

基于方向回归的高维非参数非线性系统 变量选择及辨识^{*}

孙兵^{1,2} 程长明^{1†} 蔡巧言² 彭志科^{1,3} 张涛²

(1. 上海交通大学 机械系统与振动国家重点实验室, 上海 200240)

(2. 中国运载火箭技术研究院, 北京 100076)

(3. 宁夏大学, 机械工程学院, 宁夏 750021)

摘要 变量选择问题在诸多领域中被广泛研究, 人们发展出了许多变量选择方法。然而, 有些变量选择算法存在计算耗时问题, 有些算法在检测变量是否有贡献时仅能提供必要条件, 无法提供充分必要条件。本文基于方向回归提出了一种新的高维非参数非线性系统变量选择算法, 其假设要求更低, 计算复杂度大幅降低, 性能优于现有的变量选择算法; 且为检验变量是否对系统有贡献提供了充分必要条件。此外, 由于检测变量是否有贡献的指标并不是精确的 0, 因此当指标较小时, 很难判断变量是否冗余。为解决这一问题, 本文提出了一种惩罚优化算法, 以确保集合的收敛性。仿真算例验证了所提变量选择方法的有效性。

关键词 变量选择, 非线性系统辨识, 方向回归, 核函数, 非参数系统

中图分类号:O302

文献标志码:A

Directional Regression Based Variable Selection and Identification of High-Dimensional Nonparametric Nonlinear Systems^{*}

Sun Bing^{1,2} Cheng Changming^{1†} Cai Qiaoyan² Peng Zhike¹ Zhang Tao²

(1. The State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, Shanghai 200240, China)

(2. China Academy of Launch Vehicle Technology, Beijing 100076, China)

(3. School of Mechanical Engineering, Ningxia University, Ningxia 750021, China)

Abstract The importance of discovering significant variables from a large candidate pool is now widely recognized in many fields. There exist a number of algorithms for variable selection in the literature. Some are computationally efficient but only provide a necessary condition for, not a sufficient and necessary condition for, testing whether a variable contributes or not to the system output. The others are computationally expensive. The goal of the paper is to develop a directional variable selection algorithm that performs similar to or better than the leading algorithms for variable selection, but under weaker technical assumptions and with a much reduced computational complexity. It provides a necessary and sufficient condition for testing whether a variable contributes or not to the system. In addition, since indicators for redundant variables aren't exact zeros, it is difficult to decide whether variables are redundant or not when the indicators are small. To solve this problem, a penalty optimization algorithm is proposed to ensure the convergence of the set. Simulation and experimental results verify the effectiveness of the directional variable selection method proposed in this paper.

2024-11-25 收到第 1 稿, 2024-12-05 收到修改稿。

* 国家自然科学基金资助项目(12072188, 12121002, 12372017), National Natural Science Foundation of China(12072188, 12121002, 12372017)。

† 通信作者 E-mail:cctming@sjtu.edu.cn

Key words variable selection, nonlinear system identification, directional regression, kernel function, nonparametric system

引言

非线性系统辨识^[1,2]已被广泛应用于机械工程、控制工程、电气和电子工程以及化学工程等领域。变量选择通常是高维非参数非线性系统辨识的前提条件,因此本文研究的变量选择问题非常重要。从高维非线性系统的所有候选变量中剔除多余的变量,可以有效地简化非线性系统的辨识问题,而且不会降低辨识模型的精度。例如,预先使用变量选择方法辨识 Volterra 级数中的重要时延项,可以简化 Volterra 级数的辨识。此外,变量选择方法还可用于全局敏感性分析,发现非线性系统中的重要因素或变量。

本文研究的非线性系统是一个离散非参数非线性系统,

$$Y(k) = f[X_1(k), X_2(k), \dots, X_p(k)] + v(k), \quad k=1, 2, \dots, N, \quad (1)$$

其中, $v(k)$ 表示均值为 0、方差有限的独立同分布 (independent and identically distributed, IID) 噪声序列, $v(k)$ 与 $\mathbf{X}(k)$ 相互独立, $Y(k)$ 是系统输出, N 表示测量数据的长度。回归向量 $\mathbf{X}(k) = [X_1(k), \dots, X_p(k)]^\top$ 包含所有可能贡献的输入变量。非线性函数 $f(\cdot)$ 的结构形式未知。该函数可表示一大类非线性函数。例如,如果

$$\begin{aligned} \mathbf{X}(k) &= [u(k-1), \dots, u(k-n_u)]^\top, \text{ or} \\ \mathbf{X}(k) &= [y(k-1), \dots, y(k-n_y), u(k-1), \dots, u(k-n_u)]^\top \end{aligned} \quad (2)$$

将变为著名的有限脉冲响应非线性系统^[3]和带外界输入的非线性自回归系统(NARX)^[4],式中 u 和 y 分别表示系统的输入与输出, n_u 和 n_y 分别表示所选至当前时刻的输入和输出的步数。

常用的非线性系统辨识方法包括基函数展开方法,如多项式、傅里叶级数、广义正交基和小波方法。然而,当待辨识系统缺乏先验信息时^[5,6],基函数的选择将存在困难。核函数和局部多项式方法在先验信息较少的情况下,也能辨识非线性系统,并在某些一般化假设下得到渐进收敛结果。然而,这些方法为非参数化方法,未能利用现有非线性系统的结构对系统进行表征,导致其收敛速度较慢,并

在高维情况下,常常受到维数灾难^[7]的影响。

在许多实际问题中,系统通常是稀疏的,也就是说,不是所有的变量 $X_i(k) (i=1, \dots, p)$ 都对输出 $Y(k)$ 有贡献。为了降低系统维度,需要检测和剔除非贡献变量 $X_i(k)$,这也被称为变量选择问题。从众多候选变量中找出真正有贡献的变量对非线性系统的辨识至关重要。

高维非参数非线性系统变量选择和辨识的目标首先是检测出贡献变量 $X_1(k), \dots, X_q(k)$ 或非贡献变量 $X_{q+1}(k), \dots, X_p(k)$, 并剔除非贡献变量,然后利用贡献变量使用非参数方法辨识非线性系统。对于线性系统,人们已经提出了很多著名的变量选择方法,例如,非负绞杀(nonnegative garrote)^[8]、最小绝对收缩和选择算子(LASSO)^[9]、自适应 LASSO (adaptive LASSO)^[7]、组 LASSO (group LASSO)^[10]、最小角回归(least angle regression, LARS)^[11] 和带平滑削边绝对偏离(smoothly clipped absolute deviation, SCAD)^[12]。然而,当系统是非参数非线性系统时,若直接使用以上变量选择方法,可能会得到错误的结果。

传统非线性系统变量选择方法中,通过利用局部线性估计方法得到关于任意局部变量 \mathbf{X}^0 的导数 $\partial f / \partial X_i$, 通过导数是否为 0 判断变量贡献与否。然而,当变量数量很大时,存在维数灾难问题。此外,局部变量选择方法仅使用与 \mathbf{X}^0 相邻的局部数据,如果函数 $f(\cdot)$ 是全局的,那么局部变量选择方法就会遗失大量有用信息。且局部导数是否为 0 并非变量选择的充分条件,而是必要条件。

在统计领域,人们常使用充分降维(sufficient dimension reduction, SDR)理论进行变量选择。例如,Cook^[11]通过边际坐标检验在无模型变量选择和 SDR 之间建立了重要联系。Li^[12]等人提出了基于网格化 χ^2 检验的无模型变量选择方法。然而,这些基于检验的无模型变量选择方法存在一些缺点,如计算复杂、精度低、不稳定等。另一类无模型变量选择方法是通过结合正则化方法和 SDR 方法提出的。例如,Ni 等^[13]在完全降维的情况下整合了 LASSO 收缩法,得到了一个逆回归估计器。Li 和 Yin^[14]引入了正则化来实现变量选择和降维估计。

此外, Bondell 和 Li^[15] 为整个逆回归估计器家族提出了一个通用的收缩估计策略, 允许同时进行降维和变量选择, 并证明了该新型估计器在没有任何常规模型的情况下实现了一致的变量选择。Chen 等^[16] 提出了一种坐标无关的稀疏估计方法, 该方法实现了稀疏的充分降维, 并有效筛选了不相关的冗余变量。然而, 这些方法均通过迭代程序完成, 效率较低, 且计算复杂性可能会影响其总体效率。

为了克服上述问题, 本文基于方向回归, 提出了一种高维非参数非线性系统的变量选择方法, 该方法不受维数灾难的影响。此外, 由于表征冗余变量的指标并不是精确的 0, 当指标较小时, 很难判断变量是否冗余。为解决这一问题, 本文提出了一种惩罚优化算法来保证集合的收敛性。

1 导数方法和维数灾难

每个变量 X_i 对函数 f 或输出 Y 的贡献可以通过偏导数 $\partial f / \partial X_i$ 来估计。如果 X_i 不贡献, 则 $\partial f / \partial X_i \equiv 0$ 。因此, 识别 X_i 是否有贡献的常用方法是利用局部线性估计方法^[17] 估计导数 $\partial f / \partial X_i$ 。在任何一点 $\mathbf{X}^0 = (X_1^0, \dots, X_p^0)^T$, 分别用 $\alpha = f(\mathbf{x}^0)$ 和 $\boldsymbol{\alpha}_1 = \left(\frac{\partial f}{\partial X_1}, \dots, \frac{\partial f}{\partial X_p} \right)^T \Big|_{\mathbf{x}^0}$ 表示 f 的值及其在 \mathbf{X}^0 处的导数 $\frac{\partial f}{\partial X_i}$ (其中 $i = 1, \dots, p$), 则二者的估计值可通过式(3)计算^[18]。

$$\min_{\dot{\alpha}_0, \dot{\alpha}_1} \sum_{k=1}^N (Y(k) - \{\dot{\alpha}_0 + \dot{\alpha}_1^T [\mathbf{X}(k) - \mathbf{X}^0]\}) K[\mathbf{X}(k) - \mathbf{X}^0], \quad (3)$$

其中 $K(\cdot)$ 是多变量核函数。局部线性估计方法已经被深入研究^[4, 17], 并得到了 $\dot{\alpha}_0 \rightarrow \alpha_0$ 、 $\dot{\alpha}_1 \rightarrow \boldsymbol{\alpha}_1$ 的收敛结果。该方法属于局部平均方法, 对非参数非线性系统具有重要意义。然而, 所有局部平均方法都会受到维数灾难的影响。为阐明这一问题, 以式(1)表示的系统为例, 设 $\mathbf{x}(k) = [u(k-1), \dots, u(k-p)]$, 假设输入为 $[-1, 1]$ 上均匀分布的 IID 随机序列, 希望利用输入输出数据估计函数 $f(\mathbf{x}^0)$ 在 $\mathbf{X}^0 = (0, 0, \dots, 0)^T$ 处的值。由于噪声和不确定性, 需要在 $\mathbf{X}^0 = (0, 0, \dots, 0)^T$ 周围有足够的测量值 $\mathbf{X}(k)$ 来可靠地评估 $\alpha_0 = f(\mathbf{x}^0)$ 和 $\boldsymbol{\alpha}_1 = \left(\frac{\partial f}{\partial X_1}, \dots, \frac{\partial f}{\partial X_p} \right)^T \Big|_{\mathbf{x}^0}$ 。简言之, 假设 $\mathbf{X}^0 = (0, 0, \dots, 0)^T$ 的邻域是一个以原点为中心, 半径为 0.1 的

球, 从而任意数据 $\mathbf{X}(k)$ 落在邻域的概率是 $\frac{\pi^{p/2} 0.1^p}{\Gamma(p/2+1) 2^p}$, 这里 Γ 表示 Γ 函数。假设在 \mathbf{X}^0 的邻域内至少需要 10 个数据以得到 $f(\mathbf{x}^0)$ 的可靠估计, 那么数据总量 N 需满足 $N \frac{\pi^{p/2} 0.1^p}{\Gamma(p/2+1) 2^p} \geq 10$, 即

$$N \geq 10 \times (20)^p \times p/2! \times (\pi^{p/2}) \\ = \begin{cases} 1.24 \times 10^8, & p=6 \\ 4.02 \times 10^{13}, & p=10 \end{cases}, \quad (4)$$

上式意味着即使对于适当大小的系统阶数 p , 为得到可靠的非参数估计, 所需数据总量依然巨大。

维数灾难是所有局部平均方法的一个重要问题, 并不是系统辨识领域所独有的。为了不受维数灾难的影响, 基于方向回归, 本文提出了一种新的变量选择方法。

2 基于方向回归的变量选择方法

2.1 方向回归

Li 和 Wang^[19] 提出的方向回归是一种有效的充分降维方法, 它结合了基于切片逆回归^[20, 21] 和切片平均方差估计^[22] 这两种降维估计方法的优点。方向回归方法如下。

变量 \mathbf{X} 的期望值和协方差表示为:

$$\mathbf{E}\mathbf{X} = (\mu_1, \dots, \mu_p)^T = \boldsymbol{\mu}, \\ \mathbf{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}. \quad (5)$$

定义 $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ 为标准化变量。 $(\tilde{\mathbf{Z}}, \tilde{Y})$ 与 (\mathbf{Z}, Y) 独立且同分布。在方向回归中, Li 和 Wang^[19] 定义了 $\mathbf{A}_0(Y, \tilde{Y}) = E[(\mathbf{Z} - \tilde{\mathbf{Z}})(\mathbf{Z} - \tilde{\mathbf{Z}})^T | Y, \tilde{Y}]$, 即把方向 $(\mathbf{Z} - \tilde{\mathbf{Z}})$ 回归到函数空间上。进一步, 定义了 $\mathbf{B}_0(Y, \tilde{Y}) = 2\mathbf{I}_p - \mathbf{A}_0(Y, \tilde{Y})$ 和 $\mathbf{G}_0 = E[\mathbf{B}_0(Y, \tilde{Y})^2]$ 。

根据 $S_{Y|\mathbf{Z}}$ 和 $S_{Y|\mathbf{X}}$ 不变原则^[23], 在 $E(Z | \boldsymbol{\beta}^T Z)$ 是 Z 的线性函数和 $\text{Var}(Z | \boldsymbol{\beta}^T Z)$ 是非随机的假设下, 有 $\boldsymbol{\Sigma}^{-1/2} \text{Span}(\mathbf{G}_0) \subseteq S_{Y|\mathbf{X}}$ 。当 \mathbf{X} 的分布是椭圆对称的, 如正态分布、 t 分布和拉普拉斯分布, 这些条件均得到满足。

因此, 定义

$$\mathbf{A}(Y, \tilde{Y}) = E[(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^T | Y, \tilde{Y}], \\ \mathbf{B}(Y, \tilde{Y}) = 2\boldsymbol{\Sigma} - \mathbf{A}(Y, \tilde{Y}), \quad (6)$$

其中, $(\tilde{\mathbf{X}}, \tilde{Y})$ 与 (\mathbf{X}, Y) 独立且同分布, 并定义

$$\mathbf{G} = E[\mathbf{B}(\mathbf{Y}, \tilde{\mathbf{Y}})^2].$$

将 $Y(k)$ 落在数据切片 J_k 上的概率记为 $p_k = P(Y \in J_k)$, 其中 $k = 1, \dots, H$, H 为切片的数量. 定义 $\mathbf{U}_k = E(\mathbf{X} | Y \in J_k)$, $\mathbf{V}_k = E(\mathbf{X}\mathbf{X}^T | Y \in J_k) - \Sigma$, 则矩阵 \mathbf{G} 可以表示为如下离散形式 \mathbf{F} ^[19]:

$$\begin{aligned} \mathbf{F} = & 2 \sum_{k=1}^H p_k \mathbf{V}_k^2 + 2 \left(\sum_{k=1}^H p_k \mathbf{U}_k \mathbf{U}_k^T \right)^2 + \\ & 2 \left(\sum_{k=1}^H p_k \mathbf{U}_k^T \mathbf{U}_k \right) \left(\sum_{k=1}^H p_k \mathbf{U}_k \mathbf{U}_k^T \right). \end{aligned} \quad (7)$$

定义 $\omega_1 \geq \omega_2 \geq \dots \geq \omega_K \geq \omega_{K+1} = \dots = \omega_p = 0$ 为 \mathbf{F} 的特征值, $\beta_1, \beta_2, \dots, \beta_p$ 为 \mathbf{F} 相应的特征向量.

2.2 基于方向回归的变量选择方法

假设 \mathbf{Y} 独立于 \mathbf{X}_{A^c} , 对于给定的 \mathbf{X}_{A^c} , 即

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X}_{A^c} \mid \mathbf{X}_A^* \quad (8)$$

其中, $\perp\!\!\!\perp$ 代表统计独立性, A^* 是贡献集合, A^c 是 A^* 的补集, 表示非贡献集合. 需要注意的是, 根据贡献变量的定义, 如果 X_j 是一个贡献变量, 即 $j \in A^*$, 那么 $\beta_{ij} \neq 0, 1 \leq i \leq K$; 如果 $j \in A^c$, 那么 $\beta_{ij} = 0, 1 \leq i \leq K$. 对于 $K+1 \leq i \leq p$, 特征值 ω_i 为 0.

根据特征向量 β_i 的性质, 定义新指标 δ :

$$\delta = \sum_{i=1}^p \omega_i |\beta_i|, \quad (9)$$

其中 $|\beta_i| = (|\beta_{i1}|, |\beta_{i2}|, \dots, |\beta_{ip}|)^T, 1 \leq i \leq p$. δ 可以用来检测贡献变量, 并剔除非贡献变量. δ_j 是新指标 δ 的第 j 个元素, 如果 $\delta_j = 0$, 则 X_j 不贡献; 如果 $\delta_j \neq 0$, 则 X_j 贡献, 而 $A^* = \{j : \delta_j \neq 0\}$.

将 $\hat{\mathbf{F}}$ 与 $\hat{\delta}$ 分别作为 \mathbf{F} 和 δ 的估计值, 定义 $\hat{\omega}_1, \dots, \hat{\omega}_p$ 为 $\hat{\mathbf{F}}$ 的特征值, $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ 为相应的特征向量. 在实际情况下, 由于噪声干扰和数据长度有限, 冗余变量的 $\hat{\delta}_j$ 可能是接近于 0 的数值. 由于 $\hat{\delta}_j$ 并不精确为 0, 当其值较小时, 很难判断对应变量是否冗余. 为解决这一问题, 本文提出了一种惩罚优化算法来保证集合的收敛性. 定义

$$c_i(N) = \hat{\delta}_i, \quad c_{0i} = \delta_i, \quad i = 1, \dots, p, \quad (10)$$

其中, N 清楚地表明, 评估值是从数据 $k = 1, \dots, N$ 获得的. Li 和 Wang^[19] 已根据 Lindeberg-Levy 中心极限定理证明, 在正则条件下,

$$\hat{\mathbf{F}} = \mathbf{F} + O_p(N^{-1/2}), \quad (11)$$

其中, $O_p(N^{-1/2})$ 代表随机变量在概率上与 $N^{-1/2}$ 处于同一数量级. 因此,

$$c_i(N) - c_{0i} = O_p(N^{-1/2}), \quad (12)$$

考虑如下优化问题:

$$\begin{aligned} \min_{\eta_i} J(\eta_i, N) = & \min_{\eta_i} \{ [\eta_i - c_i(N)]^2 + \lambda(N) |\eta_i| \} \\ \text{s. t. } & 0 < \lambda(N) \rightarrow 0, \\ & \frac{1}{\sqrt{N} \lambda(N)} \rightarrow 0 \text{ as } N \rightarrow \infty, \end{aligned} \quad (13)$$

其中 $\lambda(N)$ 是正则化参数. 这是一个容易解决的一维凸优化问题.

基于方向回归的变量选择方法的算法, 可以总结为如下步骤:

步骤 1 通过 $[Y_{(k)}, \mathbf{X}_{(k)}]$ 对数据集 $[Y(k), \mathbf{X}(k)]$, $k = 1, \dots, N$ 重新排序, 其中 $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(N)}$. 设 H 为切片数, m 为每个片段的数据点数, $N = Hm$. 将数据集 $Y_{(1)}, \dots, Y_{(N)}$ 均匀地划分为 H 个片段.

步骤 2 计算 $\hat{\mathbf{U}}_k, \hat{\mathbf{V}}_k (k = 1, \dots, H)$, 和

$$\begin{aligned} \hat{\mathbf{F}} = & 2 \sum_{k=1}^H p_k \hat{\mathbf{V}}_k^2 + 2 \left(\sum_{k=1}^H p_k \hat{\mathbf{U}}_k \hat{\mathbf{U}}_k^T \right)^2 + \\ & 2 \left(\sum_{k=1}^H p_k \hat{\mathbf{U}}_k^T \hat{\mathbf{U}}_k \right) \left(\sum_{k=1}^H p_k \hat{\mathbf{U}}_k \hat{\mathbf{U}}_k^T \right). \end{aligned} \quad (14)$$

步骤 3 对 $\hat{\mathbf{F}}$ 进行谱分解, 得到特征值 $\hat{\omega}_1 \geq \hat{\omega}_2 \geq \dots \geq \hat{\omega}_p \geq 0$ 和相应的特征向量 $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$.

$$\text{步骤 4} \quad \text{计算指标 } \delta = \sum_{i=1}^p \hat{\omega}_i |\hat{\beta}_i|.$$

步骤 5 求解最小化问题

$$\min_{\eta_i} \{ (\eta_i - \hat{\delta}_i)^2 + \lambda(N) |\eta_i| \}, i = 1, 2, \dots, p,$$

式中, 随着 $N \rightarrow \infty, \lambda(N) \rightarrow 0, O(N^{-1/2})/\lambda(N) \rightarrow 0$. 将上述最小化解记为 $\eta_i^*(N)$.

步骤 6 如果 $\eta_i^*(N) = 0, X_i$ 没有贡献; 如果 $\eta_i^*(N) \neq 0, X_i$ 有贡献.

3 数值仿真

在许多工程实际中, 贡献变量的数量相较于系统中可能存在的变量数量而言, 通常较少. 通过采用变量选择方法识别贡献变量和非贡献变量, 可以剔除冗余变量, 降低系统维度. 这不仅减少了系统对噪声的敏感度, 还有效地避免了维数灾难的影响.

例 1: 考虑一个 8 维非线性系统

$$\begin{aligned} Y(k) = & f[X_1(k), X_2(k), \dots, X_8(k)] + v(k) \\ = & \alpha_1 \sin[X_1(k)X_2(k)] + \alpha_2 [X_3(k) - 0.5]^2 + \\ & \alpha_3 X_4(k) + \alpha_4 X_5(k) + \alpha_5 X_6(k)X_7(k) + \\ & \alpha_6 X_7(k)^2 + \alpha_7 \cos[X_6(k)X_8(k)] + \end{aligned}$$

$$\alpha_8 e^{-|X_8(k)|} + v(k), \quad k=1,2,\dots,N, \quad (15)$$

这里 $\alpha_1 = \alpha_3 = \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = 0$, $\alpha_2 = 0.5$, $\alpha_4 = 1$.

该系统函数形式 f 是未知的。 $X_i(k)$ 是均值为 0、方差为 1 的 IID 高斯随机序列, 噪声也是均值为 0、标准差为 0.5 的 IID 高斯随机序列。设 $H=10$, $N=4000$, $\lambda(N)=2 \times N^{-2/5}$ 。对本文提出的变量选择方法进行 100 次蒙特卡罗仿真, 结果如表 1 所示。结果表明, 所提方法每次都正确选取了贡献变量 $X_3(k)$ 和 $X_5(k)$, 从未识别非贡献变量。表 2 列出了中间变量 $\hat{\delta}_i$ 和 $\eta_i^*(N)$ 的值, 表明所提算法第五步可有效解决多小是足够小的问题。表 1 和表 2 的结果表明所提出的惩罚优化算法可以保证贡献集合收敛, 即可利用指标 $\eta_i^*(N)$ 是否为 0 消除冗余变量。

表 1 在 100 次蒙特卡洛中选择 X_i 的次数

Table 1 Number of X_i was selected in 100 Monte Carlo runs

变量	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
选中次数	0	0	100	0	100	0	0	0

表 2 基于 DR 变量选择方法下 $\hat{\delta}_i$ 和 $\eta_i^*(N)$ 的值

Table 2 Values of $\hat{\delta}_i$ and $\eta_i^*(N)$ using DR-based variable selection approach

i	1	2	3	4	5	6	7	8
$\hat{\delta}_i$	0.07	0.08	3.32	0.10	2.61	0.08	0.10	0.09
$\eta_i^*(N)$	0	0	2.96	0	2.25	0	0	0

本文进一步研究了样本量对挑选出正确变量准确性的影响。图 1 显示了当噪声是一个均值为 0、标准差为 0.1 的 IID 高斯随机序列时, 变量被正确选择的正确率与总数据长度 N 的关系。计算结果

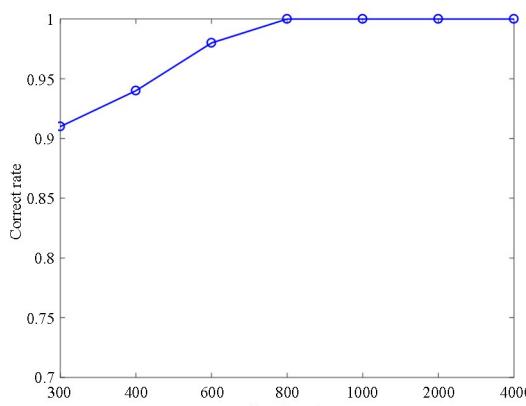


图 1 100 次蒙特卡洛下选择正确贡献变量的正确率

Fig. 1 The rates of selecting the correct contributing variables in 100 Monte Carlo runs

表明, 如果数据长度 N 大于 800, 100 次蒙特卡洛仿真结果显示识别正确变量的正确率趋于 1。

最后, 将本文所提变量选择方法和互信息法^[24]进行比较, 后者是最流行的变量选择方法之一。表 3 显示了每个输入变量 X_i 和输出 Y 之间的互信息。结果显示, 贡献变量的互信息大于非贡献变量。然而, 尽管非贡献变量的互信息很小, 但并不完全为 0。因此, 如何根据微小互信息判断变量是否贡献成为亟待解决的问题。常规的方法是应用统计假设检验, 然而假设检验的计算成本很高, 且对于高维系统并不可靠。此外, 互信息的计算也很耗时, 例如, 数据长度 $N=4000$ 时的计算时间约为 4480 s, 数据长度 $N=1000$ 时的计算时间约为 291 s。与此相反, 本文所提基于方向回归的变量选择方法是对 Y 的一维逆回归, 计算简单且高效。例如, 在数据长度 $N=4000$ 时, 本文提出的变量选择方法的计算时间约为 0.74 s, 数据长度 $N=1000$ 时的计算时间约为 0.69 s。

表 3 各输入变量和输出变量之间的互信息

Table 3 Mutual information between each input variable and output variable

$MI(X_i, Y)$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
# 互信息	0.047	0.046	0.529	0.039	0.312	0.042	0.039	0.038

为进一步说明变量选择对系统辨识有帮助, 我们用最后 50 个数据作为验证数据。对于每个验证数据 $\mathbf{X}(k)=[X_1(k), \dots, X_8(k)]^\top$, 如前所述, 从第一个训练数据 $N=3950$ 计算出两个估计输出 $\hat{Y}(k)=f[\mathbf{X}(k)]$, $k=1, \dots, 50$ 。前者是基于将例 1 直接作为一个 8 维系统进行辨识。第二个是通过使用提出的变量选择方法对二维未知系统 $g(\cdot)$ 的辨识

$$Y(k) = g[X_3(k), X_5(k)], \quad (16)$$

在本文中, 我们使用式(3)描述的局部线性估计方法来辨识非参数非线性系统。

为了能够比较这两个估计方法, 计算两种方法的拟合优度(goodness of fits, GoF):

$$G = \left\{ 1 - \sqrt{\frac{\sum [Y(k) - \hat{Y}(k)]^2}{\sum [Y(k) - \frac{1}{50} \sum Y(k)]^2}} \right\} \times 100\% \quad (17)$$

其中, G 表示拟合优度, $Y(k)$ 表示测量的输出,

$\hat{Y}(k)$ 表示预测的输出. 图2显示了实际结果(实线)与通过直接辨识 $G=0.63$ 的8维系统(点线)和辨识 $G=0.94$ 的二维系统(虚线)的估计结果. 显然,如果能够首先检测并去除非贡献变量,那么效率将得到大幅提高.

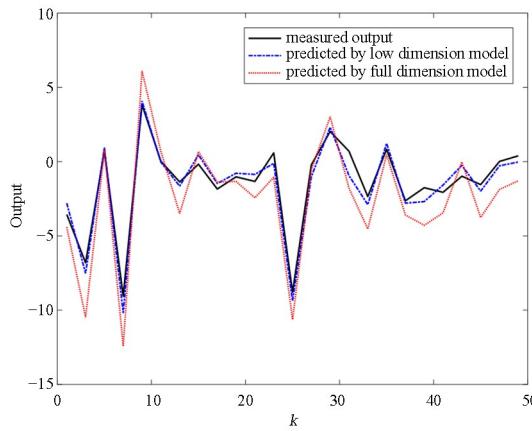


图2 全维识别(点线)和低维识别(虚线)模型的估计结果和实际(实线)测量结果

Fig.2 Predicted results for the full-dimensional identification (dotted line) and low-dimensional identification (dashed line) models and actual (solid line) measurements

4 结论

本文提出了一种新的高维非参数非线性系统变量选择算法,该算法不受维数灾难的影响,解决了传统导数方法面临的一个主要难题.此外,该变量选择方法的数值算法相当简单明了,而且具有集合收敛性.最后,由于方向回归可以导出足够的降维空间,因此基于方向回归的变量选择方法可以从候选变量中充分检测出所有贡献变量,故而本文所提方法为变量选择提供了一个充分必要条件.

参考文献

- [1] NEMETH J G, KOLLER I, SCHOUKENS J. Identification of Volterra kernels using interpolation [J]. IEEE Transactions on Instrumentation and Measurement, 2002, 51(4): 770–775.
- [2] CHENG C M, PENG Z K, ZHANG W M, et al. Volterra-series-based nonlinear system modeling and its engineering applications: a state-of-the-art review [J]. Mechanical Systems and Signal Processing, 2017, 87: 340–364.
- [3] LJUNG L. System identification : Theory for the user [M]. 2nd ed. New York: Prentice-Hall, 1999.
- [4] BAI E W, CHAN K S. Identification of an additive nonlinear system and its applications in generalized Hammerstein models [J]. Automatica, 2008, 44(2): 430–436.
- [5] TEMPO R, CALAFIORE G, DABBENE F. Randomized algorithms for analysis and control of uncertain [M]. London: Springer-Verlag, 2005.
- [6] TIBSHIRANI R. Regression shrinkage and selection via the LASSO [J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267–288.
- [7] ZOU H. The adaptive LASSO and its oracle properties [J]. Journal of the American Statistical Association, 2006, 101(476): 1418–1429.
- [8] YUAN M, LIN Y. Model selection and estimation in regression with grouped variables [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49–67.
- [9] EFRON B, HASTIE T, JOHNSTONE I, et al. Least angle regression [J]. The Annals of Statistics, 2004, 32(2): 407–499.
- [10] ZHAO W X, CHEN H F, BAI E W, et al. Kernel-based local order estimation of nonlinear nonparametric systems [J]. Automatica, 2015, 51: 243–254.
- [11] COOK R D. Testing predictor contributions in sufficient dimension reduction [J]. The Annals of Statistics, 2004, 32(3): 1062–1092.
- [12] LI L X, COOK R D, NACHTSHEIM C J. Model-free variable selection [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 285–299.
- [13] NI L Q, COOK R D, TSAI C L. A note on shrinkage sliced inverse regression [J]. Biometrika, 2005, 92(1): 242–247.
- [14] LI L X, YIN X R. Sliced inverse regression with regularizations [J]. Biometrics, 2008, 64(1): 124–131.
- [15] BONDELL H D, LI L X. Shrinkage inverse regression estimation for model-free variable selection [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2009, 71(1): 287–299.
- [16] CHEN X, ZOU C L, COOK R D. Coordinate-independent sparse sufficient dimension reduction and variable selection [J]. The Annals of Statistics, 2010, 38(6): 3696–3723.

- [17] BAI E W. Non-parametric nonlinear system identification: an asymptotic minimum mean squared error estimator [J]. IEEE Transactions on Automatic Control, 2010, 55(7): 1615—1626.
- [18] CHAN S C, ZHANG Z G. Local polynomial modeling and variable bandwidth selection for time-varying linear systems [J]. IEEE Transactions on Instrumentation and Measurement, 2011, 60(3): 1102—1117.
- [19] LI B, WANG S Q. On directional regression for dimension reduction [J]. Journal of the American Statistical Association, 2007, 102(479): 997—1008.
- [20] LI K C. Sliced inverse regression for dimension reduction [J]. Journal of the American Statistical Association, 1991, 86(414): 316—327.
- [21] COOK R D, WEISBERG S. Discussion of “Sliced inverse regression for dimension reduction” by K.-C. Li, [J]. Journal of the American Statistical Association, 1991, 86(414): 328—332.
- [22] COOK R D. On the interpretation of regression plots [J]. Journal of the American Statistical Association, 1994, 89(425): 177—189.
- [23] PENG H C, LONG F H, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and Min-redundancy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226—1238.
- [24] DOWNS J J, VOGEL E F. A plant-wide industrial process control problem [J]. Computers & Chemical Engineering, 1993, 17(3): 245—255.