程度上优于之前的特征筛选模型，但其采用核估计的方式对非参数部分进行估计的方法仍存在进一步改进的空间. 本文就从这个角度出发，将其核估计的算法修改为局部线性估计，并考虑部分特殊情况下的变量选择过程. 结果显示，基于局部线性估计的 SEVIS 方法在准确性，运行效率上都要优于基于核估计的 SEVIS 的方法.

关键词 特征筛选；局部线性估计； SEVIS; 超高维数据

MR(2000) 主题分类 62G05

中图分类 0212.7

本文 2016 年 11 月 18 日收到. 2017 年 5 月 15 日收到修改稿.

† 通讯作者.

1 引言

随着科技水平的迅速发展以及数据收集能力的大幅提高，高维数据分析已经越来越频繁地出现在许多学科的研究领域中，且扮演着越来越重要的角色. 随之应运而生的方法包括 [1–6] 等等. 它们都已经过大量的实践检验，能够在维数较高 ($p > n$) 的情况下成功选择出重要的变量. 但一旦情况变为超高维，例如特征个数 p 随着样本数 n 呈指数阶的速度增长，即 $p = O(\exp(n^{\eta}))$ 时，这些方法普遍在计算复杂性，统计准确性，算法稳定性方面存在不足 [7].

为了更好地解决这一问题，[8] 提出了开创性的两阶段变量选择方法 SIS (sure independent screening), 基于超高维数据中常见的稀疏性假定 (即大部分特征与因变量不存在相关性), 将变量筛选和模型拟合分成两个步骤, 先对每个特征 X 分别与 Y 的 Pearson 相关系数进行排序, 并取其中排名靠前的一部分构成一个活跃变量集, 旨在排除那些明显与 Y 不存在线性相关的特征，再对活跃变量集中的已经经过降维的特征对 Y 进行拟合，由于后者相当于处理低维度时候的模型拟合问题，其方法已经较为成熟，所以文章的重心主要放在第一步特征筛选的过程中. 与此同时，为了确保所有活跃变量均能够在第一步中被选入活跃变量集中， Fan 和 Lv[8] 还在该论文中提出一个判别准则，即 sure screening 性质，在假定线性模型的基础上，即随着样本总数 n 趋于无穷，所有活跃变量被选入该活跃变量集的概率趋于 1. 在此奠基性工作的基础上， [9] 将其由一般的线性模型推广到广义线性模型上. [10] 进一步使用非参数 B 样条的方法考虑了可加模型的情况. [11] 同样使用 B 样条方法，但他们研究的是变系数模型的情况，并用非参数边际相关系数的大小来判断特征的重要程度. 同样考虑变系数模型的还有 [12], 他们使用核估计的方法来估计条件相关系数并将其作为排序的依据. 与此同时，考虑到上述方法均事先假定一个特定的模型，再基于这个模型出发建立一套针对该模型的方法，尽管在该特定模型下的表现优异，但如果模型假定出现误差或与现实情况不匹配的状况发生，这些 model-based 的方法所得到的结论就难以得到保证. 因此，为了避免该误差， [13] 提出的 SIRS 方法仅假定了一般模型框架而非特定的模型结构. [14] 提出基于距离相关系数的特征筛选方法的 DC-SIS, 该方法通过联合特征函数与边际特征函数来衡量与排序特征与因变量之间的相依程度，且在不需要假定特征与因变量之间任何相依结构的情况下仍然具有 sure screening 性质，因此该方法也可以称为完全的 model-free 方法. 另一部分统计学家从分位回归的角度出发，例如 [15] 介绍了基于分位回归的非线性特征筛选方法以处理存在异方差的数据， [16] 则提出了基于条件分位回归的方法 Q-SIS, 同样地，他们也不需要假定特定的模型结构.

与大多数方法从中心性或分位数出发的角度不同， SEVIS[17] (Sure Explained Variability and Independence Screening) 从方差解释度的角度出发，考察当特征与因变量之间的关系呈现非对称性，非线性情况下的特征筛选问题，该方法在许多模型下要明显优于之前的特征筛选方法，但其基于核估计的估计方式仍然存在进一步改善的空间. 值得一提的是，现有的非参数特征筛选方法大多数基于核估计或 B 样条估计，并没有直接

采用局部线性估计的方法. 尽管局部线性估计的结果要较为类似于核估计的结果，但在一些特殊情况下，例如 $\frac{f'(x)}{f(x)}$ 的绝对值较大时，核估计存在明显较大的均方误差 [18]. 为了降低这一部分估计偏差对结果的影响并提高算法速度，本文采用局部线性估计的方式对 SEVIS 方法中的相关性度量方式 SEV 进行估计，并将其重新运用到特征筛选的过程中，得出在部分情况下采用局部线性估计的 SEVIS 方法结果要好于采用核估计方法估计的 SEVIS 方法的结论，且两者皆优于其它对照方法.

本文内容的安排如下：第 2 部分给出了 SEVIS 方法的基本概念，其局部线性估计方法和渐进性质，及对应特征筛选的理论与过程. 第 3 部分进行了两个蒙特卡洛模拟以比较各方法之间的准确性和算法速度. 第 4 部分对文章进行了总结与展望. 附录给出了局部线性估计下的渐进正态性的证明.

2　SEV 的局部线性估计

2.1　局部线性估计量

SEV (Sure Explained Variability) 是一个行之有效的针对随机变量的非线性度量方式，其核心思想来源于成对出现的广义相关系数 GMCs[19] 中的一个分支. 我们假设 (X,Y) 是一对随机变量，根据著名的方差分解公式， X 对 Y 的方差解释度可以定义为：

$$\mathrm{SEV}(Y|X)=\frac{\mathrm{var}(E(Y|X))}{\mathrm{var}(Y)}=1-\frac{E(\mathrm{var}(Y|X))}{\mathrm{var}(Y)}=1-\frac{E[\{Y-E(Y|X)\}^2]}{\mathrm{var}(Y)}. \tag{2.1}$$

也可以表示为

$$\mathrm{SEV}\,(Y|X)=\frac{\mathrm{var}(E(Y|X))}{\mathrm{var}(Y)}=\frac{E(E(Y|X=x))^2-(E(Y))^2}{\mathrm{var}(Y)}. \tag{2.2}$$

与核估计方法不同，这里我们给出一个新的局部光滑估计方法以提高该非参数估计量的准确性和有效性. 令 $(Y_i,X_i),\ i=1,\cdots,n$ 为独立同分布的随机样本. $E(Y|X=x)$ 为非参数回归

$$Y_i=f(X_i)+\varepsilon_i,\qquad i=1,\cdots,n \tag{2.3}$$

的均值函数, 其中 $f(\cdot)$ 是一个未知函数，ε_i 为一组均值为 0, 方差为 σ^2 的随机误差项. 由泰勒展开，我们有

$$f(x)\approx f(x_0)+f'(x_0)(x-x_0)\triangleq a+b(x-x_0),$$

其中 x_0 为 x 的一个临近点. 因此，我们得到局部估计量

$$\big(\widehat{E}(Y|X=x),\widehat{b}\big)=\underset{a,b}{\mathrm{argmin}}\sum_{i=1}^{n}K\Big(\frac{x-X_i}{h}\Big)\big(Y_i-a-b(x-X_i)\big)^2, \tag{2.4}$$

$K(\cdot)$ 为一个给定的核函数. 令 $K_h(x-X_i)=K((x-X_i)/h)$, 我们得到一个加权最小二乘估计的矩阵形式

$$\widehat{E}(Y|X=x)=(1,0)(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}, \tag{2.5}$$

其中

$$\mathbf{Y}=\begin{pmatrix}Y_1\\ \vdots\\ Y_n\end{pmatrix},\quad \mathbf{X}=\begin{pmatrix}1 & (x-X_1)\\ \vdots & \vdots\\ 1 & (x-X_n)\end{pmatrix},$$

且 $\mathbf{W}=\mathrm{diag}\,(K_h(x-X_1),\cdots,K_h(x-X_n))$. 因此，就可以得到 $E(E(Y|X=x))^2$ 的估计量

$$\widehat{E}\,(E(Y|X=x))^2=\frac{1}{n}\sum_{i=1}^{n}\big(\widehat{E}(Y|X=X_i)\big)^2\triangleq\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(X_i))^2. \tag{2.6}$$

记

$$S_j=\sum_{i=1}^{n}K_h(x-X_i)(x-X_i)^j,\qquad j=0,1,2 \tag{2.7}$$

和

$$m_l=\sum_{i=1}^{n}K_h(x-X_i)(x-X_i)^l\,Y_i,\qquad l=0,1. \tag{2.8}$$

通过简单的代数计算，我们可以得到 $\widehat{f}(x)$ 的一组数值解：

$$\widehat{f}(x)=(1,0)\begin{bmatrix}S_0 & S_1\\ S_1 & S_2\end{bmatrix}^{-1}\begin{pmatrix}m_0\\ m_1\end{pmatrix}=\frac{\sum\limits_{i=1}^{n}K_h(x-X_i)\big(1-S_1S_2^{-1}(x-X_i)\big)Y_i}{\sum\limits_{i=1}^{n}K_h(x-X_i)\big(1-S_1S_2^{-1}(x-X_i)\big)}. \tag{2.9}$$

再将总体均值和方差替换为其样本形式， SEV 的估计量可以记为

$$\widehat{\mathrm{SEV}}(Y|X)=\frac{n^{-1}\sum\limits_{i=1}^{n}\widehat{f}(X_i)^2-\overline{Y}^2}{S_Y^2}. \tag{2.10}$$

接下来我们考虑函数 $K_h(x-X)$ 中带宽的选择问题，这也是实践中一个重要的问题. 在这里，我们参考 [19] 中交叉验证的方法，考虑如下带宽

$$h_{\mathrm{opt}}=\underset{h>0}{\mathrm{argmin}}\sum_{k=1}^{n}(Y_k-\widehat{E}_{(h,-k)}(Y|X=X_k))^2,$$

其中， $\widehat{E}_{(h,-k)}\,(Y|X=X_k)$ 为给定参数 h, 并且排除第 k 个样本下的局部线性估计量. 值得注意的是，[19] 中的 GMCs 成对出现，而此处我们仅对其中的一个方向 $\mathrm{SEV}\,(Y|X)$ 进行估计，因此我们在带宽的选择中也仅对这一方向的均方误差进行优化，故这一式子要相比于 [19] 中的更为简洁.

2.2 SEV 局部估计的渐进性质

在这一部分中我们分析 SEV 局部估计量的渐进正态性. 因此我们需要如下的条件：

条件 1 (i) 回归函数 $f(\cdot)$ 有有界且连续的二阶导数.

(ii) 条件方差 $\sigma^2(x) = \mathrm{Var}\,(Y|X=x)$ 有界且连续.

(iii) 协变量 X 的边际密度函数 g_X 连续且在一个区间 (a_0, b_0) 上不为 0.

(iv) 核函数 $K(\cdot)$ 是一个有界的连续密度函数且满足：

$$\int_{-\infty}^{\infty} yK(y)\,\mathrm{d}y = 0, \qquad \int_{-\infty}^{\infty} y^4K(y)\,\mathrm{d}y < \infty.$$

由此我们可以得到如下引理：

引理 1 在条件 1 下，若有 $h \to 0$ 及 $nh \to \infty$, 则对任意 $x \in (a_0, b_0)$, 估计量 (2.10) 的条件 MSE 满足

$$\begin{aligned} E\left[(\widehat{f}(x) - f(x))^2|X_1, \cdots, X_n\right] =& \frac{1}{4}\Big(\widehat{f}''(x)\int_{-\infty}^{\infty} u^2K(u)\,\mathrm{d}u\Big)^2h^4 \\ &+ \frac{1}{nh}g^{-1}(x)\sigma^2(x)\int_{-\infty}^{\infty} K^2(u)\,\mathrm{d}u + o\Big(h^4 + \frac{1}{nh}\Big). \end{aligned} \tag{2.11}$$

另外我们还有估计量 (2.10) 的偏差满足：

$$E\widehat{f}(x) - f(x) = \frac{1}{2}f''(x)\int_{-\infty}^{\infty} u^2K(u)\,\mathrm{d}uh^2 + o(h^2). \tag{2.12}$$

该引理的证明可以参考 [18], 这里不再赘述. 由该引理可以得到

定理 1 在条件 1 下，若有 $nh^2 \to \infty$ 及 $nh^4 \to 0$, 我们有

$$\sqrt{n}(A^T\Sigma A)^{-\frac{1}{2}}(\widehat{\mathrm{SEV}}(Y|X) - \mathrm{SEV}(Y|X)) \Rightarrow N(0,1), \tag{2.13}$$

其中

$$\mathbf{\Sigma} = \mathrm{Cov}\Big(\sigma_Y^{-2}\int\Big(2Y_i - \frac{\phi^{Y|X}(x)}{f^X(x)}\Big)\frac{\phi^{Y|X}(x)}{f^X(x)}\frac{1}{h}K\Big(\frac{x - X_i}{h}\Big)\,\mathrm{d}x, \qquad \frac{Y_i}{\sigma_Y}, Y_i^2\Big),$$

且

$$\mathbf{A} = \Big(1,\ -\frac{2\mu_Y}{\sigma_Y} + 2\,\mu_Y\,\sigma_Y^{-3}\Big(\int\frac{(\phi^{Y|X}(x))^2}{f^X(x)}\,\mathrm{d}x - \mu_Y^2\Big),\ -\sigma_Y^{-4}\Big(\int\frac{(\phi^{Y|X}(x))^2}{f^X(x)}\,\mathrm{d}x - \mu_Y^2\Big)\Big)^T,$$

μ_Y 和 σ_Y 分别为 Y 的总体均值和总体标准差.

定理 1 给出了局部线性估计下 SEV 的渐进性质, 其结论类似于广义相关系数 GMC 的结果. 不同的是, 由于我们的局部线性估计形式 (2.10) 相比于 GMC 的估计形式较为简单, 因此我们也就得到了更为简洁的渐进性, 即我们的结论 (2.13) 中并没有他们的偏差项

$$\Big(\int\frac{(\phi^{Y|X}(x))^2}{f^X(x)}\,\mathrm{d}x - \mu_Y^2\Big)\Big(\frac{1}{\sigma_Y^2 + h^2\mathrm{Var}_K} - \frac{1}{\sigma_Y^2}\Big),$$

该项中各符号的具体含义可以参考 [19] 中的定理 3.2.

2.3 超高维特征筛选

接下来我们将 SEV 的局部线性估计量运用到特征筛选过程中. 令 $\mathbf{Y}=(Y_1,\cdots,Y_n)^T$, $\mathbf{X}=(\mathbf{X}_1,\cdots,\mathbf{X}_p)$, 其中 $\mathbf{X}_k=(X_{1k},\cdots,X_{nk})^T$, $1\le k\le p$, n 为样本大小, p 为特征数. 在整篇文章中, 我们均假设特征的维度 p 远大于样本的大小 n, 即 $n\ll p$, 并且我们先对特征进行标准化使得 $\sum\limits_{i=1}^{n}X_{ik}=0$ 且 $\frac{1}{n-1}\sum\limits_{i=1}^{n}X_{ik}^2=1$, 对 $k=1,\cdots,p$ 成立. 对所有对 Y 有影响的特征, 我们将它们组成的集合记为

$$\mathfrak{M}=\{1\le k\le p:\ \text{特征 } X_k \text{ 对因变量 } Y\text{有影响}\}. \tag{2.14}$$

对所有 $k\in\mathfrak{M}$, 我们称 X_k 为活跃变量, $\mathfrak{M}$ 即为活跃变量集. 为了辨别某个特征 X_k 是否为活跃变量, 我们用 $\mathrm{SEV}\,(Y|X_k)$ 作为标准来度量 X_k 与 Y 之间的相关关系. 定义为 $\omega_k=\mathrm{SEV}\,(Y|X_k),\ k=1,\cdots,p$. 由此我们得到一个新的集合

$$\mathfrak{M}^*=\{1\le k\le p:\ \omega_k>0\}. \tag{2.15}$$

该集合 $\mathfrak{M}^*$ 包含了度量 SEV 下, 所有对因变量 Y 有贡献的特征, 所以由定义, 我们有 $\mathfrak{M}^*=\mathfrak{M}$. 记 $\widehat{\omega}_k$ 为 ω_k 的估计量, 我们定义另一个门限集合

$$\widehat{\mathfrak{M}}^*=\{1\le k\le p:\ \widehat{\omega}_k\ge cn^{-\tau}\}. \tag{2.16}$$

其中常数 $c>0$ 和 $\tau\ge 0$ 需要事前指定并符合一定的条件. 核心问题就转化为了如何选择集合 $\widehat{\mathfrak{M}}^*$ 的大小以使其能够包含所有活跃变量构成的集合 $\mathfrak{M}^*$. 在实践中, 鉴于难以给出合适的常数 c 和 τ 的具体数值满足 $\mathfrak{M}^*\subseteq\widehat{\mathfrak{M}}^*$ 且集合 $\widehat{\mathfrak{M}}^*$ 的大小适中. 所以这里我们仍是采用特征筛选领域最常用的方法, 即先对 $\{\widehat{\omega}_k,\ k=1,\cdots,p\}$ 进行降序排序, 然后挑选前面最大的 d 个特征以构成一个行之有效的集合:

$$\widehat{\mathfrak{M}}_d^*=\{1\le k\le p:\ \widehat{\omega}_k\ \text{为所有}\ \widehat{\mathfrak{M}}^*\ \text{中最大的}\ d\ \text{个}\}. \tag{2.17}$$

在数值模拟过程中, 我们用 $\widehat{\mathfrak{M}}_d^*$ 代替 $\widehat{\mathfrak{M}}^*$, 并令 d 为一个关于 n 的函数.

另外, SEVIS 能够满足特征筛选领域中最主要的 sure screening 性质, 即当样本量 $n\to\infty$ 时, 活跃变量集 $\mathfrak{M}^*$ 以概率趋向于 1 包含于我们所选择的集合 $\widehat{\mathfrak{M}}^*$ 中, 且这一收敛速度为 n 的指数阶. 其相关证明过程已在 [17] 中进行了详细严格的推导, 这里就不再赘述.

3 数值模拟

在这一部分中, 我们利用蒙特卡洛模拟对比局部线性估计下的 SEV 方法与其在核估计下的表现情况, 为避免混淆, 本节中我们均用 SEV-lo 指代用局部线性估计的 SEVIS, 而 SEV-kn 则是指代传统的核估计方法. 除此之外, 我们也加入了几个其它的

方法作为对照，包括 SIS[8], SIRS[13], Q-SIS[16] (Q-SIS 表现要优于同样运用分位回归思想的 QaSIS 方法，这里我们采用参数 $\alpha = 0.75$, 这也是原文章作者所采用的参数之一) 以及 DC-SIS[14]. 在他们之中， SIS 为特征筛选的最初方法，而另外三个方法则各自运用了不用的统计学思想，他们的相同点在于均为 model-free 方法，同样地， SEVIS 也属于 model-free 方法. 此外，我们令 $d_1 = [n/\log(n)]$, $d_2 = 2[n/\log(n)]$, $d_3 = 3[n/\log(n)]$ 以衡量不同阀值下的模型表现. 此处 $[x]$ 代表实数 x 的整数部分. 每个模拟均重复 500 次. 我们用两种标准衡量模型的表现结果： $\mathbf{P}_a$ 和 $\mathbf{S}$. $\mathbf{P}_a$ 代表不同阀值下重要特征被选中概率的平均值，而 $\mathbf{S}$ 代表选中所有重要变量所需要的最小集合大小，这里我们给出其在 $5\%, 25\%, 50\%, 75\%, 95\%$ 分位数下的值. $\mathbf{P}_a$ 越大说明模型选中重要变量的概率越高，表现越好，理想状况下应接近于 1; $\mathbf{S}$ 越小说明重要变量均排在前列，相依性度量越准确，鉴于我们的例子中均为 3 个重要变量，所以 $\mathbf{S}$ 至少为 3, 理想状况下应接近于 3. 除此之外，我们还计算了各种方法的单次模拟平均耗时，旨在比较不同估计方式下 SEVIS 的运算效率，这也是超高维数据领域里面较为受人关注的一个方面 (考虑到我们模型中的不同参数几乎不会影响算法的效率，所以每个例子的耗时仅仅计算第一组参数下的数据).

例 1　在这个例子中，我们考虑几种方法在三角函数下的表现. 我们分别令

$$f_1(x) = \sin(2\pi x), \qquad f_2(x) = \cos(2\pi x), \qquad f_3(x) = \sin(2\pi x)^2.$$

组成可加模型

$$Y = c_1 f_1(X_1) + c_2 f_2(X_2) + c_3 f_3(X_3) + \varepsilon.$$

特征 $X = (X_1, \cdots, X_p)$ 产生于随机效应模型

$$X_j = \frac{W_j + \rho U}{1 + \rho}, \qquad j = 1, \cdots, p.$$

其中的 $W_1, \cdots, W_p, U$ 均来自于独立的 [0,1] 上的均匀分布， ε 来自于标准正态分布，显然，随着参数 ρ 的不断增大，各个特征之间的相关系数也不断上升，在这个例子中我们取 $\rho = 0.5, 1, 2$ 这三组不同的参数以衡量特征间不同相关系数时各个模型的表现. 除此之外，参数 $(c_1, c_2, c_3) = (2, 1, 4)$, 样本大小 $n = 200$, 特征维度 $p = 2000$, 如之前所述，重要的特征个数为 3.

表 1 总结了例 1 的模拟结果. 其中 3-5 列为不用阀值 d 下各个模型的表现，随后的 5 列为不同分位数下对应 $\mathbf{S}$ 的值. 可以看出在不同的相关系数下，基于局部线性估计的 SEV-lo 方法都要略优于基于核估计的 SEV-kn, 两者都要明显优于其它模型. 但从计算速度来看，由于 SEVIS 属于非参数估计的范畴，相较于其它参数估计运算速度相对较慢，但可以看出 SEV-lo 方法相对于 SEV-kn 还是在速度上有一定的提升. 对于其他对照组的情况看，尽管 SIS 有最快的运算速度，但由于其方法本身局限于一般线性模型下，所以在各组参数下均只有 1/3 左右的正确率，最小模型 $\mathbf{S}$ 也明显较大，同样表现不佳的还有 SIRS, QSIS 方法和 DCSIS 方法均有 2/3 的准确率，相对要优于 SIS 与 SIRS.

表 1 例 1 中结果 $\mathbf{P}_a$, $\mathbf{S}$ 和 time

		$\mathbf{P}_a$			$\mathbf{S}$					time
ρ	Method	d1	d2	d3	5%	25%	50%	75%	95%	
0.5	SIS	34.20%	35.33%	36.80%	456.8	969	1447.5	1719.75	1946	0.005s
	SIRS	34.33%	35.53%	36.73%	466.75	991	1356.5	1741.25	1968	9.33s
	QSIS	51.47%	60.07%	65.27%	45.9	189	372.5	720.5	1333.6	8.74s
	DCSIS	65.47%	71.07%	74.40%	23.95	99	236.5	533.25	1114.15	1.37s*
	SEV-kn	86.33%	89.40%	91.53%	3	5	19.5	113.25	516.5	35.53s
	SEV-lo	91.07%	94.40%	95.67%	3	3	9	41.5	337.35	29.60s
1	SIS	34.33%	35.60%	36.40%	441.4	1004	1414.5	1727	1952	
	SIRS	34.60%	36.27%	37.53%	384.05	891	1329.5	1691	1927.05	
	QSIS	60.27%	65.40%	68.33%	37	166.5	423	931.5	1610.9	
	DCSIS	59.73%	64.13%	67.13%	34.95	222.75	506	1053.75	1686.05	
	SEV-kn	78.07%	81.53%	83.60%	3	18	101	369	1136.95	
	SEV-lo	82.20%	85.73%	88.87%	3	9	44	173.25	774.5	
2	SIS	29.07%	32.53%	34.73%	426	992.75	1414	1707.25	1944.05	
	SIRS	27.87%	32.73%	37.13%	261.7	770.25	1245.5	1600.75	1925.1	
	QSIS	51.60%	59.87%	66.20%	21.95	155.75	372	746.75	1594.1	
	DCSIS	61.20%	67.47%	71.27%	19	125.25	369.5	816.75	1518.05	
	SEV-kn	82.20%	86.73%	88.80%	3	8	41	191	653.4	
	SEV-lo	81.27%	86.87%	90.00%	3	10	38	135	528.65	

由于 DCSIS 的代码来自于 R 语言包 energy, 其核心函数基于 C 语言编写, 其它 5 个方法均用 R 语言编写, 所以由于不同语言导致 DCSIS 算法速度可能不具有可比性, 但由于我们主要目的在于比较 SEV-kn 与 SEV-lo 的算法效率, 所以并不影响我们的结果.

例 2 [18] 中提到, 当数值 $|f'(x)/f(x)|$ 较大时, 核估计存在较大的偏差, 导致了其估计的 MSE 较大. 本例旨在探讨该情况下两种模型的结果比较. 同例 1 样, 我们仍旧构造一个 3 个重要特征的可加模型

$$Y = c_1 f_1(X_1) + c_2 f_2(X_2) + c_3 f_3(X_3) + \varepsilon.$$

其中

$$f_1(x) = x^2, \qquad f_2(x) = xI\left(x > Q_{0.8}(x)\right), \qquad f_3(x) = \exp(x).$$

$Q_{0.8}(x)$ 指代向量 x 的 80% 分位数, 旨在考察仅在部分极端情况下对因变量 Y 有影响的特征能否被选出, X^2 与 $\exp(x)$ 用来考察特征与因变量关系为非线性, 以及特征分布出现异方差情况下各个模型的表现. 与例 1 不同的是, 这里我们的特征向量由正态分布生成, 即 $X = (X_1, \cdots, X_p)$ 服从均值为 0, 协方差阵为 $\Sigma = (\sigma_{ij})_{p\times p}$ 的正态分布, 其中 $\sigma_{ii} = \sigma^2,\ i = 1, \cdots, p;\ \sigma_{ij} = \sigma^2 * \rho,\ i \neq j$, 我们用参数 σ 控制数值 $|f'(x)/f(x)|$ 的大小, ρ 控制特征间的相关系数. 本例中我们取 $\sigma = 0.5, 0.25,\ \rho = 0, 0.25$, 误差项仍旧服从标准正态分布, 参数 $(c_1, c_2, c_3) = (3, 3, 2),\ n, p$ 与例 1 相同, 模拟仍旧重复 500 次. 结果显示在表 2 中.

表 2 例 2 中结果 $\mathbf{P}_a$, $\mathbf{S}$ 和 time

			$\mathbf{P}_a$			$\mathbf{S}$					time
ρ	σ	Method	d1	d2	d3	5%	25%	50%	75%	95%	
0	0.5	SIS	69.80%	71.40%	72.47%	10	194.75	676	1347.25	1853.3	0.005s
		SIRS	67.40%	68.60%	69.87%	53.9	390.75	918	1507	1888.35	9.21s
		QSIS	98.87%	99.60%	99.87%	3	3	4	7	25.1	8.68s
		DCSIS	81.80%	89.80%	93.07%	4.95	17	41	89.5	323	1.37s
		SEV-kn	99.80%	99.93%	100.00%	3	3	3	3	7	48.05s
		SEV-lo	99.60%	99.93%	100.00%	3	3	3	3	9	29.43s
0.25	0.5	SIS	67.40%	68.87%	70.00%	37.9	344.5	1015	1685	1963.1	
		SIRS	65.73%	66.53%	67.00%	267.55	802.75	1406	1810.5	1972.1	
		QSIS	82.53%	86.40%	89.47%	3	11.75	40.5	171.5	608.4	
		DCSIS	68.27%	73.20%	76.20%	9	82	224.5	587	1212.2	
		SEV-kn	98.93%	99.80%	99.87%	3	3	3	5	29	
		SEV-lo	99.13%	99.67%	99.93%	3	3	3	4	21	
0	0.25	SIS	64.87%	67.13%	68.53%	58.95	344.75	784.5	1408.25	1890.3	
		SIRS	62.93%	65.33%	66.93%	95.95	433.75	863	1431.5	1897.75	
		QSIS	69.33%	73.20%	76.47%	16.9	82	227	540.5	1227.4	
		DCSIS	60.27%	64.67%	67.40%	50.75	232.75	532	925	1609.85	
		SEV-kn	81.93%	85.13%	87.53%	3	9	47	200	946.5	
		SEV-lo	85.87%	89.80%	92.40%	3	7	23	99.25	453.55	
0.25	0.25	SIS	62.87%	65.27%	66.87%	102.9	525.25	1047.5	1539.25	1923	
		SIRS	61.60%	63.80%	65.40%	135	570.5	1110.5	1607.75	1943.05	
		QSIS	64.87%	67.67%	69.40%	37.75	260.5	593.5	1064.25	1725.5	
		DCSIS	57.20%	63.27%	65.80%	66.75	321.25	720.5	1217.25	1760.3	
		SEV-kn	74.93%	78.53%	81.20%	4	34	145	447	1337.15	
		SEV-lo	80.87%	85.27%	88.40%	3	12	54.5	227.5	843.25	

如表 2 所示. 当特征 x 的边际标准差较高, 即 $\sigma = 0.5$ 时, SEV-kn 与 SEV-lo 的结果相差无几, 均能以较高概率选中所有重要特征, 且 3 个重要特征分列所有特征前三名的概率均至少超过 50%, 当特征之间不相关, 即 $\rho = 0$ 时, 这一概率超过了 75%. 两者的结果均要明显好于对照组的几种方法, 可以看出, 当特征相关性略微提升时, QSIS 与 DCSIS 受到的影响比较明显. 而 SIS 与 SIRS 在不同参数下均只有 2/3 的准确率. 当边际标准差降低, 即 $\sigma = 0.25$ 时, $|f'(x)/f(x)|$ 的值上升, 导致两种方法的结果都有不同程度的降低, 但此时 SEV-kn 估计准确度要明显低于 SEV-lo, 这在一定程度上印证了我们的设想及 [18] 中的结论.

4　总结与展望

在非参数估计的领域中, 局部线性估计比起 Nadaraya-Watson 核估计具有偏差较小的特点, 将其运用到特征筛选领域中无疑能在一定程度上增加变量选择的准确性. 但现有的非参数特征筛选方法中, 大多数还是基于核估计的方式对各自方法中的非参数部分进行估计. 因此, 本文将原本使用核估计方式估计的非参数特征筛选方法 SEVIS 改

为运用局部线性估计进行估计，并通过两个蒙特卡洛模拟比较了两种估计方式表现之间的差异，对未来研究者在方法选择上可能具有一定的参考意义.

在这篇文章中，我们采用一种非线性，非对称度量方式 $\mathrm{SEV}\,(Y|X)$ 来对所有特征进行筛选以排除那些明显不重要的成分，具体来说，当 $\mathrm{SEV}\,(Y|X)$ 的估计量取值较大时，我们就认为对应的特征相对于因变量具有更强的解释能力，也就在回归模型中具有更重要的地位. 在今后的研究中，我们将进一步分析和考察 $\mathrm{SEV}\,(Y|X)$ 估计量的方差对结果的影响，以及它是否适合用来进行特征筛选. $\mathrm{SEV}\,(Y|X)$ 估计量的方差牵涉到四阶矩 (峰度) 的问题，通常被认为更加稳健且适合用来考虑尾部性质，将其运用到特征筛选方法中可能在某些情况下能够得到更加准确的结果.

5 附录：定理 1 的证明

令 $f(x)=E[Y|X=x]$, 首先我们考虑 $\widehat{E}[f^2(x)]-E[f^2(x)]$. 有

$$\widehat{E}[f^2(x)]-E[f^2(x)]=(\widehat{E}[f^2(x)]-E_nE[\widehat{f}^2(x)])+(E_nE[\widehat{f}^2(x)]-E[f^2(x)]). \tag{5.1}$$

上式右边的第二部分等价于

$$E[E_n\widehat{f}^2(x)-f^2(x)]=\int(E_n\widehat{f}^2(x)-f^2(x))g(x)\,\mathrm{d}x, \tag{5.2}$$

其中， $g(x)$ 为 x 的边际密度函数，因为

$$E_n\widehat{f}^2(x)-f^2(x)=E_n[\widehat{f}(x)-f(x)]^2+2E_n[\widehat{f}(x)-f(x)]f(x). \tag{5.3}$$

由引理 1 及 (2.12), (5.3) 右边的阶数为 $O(h^4)+O(\frac{1}{nh})+O(h^2)=O\big(\frac{1}{nh}\big)+O(h^2)$ 将其代回 (5.2) 我们有

$$E[E_n\widehat{f}^2(x)-f^2(x)]\le\sup_x|E_n\widehat{f}^2(x)-f^2(x)|\int g(x)\,\mathrm{d}x=O\left(\frac{1}{nh}\right)+O(h^2). \tag{5.4}$$

另一方面， (5.1) 式右边第一部分等于

$$\frac{1}{n}\sum_{i=1}^n\widehat{f}^2(X_i)-E\widehat{f}^2(X_i).$$

因此我们有：

$$\begin{aligned}
&\widehat{\mathrm{SEV}}(Y|X)-\mathrm{SEV}(Y|X)=S_Y^{-2}(\widehat{E}[f(x)]^2-\overline{Y}^2)-\sigma_Y^{-2}(E[f(x)]^2-\mu_Y^2)\\
=&S_Y^{-2}(\widehat{E}[f(x)]^2-E[f(x)]^2)-S_Y^{-2}(\overline{Y}^2-\mu_Y^2)+(E[f(x)]^2-\mu_Y^2)(S_Y^{-2}-\sigma_Y^{-2})\\
=&S_Y^{-2}(\frac{1}{n}\sum_{i=1}^n[\widehat{f}(X_i)]^2-E[\widehat{f}(X_i)]^2)-S_Y^{-2}(\overline{Y}^2-\mu_Y^2)\\
&+(E[f(x)]^2-\mu_Y^2)(S_Y^{-2}-\sigma_Y^{-2})+O\Big(h^2+\frac{1}{nh}\Big).
\end{aligned} \tag{5.5}$$

由条件 $nh^2 \to \infty$ 和 $nh^4 \to 0$, 我们有 $\sqrt{n}(O(h^2 + \frac{1}{nh})) \to 0$. 因此，通过中心极限定理，我们有

$$\sqrt{n}\Big(\sigma_Y^{-2}\big(\widehat{f}^2(X_i) - E[\widehat{f}^2(X_i)]\big),\ \sigma_Y^{-1}\Big(\overline{Y} - \mu_Y\Big), \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - EY^2\Big) \longrightarrow N(0, \Sigma), \tag{5.6}$$

其中

$$\Sigma = \mathrm{Cov}\left(\frac{\widehat{f}^2(X_i)}{\sigma_Y^2}, \frac{Y_i}{\sigma_Y}, Y_i^2\right).$$

定义 $T_1 = \frac{\widehat{f}^2(X_i)}{\sigma_Y^2},\ T_2 = \frac{\overline{Y}}{\sigma_Y},\ T_3 = \frac{1}{n}\sum\limits_{i=1}^{n} Y_i^2$ 以及函数 $f_1(T_1) = T_1,\ f_2(T_2) = -T_2^2$, $f_3(T_2, T_3) = (T_3 - \sigma_Y^2 T_2^2)^{-1}(E[f(x)]^2 - \mu_Y^2)$.

由于我们有 $ET_2 = \frac{\mu_Y}{\sigma_Y},\ ET_3 = \mu_Y^2 + \sigma_Y^2$,

化简可得函数 f_1, f_2, f_3 对 T_1, T_2, T_3 的偏导数分别为：

$$\begin{aligned}
A_1 &= \frac{\partial f_1(T_1)}{\partial T_1}\Big|_{T_1 = ET_1} = 1;\\
A_{21} &= \frac{\partial f_2(T_2)}{\partial T_2}\Big|_{T_2 = ET_2} = -\frac{2\mu_Y}{\sigma_Y};\\
A_{22} &= \frac{\partial f_3(T_2, T_3)}{\partial T_2}\Big|_{T_2 = ET_2, T_3 = ET_3} = 2\,\mu_Y\,\sigma_Y^{-3}\Big(\int \frac{(\phi^{Y|X}(x))^2}{f^X(x)}\,\mathrm{d}x - \mu_Y^2\Big);\\
A_3 &= \frac{\partial f_3(T_2, T_3)}{\partial T_3}\Big|_{T_2 = ET_2, T_3 = ET_3} = -\sigma_Y^{-4}\Big(\int \frac{(\phi^{Y|X}(x))^2}{f^X(x)}\,\mathrm{d}x - \mu_Y^2\Big).
\end{aligned}$$

令 $\mathbf{A} = (A_1, A_{21} + A_{22}, A_3)^T$. 由多元 delta 方法，我们有渐进正态性：

$$\begin{aligned}
&\sqrt{n}\big(f_1(T_1) - f_1(ET_1) + f_2(T_2) - f_2(ET_2) + f_3(T_2, T_3) - f_3(ET_2, ET_3)\big)\\
\longrightarrow &N(0, \mathbf{A}^T\mathbf{\Sigma}\mathbf{A}).
\end{aligned} \tag{5.7}$$

由于 $S_Y^2 \to \sigma_Y^2$ 几乎必然成立. 基于 Slutsky 定理及 $nh^4 \to 0$. 由 (5.5) 我们有

$$\begin{aligned}
&\sqrt{n}(\widehat{\mathrm{SEV}}(Y|X) - \mathrm{SEV}\,(Y|X))\\
=&\sqrt{n}\Big(S_Y^{-2}\Big(\frac{1}{n}\sum_{i=1}^{n}[\widehat{f}(X_i)]^2 - E[\widehat{f}(X_i)]^2\Big) - S_Y^{-2}(\overline{Y}^2 - \mu_Y^2)\\
&+ (E[f(x)]^2 - \mu_Y^2)(S_Y^{-2} - \sigma_Y^{-2}) + O\Big(h^2 + \frac{1}{nh}\Big)\Big)\\
\longrightarrow&\sqrt{n}\Big(\frac{\widehat{f}^2(X_i)}{\sigma_Y^2} - \frac{E\widehat{f}^2(X_i)}{\sigma_Y^2} - \frac{\overline{Y}^2}{\sigma_Y^2} + \frac{\mu_Y^2}{\sigma_Y^2} + \Big(\frac{1}{S_Y^2} - \frac{1}{\sigma_Y^2}\Big)\Big(\int \frac{(\phi^{Y|X}(x))^2}{f^X(x)}\,\mathrm{d}x - \mu_Y^2\Big)\Big)\\
=&\sqrt{n}(f_1(T_1) - f_1(ET_1) + f_2(T_2) - f_2(ET_2) + f_3(T_2, T_3) - f_3(ET_2, ET_3))\\
\longrightarrow&N(0, \mathbf{A}^T\mathbf{\Sigma}\mathbf{A}).
\end{aligned} \tag{5.8}$$

所以，我们有

$$\sqrt{n}(\mathbf{A}^T\mathbf{\Sigma}\mathbf{A})^{-\frac{1}{2}}(\widehat{\mathrm{SEV}}(Y|X) - \mathrm{SEV}(Y|X)) \longrightarrow N(0, 1), \tag{5.9}$$

即证明了定理 1. 证毕.

致谢 作者衷心感谢审稿专家提出的修改意见.

参 考 文 献

[1] Tibshirani R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B.*, 1996, 58: 267–288

[2] Yuan M, Lin Y. Medel Selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 2006, 68: 49–67

[3] Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 2006, 101: 1418–1429

[4] Fan J. Comment on "Wavelets in statistics: a review" by A. Antoniadis. *J. Ital. Statist. Soc.*, 1997, 2: 131–138

[5] Zou H, Hasties T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 2005, 67: 301–320

[6] Candes E, Tao T. The Dantzig Selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, 2007, 35: 2313–2404

[7] Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research*, 2009, 10: 1829–1853

[8] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*, 2008, 70: 849–911

[9] Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 2010, 38: 3567–3604

[10] Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, 2011, 106: 544–557

[11] Song R, Yi F, Zou H. On Varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, 2012, 24: 1735–1752

[12] Liu J, Li R, Wu R. Feature selection for varying coefficient models with ultrahigh dimensional covariates. Technical Report, Departmernt of Statitics, Pennsylvania State University, 2013

[13] Zhu L P, Li L, Zhu L X. Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association*, 2011, 106: 1464–1475

[14] Li R, Zhong W, Zhu L. Feature Screening via Distance Correlation Learning. *Journal of American Statistical Association*, 2012, 107: 1129–1139

[15] He X, Wang L, Hong H G. Quantile-adaptive model-free variable scerrning for high-dimensional heterogeneous data. *The Annals of Statistics*, 2013, 41: 342–369

[16] Wu Y, Yin G. Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika*, 2014, 102: 65–76

[17] Chen M, Lian Y, Chen Z, Zhang Z. Sure Explained Variability and Independence Screening. *Journal of Nonparametric Statistics*, 2017, 1–35

[18] Fan J. Design-adaptive Nonparametric Regression. *Journal of the American Statistical Association*, 1992, 87: 998–1004

[19] Zheng S, Shi N Z, Zhang Z. Generalized Measures of Correlation for Asymmetry, Nonlinearity, and Beyond. *Journal of the American Statistical Association*, 2012, 107: 1239–1252

Local Estimation of Sure Explained Variability Independence Screening and Its Application for Ultrahigh-dimensional Data

LIAN Yimin

(*Department of Statistics and Finance, University of Science and Technology of China, Hefei* 230026*, China*)

(*E-mail: ymlian@mail.ustc.edu.cn*)

CHEN Zhao

(*Department of Statistics, Pennsylvania State University, State College, PA* 16802*, U.S.A.*)

(*E-mail: chenzhao1985@gmail.com*)

SHU Mingliang

(*Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing* 100190*, China*)

(*E-mail: shumingliang12@mails.ucas.ac.cn*)

Abstract It's quite an concerned question about how to extract the true features among ultrahigh-dimensional data, especially in today's era of big data, this question plays a key role in many related industries. The core idea of feature screening is excluding those features that significantly unrelated to response variably to solve this question. The Sure Explained Variability and Independence Screening method has obvious advantages in handling the asymmetry and nonlinearity situations compare to the methods before. But it's kernel estimation still has space for improvement. From this point, we change the kernel estimation to local estimation which been known as more accurate and effective. Some simulations about the feature screening in special situations also proof our view and show that our new algorithm is more efficient than the kernel-based one.

Key words feature screening; local estimation; SEVIS; ultrahigh-dimensional data

MR(2000) Subject Classification 62G05

Chinese Library Classification 0212.7