第 37 卷第 1 期 湖 南 工 业 大 学 学 报 Vol.37 No.1
2023 年 1 月 Journal of Hunan University of Technology Jan. 2023

doi:10.3969/j.issn.1673-9833.2023.01.012

含内生变量的高维部分线性模型特征筛选

陈海燕[1]，赵培信[1,2]

（1. 重庆工商大学 数学与统计学院，重庆 400067；2. 经济社会应用统计重庆市重点实验室，重庆 400067）

摘 要：主要研究了含有内生性协变量的超高维部分线性模型的重要特征筛选和变量选择问题。首先，为了消除数据内生性对特征筛选带来的选择性偏差，结合内生性协变量与工具变量的相关结构，给出了一个区分重要协变量和不重要协变量的充分条件，进而提出一种衡量变量边际效用的特征筛选方法。其次，利用提出的特征筛选方法，并结合剖面估计思想和两阶段正则估计方法，提出了一种识别重要协变量的变量选择方法。最后，在一定正则条件下，理论证明了所提出的变量选择方法可以消除数据内生性对变量选择带来的影响，从而保证了对协变量重要性具有排序一致性。

关键词：部分线性模型；内生性协变量；变量选择；高维数据

中图分类号：O212.7 **文献标志码**：A **文章编号**：1673-9833(2023)01-0083-08

引文格式：陈海燕，赵培信. 含内生变量的高维部分线性模型特征筛选 [J]. 湖南工业大学学报，2023，37(1)：83-90.

Feature Selection of High-Dimensional Partial Linear Models with Endogenous Covariates

CHEN Haiyan[1]，ZHAO Peixin[1,2]

（1. School of Mathematics and Statistics，Chongqing Technology and Business University，Chongqing 400067，China；
2. Chongqing Key Laboratory of Social Economy and Applied Statistics，Chongqing 400067，China）

Abstract：This study is mainly focused on the important feature selection and variable selection of ultra-high dimensional partial linear models with endogenous covariates. Firstly, in view of an elimination of the selective deviation of data endogeneity on feature selection, combined with the correlation structure of endogenous covariates and instrumental variables, a sufficient condition has thus been given to distinguish important covariates from unimportant covariates, followed by a proposed feature selection method to measure the marginal utility of variables. Secondly, a variable selection method is adopted for the identification of important covariates by using the proposed feature screening method, with a combination of the profile estimation idea and the two-stage regular estimation method. Finally, under certain regular conditions, the theory proves that the proposed variable selection method helps to eliminate the influence of data endogeneity on variable selection, thus ensuring the consistency of ranking the importance of covariates.

Keywords：partial linear model；endogenous covariate；variable selection；high-dimensional data

收稿日期：2022-03-19

基金项目：国家社会科学基金资助项目（18BTJ035）；重庆市自然科学基金资助面上项目（cstc2020jcyj-msxmX0006）

作者简介：陈海燕（1998-），女，湖南永州人，重庆工商大学硕士生，主要研究方向为复杂数据分析及应用，E-mail：2269703267@qq.com

通信作者：赵培信（1981-），男，山东曹县人，重庆工商大学教授，博士生导师，主要研究方向为复杂数据分析及应用，E-mail：zpx81@163.com

0 引言

部分线性模型 [1] 同时含有参数分量和非参数分量，在对实际问题建模过程中兼具经典参数模型和非参数模型的优点，目前已被广泛地应用于社会科学、计量经济学以及生物医学等领域。另外，随着现代数据收集技术的不断发展，研究者们能够在科学研究的各个领域以较低成本收集到大量的高维数据。这种大数据的统计推断过程中，往往会遇到超高维情况，即数据的维数远远大于样本量，从而导致经典的统计推断理论将无法直接应用。

目前关于超高维数据的统计推断问题，一般是先利用一些变量筛选方法，从大量的数据中筛选出一些重要变量，然后基于所筛选出的重要变量进行统计建模。关于超高维数据下部分线性模型的变量筛选问题，杨宜平等 [2] 结合样条方法和 Dantzig 或 Lasso 进行变量选择和未知参数估计。赖秋楠等 [3] 将超高维部分线性模型转化为高维线性模型，考虑了协变量间的相关性，提出了 profile 贪婪向前回归变量筛选方法。杨鑫等 [4] 基于 profile 最小二乘方法和保留正则化方法，提出了新的变量选择方法。但是这些文献均是在假定超高维数据为外生协变量的情况下进行讨论的。Fan J. Q.[5]、Lin W. [6] 等指出，在超高维模型中存在许多可能导致违反外生性假定的因素，例如选择偏差、测量误差和遗漏变量等。因此对超高维数据统计建模过程中假定所有变量均为外生协变量是具有限制性且往往是不现实的。在违反外生性假设时，现有的基于边际特征筛选方法可能会筛选出那些隐藏的重要变量，并产生较多的假阳性重要变量。

目前，关于超高维内生性协变量的重要变量选择问题研究还不多。针对含内生协变量的超高维线性模型，Fan J. Q. 等 [5] 通过构建惩罚聚焦广义矩法准则函数，有效实现了降维，并证明了模型存在内生性时，该方法也具有 Oracle 性质。Lin W. 等 [6] 提出了一个两阶段正则化框架，通过使用稀疏诱导惩罚函数，将经典的两阶段最小二乘法（two stage least square，2SLS）扩展到高维。Hu Q. Q. 等 [7] 提出了一种新的特征筛选工具来衡量预测变量的边际效用，然后引入两阶段正则化框架来识别重要的预测变量。但是，对超高维内生性数据下部分线性模型的重要变量选择问题目前还没有相关研究。为此，本文在假定部分协变量为内生协变量的情况下，研究超高维部分线性模型的重要变量筛选问题。

具体地，结合工具变量调整技术，本文提出了一种新的重要变量筛选方法。理论上证明了所提出的筛选方法具有排序一致性。这意味着依据效用测度，总是可以大概率地将重要变量排在不重要变量之前，从而保证可以清晰地区分重要变量和不重要变量。

1 基于工具变量调整的重要变量筛选方法

本节中，假定模型中线性部分的维数 p 远远超过样本量 n，且维数 p 随着样本量 n 呈指数型增长。本文考虑的部分线性模型结构如下：

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + g\left(U_i\right) + \varepsilon_i,\ 1 \leqslant i \leqslant n \text{。} \tag{1}$$

式中：$\boldsymbol{X}_i$ 为 p 维协变量，且

$$\boldsymbol{X}_i = \left(X_{i1}, X_{i2},\ \cdots,\ X_{ip}\right)^{\mathrm{T}},\ 1 \leqslant i \leqslant n,\ p \gg n\text{；}$$

$\boldsymbol{\beta}$ 为未知参数的 p 维向量，且

$$\boldsymbol{\beta} = \left(\beta_1,\ \beta_2,\ \cdots,\ \beta_p\right)^{\mathrm{T}}\text{；}$$

$g(\cdot)$ 为未知的非参数函数；

U_i 为一维变量；

ε_i 为模型误差。

由于部分线性模型中：$E\left(\varepsilon_i \middle| \boldsymbol{X}_i\right) \neq 0$，即模型中至少存在一个内生协变量。因此，引入如下工具变量：

$$\boldsymbol{X}_i = \boldsymbol{\Gamma}\boldsymbol{Z}_i + \boldsymbol{e}_i,\ \ 1 \leqslant i \leqslant n\text{，} \tag{2}$$

式中：$\boldsymbol{Z}_i$ 为对应的 q 维的工具变量向量，且

$$\boldsymbol{Z}_i = \left(Z_{i1},\ Z_{i2},\ \cdots,\ Z_{iq}\right)^{\mathrm{T}}\text{；}$$

$\boldsymbol{\Gamma}$ 为 $p \times q$ 维的未知参数矩阵，

$\boldsymbol{e}$ 为模型误差，且 $\boldsymbol{e}=(\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n)$，其中 $\boldsymbol{e}_i=(\boldsymbol{e}_{i1}, \boldsymbol{e}_{i2}, \cdots, \boldsymbol{e}_{ip})^{\mathrm{T}}$，且满足 $E\left(\boldsymbol{e}_i \middle| \boldsymbol{Z}_i\right) = 0$。

综上所述，考虑模型

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + g\left(U_i\right) + \varepsilon_i,\ \ \boldsymbol{X}_i = \boldsymbol{\Gamma}\mathrm{Z}_i + \boldsymbol{e}_i,\ 1 \leqslant i \leqslant n\text{，} \tag{3}$$

满足如下条件：

$$E\left(\varepsilon_i \middle| \boldsymbol{X}_i\right) \neq 0,\ E\left(\boldsymbol{e}_i \middle| \boldsymbol{Z}_i\right) = 0,\ E\left(\varepsilon_i \middle| \boldsymbol{Z}_i\right) = 0\text{。}$$

注意，$E\left(\varepsilon_i \middle| \boldsymbol{X}_i\right) \neq 0$ 意味着模型中至少存在一个内生协变量。此外，外生协变量可以是工具变量。

假设真参数 $\boldsymbol{\beta}$ 是稀疏的，即集合 $A=\{j{:}\beta_j \neq 0, 1 \leqslant j \leqslant p\}$ 很小，则本文的目标是估计集合 A。

经典的两阶段最小二乘（2SLS）[8] 和两阶段正则化（two stage regularization，2SR）[6] 将协变量替换为它们对工具变量的期望。更具体地说，变量首先在工具变量上回归，然后响应变量在变量第一阶段的预测结果上回归。然而，因为变量和工具变量的维度随着样本量呈指数增长，2SLS 方法和 2SR 方法的性能分别面临众多工具变量的维度灾难和计算成本的

问题。因此，需要探索新的方法来获取集合 A。

注意：如果响应变量在工具变量上进行回归，根据上述模型（3），可以得到如下模型：

$$Y_i = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha} + g(U_i) + \xi_i,\ 1 \leqslant i \leqslant n。\qquad（4）$$

式中：$\boldsymbol{\alpha}$ 是 $q\times 1$ 维向量，且 $\boldsymbol{\alpha}=\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\beta}$;

ξ_i 为新误差，且 $\xi_i=\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i$。

为了找到一个特征筛选工具来估计活跃集 A，首先考虑一个例子。在模型（4）中，很容易得到：

$$\operatorname{cov}(\boldsymbol{X}_j, \boldsymbol{Y}) = \operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha} + g(\boldsymbol{U}) + \boldsymbol{\xi}) = \operatorname{cov}(\boldsymbol{X}_j, \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}) + \operatorname{cov}(\boldsymbol{X}_j, g(\boldsymbol{U})) + \operatorname{cov}(\boldsymbol{X}_j, \boldsymbol{\xi}),$$

另一方面，

$$\operatorname{cov}(\boldsymbol{X}_j, \boldsymbol{Y}) = \operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + g(\boldsymbol{U}) + \boldsymbol{\varepsilon}) = \operatorname{cov}(\boldsymbol{X}_j, \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) + \operatorname{cov}(\boldsymbol{X}_j, g(\boldsymbol{U})) + \operatorname{cov}(\boldsymbol{X}_j, \boldsymbol{\varepsilon})。$$

结合上面的方程，可以得到：

$$\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) + \operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{\varepsilon}) = \operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}) + \operatorname{cov}(\boldsymbol{X}_j, \boldsymbol{\xi}),$$

进一步展开，得到：

$$\operatorname{cov}(\boldsymbol{X}_j, \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}) = \operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\beta}) = \left(\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{X}^{\mathrm{T}}) - \operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{e}^{\mathrm{T}})\right)\boldsymbol{\beta}。\qquad（5）$$

若 β_j=0，则有 $\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}) = 0$；若 $\beta_j \neq 0$，即 $j \in A$，下面对 $\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha})$ 的取值情况进行分析。

式（5）可表示成

$$\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}) = \left(\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}) - \operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{e}_A^{\mathrm{T}})\right)\boldsymbol{\beta}_A = \left(\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}) - \operatorname{cov}(\boldsymbol{e}_j,\ \boldsymbol{e}_A^{\mathrm{T}})\right)\boldsymbol{\beta}_A。$$

记 $\operatorname{Var}(\boldsymbol{X}_j) = \sigma_{X_j}^2$，$\operatorname{Var}(\boldsymbol{e}_j) = \sigma_{e_j}^2$，并且假设 $\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{X}_k^{\mathrm{T}}) = 0$，$\operatorname{cov}(\boldsymbol{e}_j,\ \boldsymbol{e}_k^{\mathrm{T}}) = 0$，其中，$j \neq k$，$j \in \{1, 2, \cdots, p\}, k \in A$。

通过这些假设，可以得到如下结论：

1）当 $j \notin A$ 时，$\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}) = 0$；

2）当 $j \in A$ 时，$\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}) = \left(\sigma_{X_j}^2 - \sigma_{e_j}^2\right)\beta_j$。

由于 $\boldsymbol{X}_i=\boldsymbol{\Gamma}\boldsymbol{Z}_i+\boldsymbol{e}_i$, $1 \leqslant i \leqslant n$，且 $\boldsymbol{Z}_i$ 和 $\boldsymbol{e}_i$ 是不相关的，因此，可以得到 $\sigma_{X_j}^2 > \sigma_{e_j}^2$，进一步可以得到 $\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}) > 0$。

定义：

$$\varPhi_j = \operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}),\ \phi_j = \varPhi_j^2,\ 1 \leqslant j \leqslant p。$$

结合上述分析，可以得到：

假设 $\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{X}_k^{\mathrm{T}}) = 0$，$\operatorname{cov}(\boldsymbol{e}_j,\ \boldsymbol{e}_k^{\mathrm{T}}) = 0$，其中，$j \neq k$，$j \in \{1, 2, \cdots, p\}, k \in A$。当且仅当 β_j=0 时，有 ϕ_j=0。因此，在上述观察的驱动下，考虑使用 ϕ_j 作为特征筛选工具来估计活跃集 A。值得指出的是，本文的目标是获得一个有用的筛选工具，用于分离活跃和非活跃预测变量。为了保证 ϕ_j 可以被用作筛选工具，并不需要任何的 ϕ_j $\left(j \in \Theta,\ \Theta = \{1, 2, \cdots, p\}\backslash A\right)$ 精确为 0，只需要非活跃预测变量的 ϕ_j 总是比活跃预测变量的 ϕ_j 小，这足以达到给预测变量排序的目的。因此，可以用一个更弱的条件：$\operatorname{cov}(\boldsymbol{X}_j,\ \boldsymbol{X}_k^{\mathrm{T}}) = 0$，$j \neq k, j \in \{1, 2, \cdots, p\}, k \in A$ 代替假设，由此确保 ϕ_j 是一个筛选工具。

2　重要变量筛选迭代算法

根据部分线性模型的剖面估计思想，首先假定 $\boldsymbol{\beta}$ 已知，则模型（1）可被看作是一个非参数回归模型：

$$Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} = g(U_i) + \varepsilon_i,\ 1 \leqslant i \leqslant n,\qquad（6）$$

由 $g(u) = E\left(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} \middle| U_i = u\right)$，有

$$Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} = E\left(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} \middle| U_i\right) + \varepsilon_i,$$

经整理得：

$$Y_i - E(Y_i|U_i) = \left(\boldsymbol{X}_i - E(\boldsymbol{X}_i|U_i)\right)^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i,$$

定义：$\tilde{Y}_i = Y_i - E(Y_i|U_i)$，

$$\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i - E(\boldsymbol{X}_i|U_i) = \left(\tilde{X}_{i1}, \tilde{X}_{i2}, \cdots, \tilde{X}_{ip}\right)^{\mathrm{T}},$$

即 $\tilde{X}_{ij} = X_{ij} - E(X_{ij}|U_i),\ 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant p$，则：

$$\tilde{Y}_i = \tilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i。$$

对于非参数 $g(u)$ 可使用局部线性光滑方法构造其估计量，它能减少 Nadaraya-Watson 核估计的偏差和 Gasser-Müller 估计的方差，并能够避免核估计的边界效应，在边界点和内点有相同的收敛速度。设回归函数 $g(u)$ 在 u 的邻域内有连续的一阶导数，如果 U_i 在 u 的一个小邻域内，可用一个线性函数局部地逼近回归函数 $g(U_i)$，有：

$$g(U_i) \approx g(u) + g'(u)(U_i - u) \triangleq a + b(U_i - u),$$

式中 a、b 为回归系数。

因为假定 $\boldsymbol{\beta}$ 已知，可通过极小化下式加权最小二乘目标函数求 a 和 b，

$$\sum_{i=1}^{n}\left\{Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - a - b(U_i - u)\right\}^2 K_h(U_i - u),$$

式中：$K_h(\cdot)=K(\cdot/h)/h$，其中 $K(\cdot)$ 为核函数，h 为窗宽，

且 $h>0$。

记 $\hat{a}$ 为 a 的估计，经简单计算得：

$$\hat{a}=\frac{\sum_{i=1}^{n} w_i(u)\left(Y_i-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)}{\sum_{i=1}^{n} w_i(u)},$$

式中：

$$w_i(u)=K_h(U_i-u)\left[S_{n,2}(u)-(U_i-u)S_{n,1}(u)\right],$$

其中 $S_{n,l}(u)=\frac{1}{n}\sum_{i=1}^{n}(U_i-u)^l K_h(U_i-u),\ l=0,1,2$。因此，有

$$\hat{g}(u,\ \boldsymbol{\beta})=\sum_{i=1}^{n} w_{ni}(u)\left(Y_i-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right),$$

式中：$w_{ni}(\cdot)$ 为权函数，且 $w_{ni}(u)=w_i(u)\Big/\sum_{k=1}^{n} w_k(u)$。基于该权函数，可以分别定义 $g_0(u)=E(\boldsymbol{Y}|U=u)$ 和 $\boldsymbol{g_X}(u)=\left(g_1(u),g_2(u),\cdots,g_p(u)\right)^{\mathrm{T}}=E(\boldsymbol{X}|U=u)$ 的局部线性估计如下：

$$\hat{g}_0(u)=\sum_{i=1}^{n} w_{ni}(u)Y_i,$$

$$\hat{\boldsymbol{g}}_{\boldsymbol{X}}(u)=\sum_{i=1}^{n} w_{ni}(u)\boldsymbol{X}_i。$$

接下来，使用两阶段方法获取 ϕ_j 的样本估计量。给定一个来自 $\{\boldsymbol{X},\ \boldsymbol{Z},\ \boldsymbol{Y}\}$ 的随机样本 $\{\boldsymbol{X}_i,\ \boldsymbol{Z}_i,\ Y_i\}$。为了便于表述，假设样本预测变量都是标准化的，即 $p^{-1}\sum_{j=1}^{p}X_{ij}=0,\ p^{-1}\sum_{j=1}^{p}X_{ij}^2=1,\ 1\leqslant i\leqslant n$，其中，$X_{ij}$ 表示 $\boldsymbol{X}_i$ 的第 j 个元素。

接下来设计两阶段方法。

1）第一阶段。当响应变量在工具变量上回归时得到预值 $\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{a}}$，其中 $\hat{\boldsymbol{a}}$ 是 $\boldsymbol{a}$ 的估计值。因为工具变量的数量随着样本量呈指数级增长，所以大多数变量选择方法可能会失效或面临巨大的计算挑战。因此，需要一个特征筛选程序来筛选变量。响应变量在工具变量上回归时，工具变量就是外生变量，可以使用现有的特征筛选程序来筛选变量，例如：准确独立排序筛选（sure independent ranking and screening，SIRS）[9]。工具变量的活跃集记作：$B=\{k=1,2,\cdots,q{:}\alpha_k\neq 0\}$。$B_s$ 是通过筛选过程得到的估计活跃集，$d=|B_s|$ 是 B_s 的大小。则第一阶段估计量由正则化方法定义为

$$\hat{\boldsymbol{\alpha}}=\arg\min_{\boldsymbol{\alpha},B_s\in R^d}\frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{Z}_{B_s}^{\mathrm{T}}\boldsymbol{\alpha}_{B_s}\right\|_2^2+p_\lambda\left(\boldsymbol{\alpha}_{B_s}\right)。$$

式中：$p_\lambda(\cdot)$ 为惩罚函数；λ 为调和参数，且 $\lambda\geqslant 0$。

例如，在 Lasso[10] 中，作者选择了 $p_\lambda\left(\boldsymbol{\alpha}_{B_s}\right)=\sum_{j\in B_s}\lambda\left|\alpha_j\right|$ 作为惩罚函数。通过使用 Lasso 或其他变量选择方法可以得到估计值 $\hat{B}$ 和 $\hat{\boldsymbol{\alpha}}$。$\boldsymbol{Y}$ 的预测值由 $\sum_{i\in\hat{B}}\boldsymbol{Z}_i^{\mathrm{T}}\hat{\alpha}_i+g(U_i)$ 组成。

注意到 $\boldsymbol{Y}=\boldsymbol{X}_A^{\mathrm{T}}\boldsymbol{\beta}_A+g(\boldsymbol{U})+\boldsymbol{\varepsilon},\ \boldsymbol{X}_A=\boldsymbol{\Gamma}_A\boldsymbol{Z}+\boldsymbol{e}_A$，因此，在第一阶段可以得到如下模型：

$$\boldsymbol{Y}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Gamma}_A^{\mathrm{T}}\boldsymbol{\beta}_A+\boldsymbol{e}_A^{\mathrm{T}}\boldsymbol{\beta}_A+g(\boldsymbol{U})+\boldsymbol{\varepsilon},\tag{7}$$

式中：$\boldsymbol{\Gamma}_A^{\mathrm{T}}$ 是由矩阵 $\boldsymbol{\Gamma}^{\mathrm{T}}$ 中与 A 对应的列向量构成的矩阵。

为了保证模型（7）有意义，显然需要 $\alpha_A=\boldsymbol{\Gamma}_A^{\mathrm{T}}\boldsymbol{\beta}_A\neq 0$，这意味着 $Z_B\cap IV_A\neq\varnothing$，其中，$IV_A$ 表示所有解释变量 $\boldsymbol{X}_j, j\in A$ 的工具变量。假设 $\mathrm{cov}(\boldsymbol{X}_j,\boldsymbol{X}_k)=0$，$j\neq k$，$j\in\{1,2,\cdots,n\}$，$k\in A$，可得：$IV_j\cap IV_k=\varnothing,\ \alpha_t=\Gamma_{tj}\beta_j$，其中，$IV_j$ 为包括所有解释变量 $\boldsymbol{X}_j$ 的工具变量（instrumental variable，IV）。因此，当 $\Gamma_{tj}\neq 0$ 且 $\beta_j\neq 0$ 时，$\alpha_t=\Gamma_{tj}\beta_j\neq 0$。这意味着，解释重要变量的工具变量肯定是第 1 阶段中的活跃变量，即 $IV_A\subset Z_B$。此外，β_j 越大，a_t 越大，这使得变量更容易在第一阶段被选择出来。因此，解释重要变量的工具变量更容易被选择出来。注意到：如果 Γ_{tj} 很小（即 $\boldsymbol{Z}_t$ 是 $\boldsymbol{X}_j$ 的弱工具变量），那么 a_t 会很小，而小的 a_t 可能会影响选择结果。因此，上面提出的程序更加倾向于强工具变量。

值得注意的是，没有必要对解释非活跃预测变量的工具变量做出任何条件假设。因此，所有非活跃预测变量都可以使用一个工具变量，即使这个工具变量非常弱，这意味着虽然预测变量 $\boldsymbol{X}$ 是高维的，但工具变量 $\boldsymbol{Z}$ 并不需要是高维的。通常情况下，要求工具变量的数量应不小于用于识别的预测变量的数量，但是，系数的识别对筛选目标并不重要。即使系数不确定，仍然可以确定活动回归量。因此，当工具变量的维度不太高时，可以忽略第 1 阶段的特征筛选。从理论上讲，当工具变量的维数小于样本量时，可以使用“普通最小二乘法”代替变量选择过程。

2）第二阶段。以第一阶段的预测值 $\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{a}}$ 代替 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{a}$，可以得到 ϕ_j 的估计值：

$$\hat{\phi}_j=\hat{\Phi}_j^2,\ 1\leqslant j\leqslant n,$$

式中：$\hat{\Phi}_j=\frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\hat{B}}X_{ij}Z_{it}\hat{\alpha}_t$，$Z_{it}$ 表示 $\boldsymbol{Z}_i$ 的第 t 个元素。

总之，建议将所有候选预测变量 $\boldsymbol{X}_j$ 根据 $\hat{\phi}_j$ 从大到小的顺序进行排序，然后选择排名靠前的作为活跃预测变量，具体定义如下：

$$\hat{A}=\left\{j:\hat{\phi}_j \geqslant \psi_n,\ 1\leqslant j\leqslant n\right\},$$

式中，ψ_n 是给定的阈值参数。

值得注意的是，无论模型中是否存在内生协变量，本文所提出的筛选程序都是可行的。

3　理论结果

本节将讨论所提出的筛选程序的理论性质。下列条件是为了方便技术证明，尽管它们可能不是最弱的条件。

C1）随机误差 $\boldsymbol{e}$ 的条件。给定为

$$\operatorname{cov}\left(\boldsymbol{e}_j,\boldsymbol{e}_k\right)=0,\ j\neq k,\ j\in\{1,2,\cdots,n\},\ k\in A。$$

C2）协变量 $\boldsymbol{X}$ 的条件。给定为

$$\max_{j\in\Theta}\lambda_{\max}\left\{\operatorname{cov}\left(\boldsymbol{X}_A,\boldsymbol{X}_j\right)\operatorname{cov}\left(\boldsymbol{X}_j,\boldsymbol{X}_A^{\mathrm{T}}\right)\right\}<\min_{j\in A}\lambda_{\min}\left\{\operatorname{cov}\left(\boldsymbol{X}_A,\boldsymbol{X}_j\right)\operatorname{cov}\left(\boldsymbol{X}_j,\boldsymbol{X}_A^{\mathrm{T}}\right)-\sigma_{e_j}^4\boldsymbol{E}_j\right\}。$$

式中：$\lambda_{\max}(\boldsymbol{D})$ 和 $\lambda_{\min}(\boldsymbol{D})$ 分别为矩阵 $\boldsymbol{D}$ 最大和最小的特征值；$\boldsymbol{E}_j$ 是一个 $|A|\times|A|$ 的矩阵，对所有 $s\neq j',\ t\neq j'$ 有 $\left(\boldsymbol{E}_j\right)_{j'j'}=1,\ \left(\boldsymbol{E}_j\right)_{st}=0$。$|A|$ 是 A 的大小，j' 表示 $\boldsymbol{X}_j$ 在 $\boldsymbol{X}_A$ 中的排名。为了简单起见，当 $\boldsymbol{D}$ 是一个数时记为 D，仍然使用符号 $\lambda_{\max}(\boldsymbol{D})$ 和 $\lambda_{\min}(\boldsymbol{D})$，在这种情况下，它们都等于 D。

C3）工具变量 $\boldsymbol{Z}$ 的条件。

C3-a）存在正常数 K_1、K_2 和 κ，使得

$$\max_{1\leqslant k\leqslant q}P\left(\left|\boldsymbol{Z}_k\right|>t\right)\leqslant K_1\exp\left(-K_2t^{\kappa}\right);$$

C3-b）$$\frac{\lambda_{\max}\left\{\operatorname{cov}\left(\boldsymbol{Z}_B^{\mathrm{T}},\boldsymbol{Z}_{B^c}\right)\operatorname{cov}\left(\boldsymbol{Z}_{B^c}^{\mathrm{T}},\boldsymbol{Z}_B\right)\right\}}{\lambda_{\min}^2\left\{\operatorname{cov}\left(\boldsymbol{Z}_B^{\mathrm{T}},\boldsymbol{Z}_B\right)\right\}}<\frac{\min_{t\in B}\tilde{\omega}_t}{\lambda_{\max}\left\{\Omega_B\right\}},$$

式中 $\Omega_B=E\left\{\Omega_B\left(\boldsymbol{Y}\right)^{\mathrm{T}}\Omega_B\left(\boldsymbol{Y}\right)\right\}$，其中 $\Omega_B(\boldsymbol{Y})$ 是 $1\times|B|$ 维向量，包括所有 $\Omega_t(y),t\in B$，$\Omega_t(y)=\operatorname{cov}\left(\boldsymbol{Z}_t,\mathbf{1}\left(\boldsymbol{Y}<y\right)\right)$，$\tilde{\omega}_t=E\left(\Omega_t^2\left(\boldsymbol{Y}\right)\right),\ t\in B$。

C3-c）线性条件为

$$E\left\{\boldsymbol{Z}^{\mathrm{T}}\middle|\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}\right\}=\operatorname{cov}\left(\boldsymbol{Z}^{\mathrm{T}},\boldsymbol{Z}^{\mathrm{T}}\right)\boldsymbol{\alpha}\left\{\operatorname{cov}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}\right)\right\}^{-1}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}},\ \forall\boldsymbol{\alpha}。$$

C3-d）对于一些 $\varphi_0>0$，所有满足 $\left\|\boldsymbol{\alpha}_{B^c}\right\|_1\leqslant3\left\|\boldsymbol{\alpha}_B\right\|_1$ 的 $\boldsymbol{\alpha}$，有以下式子成立：

$$\left\|\boldsymbol{\alpha}_B\right\|_1^2\leqslant\left(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\,\boldsymbol{\alpha}\right)|B|/\varphi_0^2,$$

式中：$\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ 是 $\boldsymbol{Z}$ 的总体协方差，$|B|$ 是 B 的大小。

C4）活跃集 A 和 B 之间的关系为

$$Z_B\bigcap IV_A\neq\varnothing。$$

条件 C2 描述了预测变量之间的相关性。更具体地说，C2 中不等式左侧的项测量了 $\boldsymbol{X}_A$ 中的活跃预测变量和 $\boldsymbol{X}_\Theta$ 中的非活跃预测变量之间的相关性，右边的项评估了 $\boldsymbol{X}_A$ 中的活跃预测变量之间的相关性。对这种情况作如下评论：首先，C2 对活跃变量之间的相关性没有任何要求；其次，可以验证，当 $\boldsymbol{X}_A$ 和 $\boldsymbol{X}_\Theta$ 不相关时，条件 C2 可以替换为更弱的条件。因为 $\boldsymbol{X}_A$ 和 $\boldsymbol{X}_\Theta$ 不相关时，$\phi_j=0$（$j\in\Theta$），只需要个别的活跃变量 $\min_{j\in A}\phi_j>0$ 的信号强度，这显然是比条件 C2 更弱的条件。条件 C3-a 中，工具变量的尾部分布是克莱默型大偏差的常规技术要求。例如，工具变量 $\boldsymbol{Z}_k(1\leqslant k\leqslant n)$ 是正态或亚高斯分布，就有 $\kappa=2$。使用 SIRS（第一阶段的筛选部分）需要条件 C3-b 和 C3-c，细节具体见文献 [9]。当工具变量的维度不高，无法使用变量选择方法 (如 Lasso) 处理时，这些条件可以忽略不计。条件 C3-d 是 Lasso 估计中 l_1 误差界的正则 $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ 相容条件。基于条件 C3-d，$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}$ 相容条件也大概率成立，其中 $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}$ 是样本协方差矩阵 $\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}/n$，$\boldsymbol{Z}=(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_n)$[11]。条件 C4 是为了保证模型（7）有意义。

接下来，介绍所提出的筛选程序的理论性质，这些理论性质是新筛选方法的主要理论基础。

定理 1　在条件 C1、C2、C4 下，有如下不等式关系成立：

$$\max_{j\in\Theta}\phi_j<\min_{j\in A}\phi_j。$$

定理 1 表明，通过对效用测度 ϕ_j（$j=1,2,\cdots,n$）进行排序，从理论上总可以将活跃变量排在非活跃变量之前。

定理 2　在 C1~C4 条件下，如果有 $p=o\left\{\exp\left(an\right)\right\}$，$q=o\left\{\exp\left(a'n\right)\right\}$，对任意固定的 $a>0,\ a'>0$，对所有的 $\tau>0$，存在常数 R_1,R_2,R_3,R_4 和 δ_τ，使得：

$$\Pr\left(\sup_{j=1,2,\cdots,n}\left|\phi_j-\hat{\phi}_j\right|>\tau\right)\leqslant p\left(d\exp\left(-n\delta_\tau^2R_1\right)+|B|\exp\left(-n\delta_\tau^2R_2-nR_3\right)+\exp\left(-n\delta_\tau^2R_2-nR_4\right)\right),$$

进一步记 $\epsilon=\min_{j\in A}\phi_j-\max_{j\in\Theta}\phi_j$，此时，存在一个常数 $\delta_{\epsilon/2}^2$，使得：

$$\Pr\left(\max_{j\in\Theta}\hat{\phi}_j<\min_{j\in A}\hat{\phi}_j\right)\geqslant1-2p\left(d\exp\left(-n\delta_{\epsilon/2}^2R_1\right)+|B|\exp\left(-n\delta_{\epsilon/2}^2R_2-nR_3\right)+\exp\left(-n\delta_{\epsilon/2}^2R_2-nR_4\right)\right)。$$

式中：$d=|B_s|$，δ_τ 和 $\delta^2_{\epsilon/2}$ 分别取决于 τ 和 ϵ；R_1、R_2、R_3、R_4 取决于 τ、ϵ、φ_0、ξ 的协方差 σ^2_ξ、$\max\limits_{t\in B}|\hat{\alpha}_t|$ $\max\limits_{j=1,2,\cdots,n;t=1,2,\cdots,n}\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_t\right)\right|$ 以及 $\hat{B}$ 的大小。

定理 2 表明，通过对效用测度的估计 $\hat{\phi}_j$ (j=1, 2 , ⋯ , n) 进行排序，可以使得活跃变量以指数级速度渐近趋于 1 的概率排在非活跃变量之前。

4 定理证明

定理 1 的证明 基于模型（3）和模型（4），设 $\boldsymbol{\beta}=\left(\beta_1,\beta_2,\cdots,\beta_p\right)^{\mathrm{T}}$ 为真系数，可以得到：

$$\begin{aligned}\phi_j&=\left[\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Y}-g(\boldsymbol{U})-\boldsymbol{\xi}\right)\right]^2=\\&\left[\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Y}-g(\boldsymbol{U})-\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{\varepsilon}\right)\right]^2=\\&\left[\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Y}-g(\boldsymbol{U})-\boldsymbol{\varepsilon}\right)-\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\beta}\right)\right]^2=\\&\left[\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}\boldsymbol{\beta}_A\right)-\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{e}_A^{\mathrm{T}}\boldsymbol{\beta}_A\right)\right]^2,\end{aligned}$$

式中：$\boldsymbol{\beta}_A$ 由所有的 β_j，$j\in A$ 组成；$\boldsymbol{e}_A^{\mathrm{T}}$ 由所有的 $\boldsymbol{e}_j^{\mathrm{T}}$，$j\in A$ 组成。

由于 $\boldsymbol{X}_j=\boldsymbol{\Gamma}\boldsymbol{Z}_j+\boldsymbol{e}_j$，其中 $\boldsymbol{Z}_j$ 是 $\boldsymbol{Z}$ 的第 j 列，基于条件 C1，可以得到：当 $j\in\Theta$ 时，$\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{e}_A^{\mathrm{T}}\right)=0$；当 $j\in A$ 时，$\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{e}_A^{\mathrm{T}}\boldsymbol{\beta}_A\right)=\sigma^2_{e_j}\beta_j$。因此，如果 $j\in\Theta$，则有

$$\begin{aligned}\phi_j&=\left[\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}\boldsymbol{\beta}_A\right)\right]^2=\\&\boldsymbol{\beta}_A^{\mathrm{T}}\left[\mathrm{cov}\left(\boldsymbol{X}_A,\ \boldsymbol{X}_j\right)\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}\right)\right]\boldsymbol{\beta}_A\leqslant\\&\lambda_{\max}\left\{\mathrm{cov}\left(\boldsymbol{X}_A,\ \boldsymbol{X}_j\right)\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}\right)\right\}\boldsymbol{\beta}_A^{\mathrm{T}}\boldsymbol{\beta}_A。\end{aligned}$$

另一方面，如果 $j\in A$，可以得到：

$$\begin{aligned}\phi_j&=\left[\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}\boldsymbol{\beta}_A\right)-\sigma^2_{e_j}\beta_j\right]^2=\\&\boldsymbol{\beta}_A^{\mathrm{T}}\left[\mathrm{cov}\left(\boldsymbol{X}_A,\ \boldsymbol{X}_j\right)\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}\right)-\sigma^4_{e_j}\boldsymbol{r}_j^{\mathrm{T}}\boldsymbol{r}_j\right]\boldsymbol{\beta}_A=\\&\boldsymbol{\beta}_A^{\mathrm{T}}\left[\mathrm{cov}\left(\boldsymbol{X}_A,\ \boldsymbol{X}_j\right)\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}\right)-\sigma^4_{e_j}\boldsymbol{R}_j\right]\boldsymbol{\beta}_A\geqslant\\&\lambda_{\min}\left\{\mathrm{cov}\left(\boldsymbol{X}_A,\ \boldsymbol{X}_j\right)\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{X}_A^{\mathrm{T}}\right)-\sigma^4_{e_j}\boldsymbol{R}_j\right\}\boldsymbol{\beta}_A^{\mathrm{T}}\boldsymbol{\beta}_A,\end{aligned}$$

式中：$\boldsymbol{R}_j$ 是一个 $|A|\times|A|$ 矩阵，对所有 $s\neq j'$，$t\neq j'$，有 $\left(\boldsymbol{R}_j\right)_{j'j'}=1$ 和 $\left(\boldsymbol{R}_j\right)_{st}=0$，$|A|$ 是 A 的大小，j' 表示 $\boldsymbol{X}_j$ 在 $\boldsymbol{X}_A$ 中的排名。因为 $\lambda_{\min}(\boldsymbol{C})\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\leqslant\boldsymbol{x}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}\leqslant\lambda_{\max}(\boldsymbol{C})\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ 适用于任何对称矩阵 $\boldsymbol{C}$ 和向量 $\boldsymbol{x}\neq 0$，故此处不等式成立。根据条件 C2，可以直接得到定理 1 的结果。

定理 2 的证明 为了提高可读性，将证明分为如下两个主要步骤。

步骤 1 首先

$$\begin{aligned}&\Pr\left(\sup_{j=1,2,\cdots,n}\left|\phi_j-\hat{\phi}_j\right|>\tau\right)\leqslant\\&\qquad p\left(q\exp\left(-n\delta_\tau^2R_1\right)+\exp\left(-n\delta_\tau^2R_2-nR_3\right)\right)。\end{aligned}$$

注意到：$\phi_j=\Phi_j^2$ 和 $\hat{\phi}_j=\hat{\Phi}_j^2$，基于 $h(x)=x^2$ 的连续性，对每一个 $\tau>0$，存在 $\delta>0$，使得：$\left|\Phi_j-\hat{\Phi}_j\right|<\delta\Rightarrow\left|\phi_j-\hat{\phi}_j\right|<\tau$，因此，对于所有的 j，可以得到：

$$\Pr\left(\left|\phi_j-\hat{\phi}_j\right|>\tau\right)\leqslant\Pr\left(\left|\Phi_j-\hat{\Phi}_j\right|>\delta\right),$$

即 $\max\limits_{j=1,2,\cdots,n}\Pr\left(\left|\phi_j-\hat{\phi}_j\right|>\tau\right)\leqslant\max\limits_{j=1,2,\cdots,n}\Pr\left(\left|\Phi_j-\hat{\Phi}_j\right|>\delta\right)$。

对于 $\Pr\left(\left|\Phi_j-\hat{\Phi}_j\right|>\delta\right)$，因为：$\Phi_j=\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\alpha}\right)=\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_B^{\mathrm{T}}\boldsymbol{\alpha}_B\right)$ 和 $\hat{\Phi}_j=\dfrac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{t\in\hat{B}}X_{ij}\boldsymbol{Z}_{it}\hat{\alpha}_t$，可得：

$$\begin{aligned}&\Pr\left(\left|\Phi_j-\hat{\Phi}_j\right|>\delta\right)=\\&\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_B^{\mathrm{T}}\boldsymbol{\alpha}_B\right)-\frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\hat{B}}X_{ij}\boldsymbol{Z}_{it}\hat{\alpha}_t\right|>\delta\right)=\\&\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_B^{\mathrm{T}}\boldsymbol{\alpha}_B\right)-\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_{\hat{B}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}}\right)+\right.\right.\\&\left.\left.\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_{\hat{B}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}}\right)-\frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\hat{B}}X_{ij}\boldsymbol{Z}_{it}\hat{\alpha}_t\right|>\delta\right)\leqslant\\&\Pr\left(\left\{\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_B^{\mathrm{T}}\boldsymbol{\alpha}_B\right)-\mathrm{cov}\left(\boldsymbol{X}_j,\boldsymbol{Z}_{\hat{B}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}}\right)\right|+\right.\right.\\&\left.\left.\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_{\hat{B}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}}\right)-\frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\hat{B}}X_{ij}\boldsymbol{Z}_{it}\hat{\alpha}_t\right|\right\}>\delta\right)\leqslant\\&\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_B^{\mathrm{T}}\boldsymbol{\alpha}_B\right)-\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_{\hat{B}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}}\right)\right|>\delta/2\right)+\\&\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_{\hat{B}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}}\right)-\frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\hat{B}}X_{ij}\boldsymbol{Z}_{it}\hat{\alpha}_t\right|>\delta/2\right):=\\&I_1+I_2。\end{aligned}$$

注意：

$$\begin{aligned}I_1&=\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\boldsymbol{Z}_B^{\mathrm{T}}\boldsymbol{\alpha}_B\right)-\mathrm{cov}\left(\boldsymbol{X}_j,\boldsymbol{Z}_{\hat{B}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}}\right)\right|>\delta/2,B\subset B_s\right)+\\&\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\boldsymbol{Z}_B^{\mathrm{T}}\boldsymbol{\alpha}_B\right)-\mathrm{cov}\left(\boldsymbol{X}_j,\boldsymbol{Z}_{\hat{B}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}}\right)\right|>\delta/2,B\subsetneq B_s\right)\leqslant\\&\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\boldsymbol{Z}_{B_s}^{\mathrm{T}}\boldsymbol{\alpha}_{B_s}\right)-\mathrm{cov}\left(\boldsymbol{X}_j,\boldsymbol{Z}_{\hat{B}_s}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}_s}\right)\right|>\delta/2,B\subset B_s\right)+\\&\Pr\left(B\subsetneq B_s\right)=\\&\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_{B_s}^{\mathrm{T}}\right)\left(\boldsymbol{\alpha}_{B_s}-\hat{\boldsymbol{\alpha}}_{\hat{B}_s}\right)\right|>\delta/2,B\subset B_s\right)+\\&\Pr\left(B\subsetneq B_s\right)\leqslant\\&\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_{B_s}^{\mathrm{T}}\right)\left(\boldsymbol{\alpha}_{B_s}-\hat{\boldsymbol{\alpha}}_{\hat{B}_s}\right)\right|>\delta/2,\ B\subset B_s\right)+\\&\Pr\left(B\subsetneq B_s\right)\leqslant\Pr\left(M\left\|\boldsymbol{\alpha}_{B_s}-\hat{\boldsymbol{\alpha}}_{\hat{B}_s}\right\|_1>\delta/2,\ B\subset B_s\right)+\end{aligned}$$

$$\Pr\left(B\subsetneq B_s\right)=\Pr\left(\left\|\boldsymbol{\alpha}_{B_s}-\hat{\boldsymbol{\alpha}}_{\hat{B}_s}\right\|_1>\delta/(2M),\ B\subset B_s\right)+$$
$$\Pr\left(B\subsetneq B_s\right)=I_{11}+I_{12},$$

式中 $M=\max_{j=1,2,\cdots,n;t=1,2,\cdots,q}\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_t^{\mathrm{T}}\right)\right|$。

$\hat{\boldsymbol{\alpha}}$ 是第一阶段 SIRS 后的 Lasso 估计量，根据 P. Bühlmann 等 [11] 的推论 6.8，在条件 C3 的第二部分和 $\left\|\boldsymbol{\Sigma}_{\boldsymbol{Z}}-\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}\right\|_\infty\leqslant\tilde{\lambda}$，其中 $32\tilde{\lambda}/\varphi_0\leqslant1$，则 $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}$!'a相容性条件也成立。其次，根据条件 C3 的第一部分，可以得到存在正常数 K_3、K_4、K_5，对所有的 t 有

$$P\left(\left\|\boldsymbol{\Sigma}_{\boldsymbol{Z}}-\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}\right\|_\infty\geqslant K_3t+K_4\sqrt{t}+K_5\right)\leqslant\exp(-nt)。$$

因此，存在常数 $\tilde{c}$，使得：

$$P\left(\left\|\boldsymbol{\Sigma}_{\boldsymbol{Z}}-\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}\right\|_\infty\geqslant\tilde{\lambda}(\tilde{c})\right)\leqslant\exp(-n\tilde{c}),$$

其中，$32\tilde{\lambda}(\tilde{c})/\varphi_0\leqslant1$。根据 P. Bühlmann 等 [11] 的结论，对于任意 $t>0$，$\lambda_0=2\sigma_\xi\sqrt{\left(t^2+2\log d\right)/n}$，在条件 C3 下，对于 $\lambda\geqslant8\lambda_0$，存在一些常数 $\tilde{c}$，使得：

$$\Pr\left(\left\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\right\|_1>32\lambda|B|/\varphi_0^2\right)\leqslant\exp\left(-t^2/2\right)+\exp(-\tilde{c}n),$$

式中，$\tilde{c}$ 取决于 φ_0 和 $\hat{B}$ 的大小。选择 t 使得 $t^2\leqslant\dfrac{n\sigma^2\varphi_0^4}{(32\times8)^2M^2\sigma_\xi^2}-2\log d$，则有

$$I_{11}=\Pr\left(\left\|\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right\|_1>\delta/(2M)\right)\leqslant\exp\left(-t^2/2\right),$$

因此，选择 $t^2=\dfrac{n\sigma^2\varphi_0^4}{(32\times8)^2M^2\sigma_\xi^2}-2\log d$，可得：

$$I_{11}=\Pr\left(\left\|\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right\|_1>\delta/(2M)\right)=$$
$$\exp\left\{\frac{-n\sigma^2\varphi_0^4}{\left(2\times(32\times8)^2M^2\sigma_\xi^2\right)}+2\log d\right\}+\exp(-\tilde{c}n)。$$

基于文献 [9] 中的定理 2，可以得知存在一些常数 $\hat{c}$，使得：

$$I_{12}=\Pr\left(B\subsetneq B_s\right)\leqslant|B|\exp(-\hat{c}n)。$$

另一方面：

$$I_2=\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_{\hat{B}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\hat{B}}\right)-\frac{1}{n}\sum_{i=1}^n\sum_{t\in\hat{B}}X_{ij}Z_{it}\hat{\alpha}_t\right|>\delta/2\right)\leqslant$$
$$\Pr\left(\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\sum_{t\in\hat{B}}\left(\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_t\right)-\frac{1}{n}\sum_{i=1}^n\sum_{t\in\hat{B}}X_{ij}Z_{it}\right)\right|>\delta/2\right)\leqslant$$
$$\Pr\left(\left|\sum_{t\in\hat{B}}\left(\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_t\right)-\frac{1}{n}\sum_{i=1}^n\sum_{t\in\hat{B}}X_{ji}Z_{it}\right)\right|>\frac{\delta}{2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|}\right)\leqslant$$
$$\Pr\left(\sum_{t\in\hat{B}}\left|\left(\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_t\right)-\frac{1}{n}\sum_{i=1}^n\sum_{t\in\hat{B}}X_{ji}Z_{it}\right)\right|>\frac{\delta}{2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|}\right)\leqslant$$
$$\Pr\left(\max_{t\in\hat{B}}\left|\left(\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_t\right)-\frac{1}{n}\sum_{i=1}^n\sum_{t\in\hat{B}}X_{ji}Z_{it}\right)\right|>\frac{\delta}{2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|}\right)\leqslant$$
$$\exp\left(-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right)。$$

可得：

$$\Pr\left(\left|\Phi_j-\hat{\Phi}_j\right|>\delta\right)=$$
$$\Pr\left(\left|\mathrm{cov}\left(\boldsymbol{X}_j,\ \boldsymbol{Z}_B^{\mathrm{T}}\boldsymbol{\alpha}_B\right)-\frac{1}{n}\sum_{i=1}^n\sum_{t\in\hat{B}}X_{ij}Z_{it}\hat{\alpha}_t\right|>\delta\right)\leqslant$$
$$\exp\left\{-\left(n\sigma^2\varphi_0^4\right)\Big/\left(2\times(32\times8)^2M^2\sigma_\xi^2\right)+\log d-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right\}+$$
$$\exp\left(-\tilde{c}n-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right)+$$
$$\exp\left(-\hat{c}n+\log|B|-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right)。$$

因此，

$$\max_{j=1,2,\cdots,n}\Pr\left(\left|\phi_j-\hat{\phi}_j\right|>\tau\right)\leq\max_{j=1,2,\cdots,n}\Pr\left(\left|\Phi_j-\hat{\Phi}_j\right|>\delta\right)\leqslant$$
$$\exp\left\{-n\sigma^2\varphi_0^4/2\times(32\times8)^2M^2\sigma_\xi^2+\log d-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right\}+$$
$$\exp\left(-\tilde{c}n-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right)+$$
$$\exp\left(-\hat{c}n+\log|B|-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right),$$

然后，得到：

$$\Pr\left(\sup_{j=1,2,\cdots,n}\left|\phi_j-\hat{\phi}_j\right|>\tau\right)\leqslant$$
$$p\exp\left(-n\sigma^2\varphi_0^4\Big/\left(2\times(32\times8)^2M^2\sigma_\xi^2\right)+\log d-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right)+$$
$$p\exp\left(-\tilde{c}n-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right)+$$
$$p\exp\left(-\hat{c}n+\log|B|-nc_1\delta^2\Big/\left(2\max_{t\in\hat{B}}\left|\hat{\alpha}_t\right|\left|\hat{B}\right|\right)^2\right)\leqslant$$
$$p\left(d\exp\left(-n\delta_\tau^2R_1\right)+|B|\exp\left(-n\delta_\tau^2R_2-nR_3\right)+\exp\left(-n\delta_\tau^2R_2-nR_4\right)\right),$$

其中，$\delta_t=\delta$ 是强调 δ 取决于 τ，

$$R_1 = \varphi_0^4 \Big/ \left(2\times(32\times 8)^2 M^2 \sigma_\xi^2\right) + c_1 \Big/ \left(2\max_{t\in\hat{B}} |\hat{\alpha}_t| |\hat{B}|\right)^2,$$

$$R_2 = c_1 \Big/ \left(2\max_{t\in\hat{B}} |\hat{\alpha}_t| |\hat{B}|\right)^2, \quad R_3 = \widehat{c}, \quad R_4 = \tilde{c}。$$

步骤 2　接下来证明

$$\Pr\left(\max_{j\in\Theta}\hat{\phi}_j < \min_{j\in A}\hat{\phi}_j\right) \geqslant 1 - 2p\Big(d\exp\left(-n\delta_\tau^2 R_1\right) + |B|\exp\left(-n\delta_\tau^2 R_2 - nR_3\right) + \exp\left(-n\delta_\tau^2 R_2 - nR_4\right)\Big)。$$

根据假设 $\epsilon = \min_{j\in A}\phi_j - \max_{j\in\Theta}\phi_j > 0$，有

$$\Pr\left(\max_{j\in\Theta}\hat{\phi}_j < \min_{j\in A}\hat{\phi}_j\right) \geqslant \Pr\left(\min_{j\in A}\hat{\phi}_j - \min_{j\in A}\phi_j + \epsilon < \max_{j\in\Theta}\hat{\phi}_j - \max_{j\in\Theta}\phi_j\right) \geqslant \Pr\left(\sup_{j\in A}\left|\hat{\phi}_j - \phi_j\right| \geqslant \epsilon/2\right) + \Pr\left(\sup_{j\in\Theta}\left|\hat{\phi}_j - \phi_j\right| \geqslant \epsilon/2\right)。$$

令 $\tau = \epsilon/2$，利用步骤 1 即可得定理 2 的结果。

5　结语

针对超高维内生协变量的变量选择问题，结合内生协变量和工具变量的相关结构，提出了一种新的用于超高维线部分线性工具变量回归模型的两阶段特征筛选方法，其中内生协变量和工具变量的维数可以随样本量呈指数级增长。理论结果表明，该特征筛选方法在排序上具有一致性。

本文只考虑了工具变量的各分量之间相关性较弱的情况。当工具变量的各分量之间存在高度相关性时，可以使用 Hu Q. Q. 等 [12] 给出的条件特征筛选程序来处理。然而，在对内生性协变量的工具变量调整过程中，如何事先确定一个工具变量的备选集合，然后从中筛选重要的工具变量，是当前内生性数据统计建模中常遇到的难题之一。另外，值得进一步研究的问题是如何在不事先假定模型结构的前提下，完全基于内生变量与工具变量的相关结构来构造特征筛选方法。这些问题都有待进一步深入研究。

参考文献：

[1] ENGLE R F，GRANGER C W J，RICE J，et al. Semiparametric Estimates of the Relation Between Weather and Electricity Sales[J]. Journal of the American Statistical Association，1986，81：310-320.

[2] 杨宜平，薛留根，王学娟. 高维部分线性模型中的变量选择 [J]. 北京工业大学学报，2011，37(2)：291-295.
YANG Yiping，XUE Liugen，WANG Xuejuan. Variable Selection in High-Dimensional Partially Linear Models[J]. Journal of Beijing University of Technology，2011，37(2)：291-295.

[3] 赖秋楠，李玉杰，李高荣. 超高维部分线性模型的 PGFR 变量筛选 [J]. 应用概率统计，2017，33(6)：608-624.
LAI Qiunan，LI Yujie，LI Gaorong. Profile Greedy Forward Regression Variable Screening for Ultra-High Dimensional Partially Linear Model[J]. Chinese Journal of Applied Probability and Statistics，2017，33(6)：608-624.

[4] 杨　鑫，李冰月，田　萍. 超高维部分线性模型的 PRAR 变量选择 [J]. 应用概率统计，2021，37(6)：551-568.
YANG Xin，LI Bingyue，TIAN Ping. Profile Regularization After Retention Variable Selection for Ultrahigh Dimensional Partially Linear Models[J]. Chinese Journal of Applied Probability and Statistics，2021，37(6)：551-568.

[5] FAN J Q，LIAO Y. Endogeneity in High Dimensions[J]. Annals of Statistics，2014，42(3)：872-917.

[6] LIN W，FENG R，LI H Z. Regularization Methods for High-Dimensional Instrumental Variables Regression with an Application to Genetical Genomics[J]. Journal of the American Statistical Association，2015，110：270-288.

[7] HU Q Q，LIN L. Feature Screening in High Dimensional Regression with Endogenous Covariates[J]. Computational Economics，2021：1-21. https://doi.org/10.1007/s10614-021-10174-x.

[8] ANDERSON T W. Origins of the Limited Information Maximum Likelihood and Two-Stage Least Squares Estimators[J]. Journal of Econometrics，2005，127(1)：1-16.

[9] 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，2011，106：1464-1475.

[10] TIBSHIRANI R. Regression Shrinkage and Selection via the Lasso[J]. Journal of the Royal Statistical Society：Series B (Methodological)，1996，58(1)：267-288.

[11] BÜHLMANN P，VAN DE GEER S. Statistics for High-Dimensional Data[M]. Heidelberg：Springer Berlin Heidelberg，2011：146-180.

[12] HU Q Q，LIN L. Conditional Sure Independence Screening by Conditional Marginal Empirical Likelihood[J]. Annals of the Institute of Statistical Mathematics，2017，69(1)：63-96.

（责任编辑：廖友媛）