中国科学 : 数学　2016 年　第 46 卷　第 10 期 : 1383 ∼ 1400
SCIENTIA SINICA　Mathematica

《中国科学》杂志社
SCIENCE CHINA PRESS

非参数非线性系统的变量选择及研究进展

献给陈翰馥教授 80 华诞

白尔维[①*], 李慷[②], 赵文虎[③], 牟必强[③], 郑卫新[④]

① Department of Electrical and Computer Engineering, University of Iowa, Iowa City 52242, USA;
② School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BTT INN, UK;
③ 中国科学院数学与系统科学研究院系统控制重点实验室, 北京 100190;
④ School of Computing, Engineering and Mathematics, Western Sydney University, Sydney 2751, Australia
E-mail: er-wei-bai@uiowa.edu, k.li@qub.ac.uk, wxzhao@amss.ac.cn, bqmu@amss.ac.cn, w.zheng@westernsydney.edu.au

收稿日期: 2016-01-01; 接受日期: 2016-02-26; 网络出版日期: 2016-09-30; * 通信作者
国家自然科学基金 (批准号: 61573345)、国家重点基础研究发展计划 (批准号: 2014CB845301 和 2016YFB0901902) 和中国科学院数学院院长基金 (批准号: 2015-hwyxqnrc-mbq) 资助项目

摘要　本文考虑非参数非线性系统的变量选择问题, 所谓变量选择是指, 判断哪些变量对系统的输出有实质性影响并构造算法将它们辨识出来. 由于 "维数灾难", 非线性系统随着维数增加而使得辨识需要的数据量呈指数速度上升, 因而有必要从数据量趋于无穷、数据量有限和系统结构等不同角度来研究非参数非线性系统的变量选择问题. 针对上述不同情形, 本文分别给出变量选择算法和收敛性结论, 并给出数值例子以验证理论结果的有效性, 结果显示数值模拟与理论分析一致.

关键词　变量选择　非参数非线性系统　可加非线性系统　局部变量　全局变量
MSC (2010) 主题分类　93B30, 93E12, 93C10

1　引言

考虑离散时间标量非参数非线性系统

$$y(k) = f(x(k)) + v(k), \quad k = 1, 2, \ldots, N, \tag{1.1}$$

其中 $\{y(k)\}_{k\geqslant 1}$ 为系统输出, $\{v(k)\}_{k\geqslant 1}$ 为系统噪音, 本文假设 $\{v(k)\}_{k\geqslant 1}$ 为独立同分布 (i.i.d.) 随机序列, 回归向量记为 $x(k) = [x_1(k), \ldots, x_p(k)]^{\mathrm{T}}$, 它的维数为 p, 回归向量 $x(k)$ 包含了所有可能起作用的变量, $f(\cdot)$ 为未知函数. 系统 (1.1) 的辨识就是利用数据集 $\{y(k), x(k)\}_{k=1}^{N}$ 来估计函数 $f(\cdot)$ 和回归向量 $x(k)$ 中起作用的变量 (参见文献 [1,2]).

因为函数 $f(\cdot)$ 未知, 一类具有代表性的辨识方法是将 $f(\cdot)$ 参数化, 例如, 将 $f(\cdot)$ 表示为基函数的组合 $f(x) = \sum \alpha_i \phi_i(x)$, 或者更一般地, $f(x) = g(\alpha, \phi_1(x), \phi_2(x), \ldots)$, 其中 α 和 α_i 为未知参数, 函数 $g(\cdot)$ 和 $\phi_i(x)$ 均为已知, 从而辨识转化为给定准则函数条件下对参数 α 和 α_i 的优化. 注意到 $f(\cdot)$ 未

引用格式: Bai E W, Li K, Zhao W X, et al. Variable selection in nonlinear nonparametric system identification (in Chinese). Sci Sin Math, 2016, 46: 1383–1400, doi: 10.1360/N012016-00002

　

知, 对函数作如上假设通常需要获得较多的系统先验信息; 另一方面, 由于非线性因素的存在, 相应优化问题的求解可能陷于局部极小值点而无法获得渐近一致的参数估计. 上面这些因素, 使参数化辨识的应用范围受到局限. 另一类辨识方法称为非参数方法 (参见文献 [3–9]), 例如, 直接估计函数 $f(\cdot)$ 在给定点 x^0 的函数值 $f(x^0)$, 注意到 $f(\cdot)$ 的非线性, 这类算法通常基于系统在 x^0 附近数据的加权平均, 因而容易受到所谓 "维数灾难" 问题的影响. 请看下面例子: 假设系统 (1.1) 为有限脉冲响应的非线性系统, 系统的阶数为 p, 从而, $x(k)=[u(k-1),\ldots,u(k-p)]^{\mathrm{T}}$, 取输入信号 $\{u(k)\}_{k\geqslant 1}$ 为 $[-1,1]$ 上均匀分布的独立同分布随机序列, 希望利用输入输出数据 $\{u(k),y(k)\}_{k\geqslant 1}$ 估计函数 $f(\cdot)$ 在点 $x^0=[0,\ldots,0]^{\mathrm{T}}$ 的函数值. 选取以 x^0 为球心、以 0.1 为半径的球作为 x^0 的邻域, 从而任意数据 $x(k)$ 落于此邻域的概率为 $\frac{\pi^{p/2}0.1^p}{\Gamma(p/2+1)}\frac{1}{2^p}$, 其中 $\Gamma(\cdot)$ 为 Γ- 函数. 假设在 x^0 的邻域内至少需要 10 组数据以得到 $f(x^0)$ 的可靠估计, 那么数据总量 N 需满足 $N\frac{\pi^{p/2}0.1^p}{\Gamma(p/2+1)}\frac{1}{2^p}\geqslant 10$, 即

$$N\geqslant\frac{10\cdot(20)^p\cdot(p/2)!}{\pi^{p/2}}=\begin{cases}1.24\cdot 10^8, & p=6,\\ 4.02\cdot 10^{13}, & p=10.\end{cases}$$

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, 所需数据总量依然十分巨大.

对一些实际系统而言, 回归向量 $x(k)$ 中的各个变量是 "稀疏" 的, 即是说并非所有 $x_i(k)$ $(i=1,2,\ldots,p)$ 仅是其中一部分对系统的输出 $y(k)$ 起作用. 如果能够辨识并消除对于系统输出 $y(k)$ 不起作用的那些变量, 就能实现系统的降维, 并显著减少辨识所需的数据总量. 在文献中, 这个问题常称为变量选择[10]. 总的来讲, 变量选择算法可分为两类: 前一类算法中, 假若某些变量对系统实际输出的影响为零, 那么这些变量被舍去, 其他的变量被保留, 从而实现系统的变量选择; 后一类算法中, 不仅舍去对系统输出影响为零的那些变量, 对系统输出影响很小的那些变量也将被舍去, 其他的变量被保留. 二者相比较, 后一类算法可以得到更简化的系统模型, 便于进一步分析与设计, 得到的结果更具稳健性. 显然, 第二类算法的首要问题是如何定义哪些变量起作用、哪些变量不起作用, 进一步如何辨识. 这些问题后文将作详细分析.

前文假设回归向量 $x(k)=[x_1(k),\ldots,x_p(k)]^{\mathrm{T}}$ 的维数为 p, $x_i(k)$ $(i=1,\ldots,p)$ 中包含了起作用的变量和冗余变量. 依对 $x(k)$ 中各个变量的分析角度, 变量选择又可分为自上而下的算法和自下而上的算法两大类: 前一类着眼于 p 维系统, 构造算法判定向量 $x(k)$ 中的各个变量是否起作用以及对系统的贡献大小; 而后一类算法首先致力于一维系统, 即判定 $x_i(k)$ $(i=1,\ldots,p)$ 中对系统贡献最大的变量, 而后判定次要变量, 以此类推, 逐个判定变量的重要性以及是否冗余, 从而实现变量选择. 二者相比较, 前一类直接基于高维系统构造算法, 可借鉴非参数统计的思路, 便于理论分析与推导, 另一方面由于非参数辨识的维数灾难, 为保证算法的收敛性常需要较大的数据总量, 对于数据量有限的实际问题, 这类算法难以奏效; 后一类基于低维系统构造算法, 对于数据量有限的实际问题, 可有效地避开维数灾难, 但与前一类算法相比, 理论分析是其难点. 本文依据这个思路, 针对系统 (1.1) 分别给出了两类变量选择算法, 并给出理论结果与仿真验证.

变量选择在系统控制、信号处理、统计和机器学习等领域扮演着重要角色 (参见文献 [11–16]), 与之密切相关的一个问题是阶的估计. 动态系统的阶估计已有很多研究成果 (参见文献 [17–21]). 变量选择不仅要估计系统的真实阶次 (p_0,q_0), 还要判定回归向量中真正起作用的变量, 目前, 相关文献不断涌现, 如文献 [22–27], 其中既有针对线性系统的变量选择[28,29], 也有针对非线性系统的相关研究[30–34]. 上述研究工作均假设起作用的变量在系统的工作区域中是全局的、唯一的, 对线性系统来讲, 这是显而易见的事实, 但对非线性系统来讲却可能存在反例. 考虑如下有限脉冲响应非线性系统:

$$y_{k+1}=f(u_k,u_{k-1},u_{k-2},u_{k-3})+\varepsilon_{k+1}, \tag{1.2}$$

$$f(u_k,\ldots,u_{k-3})=\begin{cases}u_{k-3}, & \text{若 } u_k\geqslant 0,\\ u_{k-3}u_{k-1}, & \text{若 } u_k<0, \quad u_{k-1}>1,\\ u_{k-3}u_{k-2}, & \text{若 } u_k<0, \quad u_{k-1}<-1,\\ u_k, & \text{其他.}\end{cases} \tag{1.3}$$

从 (1.3) 可见, 系统在不同区域内有着不同的工作变量, 且没有一个变量在所有区域中都对系统输出产生实质作用. 从宏观上看, $u_k,u_{k-1},u_{k-2},u_{k-3}$ 这四个变量都不可少, 系统是 "稠密的", 但从局部来看, 仅是其中一部分发生作用, 系统又是 "稀疏的". 因而, 契合非线性系统 "全局稠密" 和 "局部稀疏" 的特点, 为得到更精准的数学模型, 有必要从局部的角度来研究它的变量选择 (参见文献 [35]). 目前, 基于这种思想的非参数非线性系统的变量选择还未见到相关研究文献, 本文将介绍我们在这个方向上近期的研究成果.

从方法上看, 直接借鉴线性系统的变量选择算法来处理非线性系统的相关问题可能难以得到有效的辨识结果, 请看下例: 相关系数算法是线性系统变量选择的一类有效算法 (参见文献 [36,37]), 考察线性系统 $y(k)=u(k-1)$, 选取输入信号 $\{u(k)\}_{k\geqslant 1}$ 为 $[-1,1]$ 上均匀分布的独立同分布随机序列, 容易验证输入 $u(k-1)$ 和输出 $y(k)$ 的相关系数为 1, 这意味着系统输出 $y(k)$ 线性地依赖于系统输入 $u(k-1)$; 而对非线性系统 $y(k)=u(k-1)^2$, 在相同的输入条件下可得输入 $u(k-1)$ 和输出 $y(k)$ 的相关系数为 0, 若以此断言输入 $u(k-1)$ 对系统输出 $y(k)$ 没有影响就忽视了系统的本质特性, 得不到正确结论. 综上所述, 由于非线性带来的 "维数问题"、"全局稠密、局部稀疏" 等特点, 因此需要结合系统的具体特点来研究变量选择. 本文围绕上述问题, 从全局变量的辨识、局部变量的辨识和可加非线性系统的变量选择等不同角度, 介绍我们近期的相关研究成果, 其中第 3.3 小节的内容不见于其他文献.

本文具体组织结构如下: 第 2 和 3 节主要针对系统 (1.1), 介绍局部变量选择算法和全局变量选择算法; 第 4 节针对 (1.1) 的特例—可加非线性系统, 介绍相关算法和主要结论; 第 5 节针对第 3 节的算法给出数值模拟以验证算法的有效性; 最后, 第 6 节作总结与讨论.

本文常用数学符号: 记 $(\Omega,\mathcal{F},\mathrm{P})$ 为基本概率空间, $\omega\in\Omega$ 为随机事件. $\mathbb{R}$ 表示实数域, 集 A 的余集记为 A^c, 矩阵 M 的 2- 范数记为 $\|M\|$. 对正序列 $\{a_k\}$ 和 $\{b_k\}$, $a_k=O(b_k)$ 和 $a_k=o(b_k)$ 分别意味着 $|\frac{a_k}{b_k}|\leqslant C$ 和 $|\frac{a_k}{b_k}|\to 0$, $k\to\infty$, 其中常数 $C>0$. 对随机序列 $\{x_k\}$, $x_k=O_P(1)$ 是指存在正数 C 使得 $\mathrm{P}\{|x_k|\geqslant C\}\to 0$, $k\to\infty$.

2 局部变量的选择算法

任意固定 $x^0=[x_1^0,x_2^0,\ldots,x_p^0]^{\mathrm{T}}\in\mathbb{R}^p$, 本节考虑系统 (1.1) 在点 x^0 及其邻域范围的变量选择. 首要问题是, 如何定义点 x^0 及其邻域范围的作用变量和冗余变量.

假设 $f(\cdot)$ 在 x^0 连续可微, 依 Taylor 级数展开, 易知在 x^0 附近各个变量 $x_i(k)$ $(i=1,\ldots,p)$ 的重要性可用 $f(\cdot)$ 各个分量的偏导数 $\frac{\partial f}{\partial x_i}|_{x=x^0}$ 来衡量: 假若 $\frac{\partial f}{\partial x_i}|_{x=x^0}\neq 0$, 则认为第 i 个变量在 x^0 附近起作用; 假若 $\frac{\partial f}{\partial x_i}|_{x=x^0}=0$, 则认为第 i 个变量在 x^0 附近不起作用. 基于此, 系统在 x^0 及其邻域的变量选择问题关键在于, 首先构造系统在 x^0 及其邻域的局部线性模型, 进而判断线性模型中的系数哪些非零、哪些为零, 非零元对应作用变量, 而零元则对应冗余变量, 变量的重要性程度则用偏导数的绝对值来刻画.

根据上述思路, 算法的关键在于建立函数 $f(\cdot)$ 在给定点 x^0 的局部线性模型. 首先介绍局部线性模型的辨识算法. 在给定点 x^0 处, 当 $\|x(k)-x^0\|\leqslant h$ 时, 依 Taylor 级数展开有下式成立:

$$f(x(k))=f(x^0)+(x(k)-x^0)^{\mathrm{T}}\frac{\partial f}{\partial x}\bigg|_{x^0}+O(h^2),$$

其中常数 $h>0$ 决定了点 x^0 的邻域大小, 称为带宽常数 (bandwidth).

记函数 $f(\cdot)$ 在给定点的函数值和梯度如下:

$$[f(x^0),\beta^{*\mathrm{T}}]^{\mathrm{T}}\triangleq[f(x^0),\beta_1^*,\ldots,\beta_p^*]^{\mathrm{T}}=\left[f(x^0),\frac{\partial f}{\partial x_1}\bigg|_{x=x^0},\ldots,\frac{\partial f}{\partial x_p}\bigg|_{x=x^0}\right]^{\mathrm{T}}.$$

考虑如下算法:

$$\min_{\gamma_0\in\mathbb{R},\gamma_1\in\mathbb{R}^p}\sum_{k=1}^{N}\left\{y(k)-[1,(x(k)-x^0)^{\mathrm{T}}]\begin{bmatrix}\gamma_0\\ \gamma_1\end{bmatrix}\right\}^2\cdot K_Q(x(k)-x^0),\tag{2.1}$$

其中 $K_Q(x)=\frac{1}{|Q|}K(Q^{-1}x)$, 矩阵 $Q=hI$, $h>0$ 为带宽常数, I 为单位阵, $K(\cdot)$ 为多变量核函数, 通常可取为任意概率密度函数, 本文选取核函数 $K(\cdot)$ 满足 $K_Q(x(k)-x^0)=0,\forall\,\|x(k)-x^0\|>h$.

从核函数的构造可见, 对数据集 $\{x(1),\ldots,x(N)\}$, 假若某 $x(k)$ 落于 x^0 的邻域之内, 则核函数为正, 否则为零. 从而, 算法 (2.1) 实为利用 x^0 近旁数据的局部最小二乘算法, 相应地, 其极小值点可作为系统 (1.1) 在给定点 x^0 的局部线性模型的合理估计. 记算法 (2.1) 的极小值点为

$$\begin{bmatrix}\hat f(x^0)\\ \hat\beta\end{bmatrix}=\begin{bmatrix}\hat f(x^0)\\ \hat\beta_1\\ \vdots\\ \hat\beta_p\end{bmatrix}=\begin{bmatrix}\hat f(x^0)\\ \frac{\partial\hat f}{\partial x_1}|_{x=x^0}\\ \vdots\\ \frac{\partial\hat f}{\partial x_p}|_{x=x^0}\end{bmatrix}.\tag{2.2}$$

根据核函数的构造, 假若 $\|x(k)-x^0\|>h$, 则有 $K_Q(x(k)-x^0)=0$. 给定数据长度 N 和带宽参数 h, 假设落于点 x^0 邻域内的数据集合为 $\{x(N_i),i=1,\ldots,M\}$, 相应地, 系统输出和噪音的集合记为 $\{y(N_i),i=1,\ldots,M\}$ 和 $\{v(N_i),i=1,\ldots,M\}$. 易见, M 依赖于数据长度 N 和带宽参数 h, 即 $M=M(N,h)$. 根据 (2.2), $\hat f(x^0)$ 为函数值 $f(x^0)$ 的估计, 定义

$$Z_M=\begin{bmatrix}y(N_1)-f(\hat{x}^0)\\ y(N_2)-f(\hat{x}^0)\\ \vdots\\ y(N_M)-f(\hat{x}^0)\end{bmatrix},\quad \Phi_M=\begin{bmatrix}(x(N_1)-x^0)^{\mathrm{T}}\\ (x(N_2)-x^0)^{\mathrm{T}}\\ \vdots\\ (x(N_M)-x^0)^{\mathrm{T}}\end{bmatrix},\quad v_M=\begin{bmatrix}v(N_1)\\ v(N_2)\\ \vdots\\ v(N_M)\end{bmatrix}.\tag{2.3}$$

由局部最小二乘算法的性质 (参见文献 [38]), 可得

$$Z_M=\Phi_M\beta^*+v_M+O(h^2)+O\left(\frac{1}{\sqrt{Nh^{p+l}}}\right).\tag{2.4}$$

注意到系统 (1.1) 为 p 维系统, 假设系统 (1.1) 在点 x^0 有 q ($1\leqslant q\leqslant p$) 个作用变量, $p-q$ 个冗余变量, 即 β^* 中有 q 个分量非零, $p-q$ 个分量为零. 不失一般性, 假设

$$\beta^*=[\beta_1^*,\ldots,\beta_q^*,\beta_{q+1}^*,\ldots,\beta_p^*]^{\mathrm{T}},\quad |\beta_1^*|>0,\ldots,|\beta_q^*|>0,\quad \beta_{q+1}^*=\cdots=\beta_p^*=0.\tag{2.5}$$

定义集合

$$A^* = \{j : |\beta_j^*| > 0\} = \{1, 2, 3, \ldots, q\}. \tag{2.6}$$

显然, 集合 A^* 是系统 (1.1) 在 x^0 处作用变量的指标集, 系统 (1.1) 在 x^0 处的变量选择一方面要正确地判断哪些分量非零, 即估计集合 A^*, 称为集收敛 (set convergence), 还要得到 β_j^* $(j = 1, \ldots, q)$ 的一致参数估计, 称为参数收敛 (parameter convergence). 构造变量选择算法如下:

依算法 (2.1) 所得估计值 $[f(\hat{x}^0), \hat{\beta}^{\mathrm{T}}]^{\mathrm{T}} = [f(\hat{x}^0), \hat{\beta}_1, \ldots, \hat{\beta}_p]^{\mathrm{T}}$, 定义 $w_j = \frac{1}{|\hat{\beta}_j|}$, $w = [w_1, w_2, \ldots, w_p]^{\mathrm{T}}$, 以及如下准则函数:

$$\overline{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \|Z_M - \Phi_M \beta\|^2 + \lambda_M \sum_{j=1}^{p} w_j |\beta_j| \right\}, \tag{2.7}$$

其中 Z_M 和 Φ_M 由 (2.3) 定义, λ_M 为加权系数.

从 (2.7) 的构造可见, 假若算法 (2.1) 所得的某个估计值 $\hat{\beta}_j \to 0$, 对应地, $w_j \to \infty$, 则算法 (2.7) 对应的极小值点必然为零, 可判定第 j 个分量为冗余变量, 从而实现了变量选择. 基于此, 定义集合

$$A_N = \{j : |\overline{\beta}_j| > 0\}, \tag{2.8}$$

其中 $\overline{\beta} = [\overline{\beta}_1, \ldots, \overline{\beta}_p]^{\mathrm{T}}$ 是优化准则函数 (2.7) 得到的极小值点, A_N 中的指标即对应系统 (1.1) 在点 x^0 处的作用变量.

算法 2.1 考虑系统 (1.1), 给定点 x^0 和数据集 $\{y(k), x(k)\}_{k=1}^N$, 固定带宽 h,

(II.1) 计算局部最小二乘算法的极小值点 $[f(\hat{x}^0), \hat{\beta}^{\mathrm{T}}]^{\mathrm{T}} = [f(\hat{x}^0), \hat{\beta}_1, \ldots, \hat{\beta}_p]^{\mathrm{T}}$;

(II.2) 依带宽 h, 计算点 x^0 邻域内的输入数据集, 其容量记为 M, 依 (2.3) 构造 Z_M 和 Φ_M;

(II.3) 计算 $w_j = \frac{1}{|\hat{\beta}_j|}$, $j = 1, 2, \ldots, p$, 计算准则函数 (2.7) 的极小值点 $\overline{\beta} = [\overline{\beta}_1, \ldots, \overline{\beta}_p]^{\mathrm{T}}$, 并依 (2.8) 得到集合 A_N 作为集合 A^* 的估计;

(II.4) 依 (II.1) 和 (II.3) 所得估计, 定义 $\tilde{\beta} = [\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_p]^{\mathrm{T}}$:

$$\tilde{\beta}_j = \begin{cases} 0, & \text{若 } j \notin A_N, \\ \hat{\beta}_j, & \text{若 } j \in A_N, \end{cases} \tag{2.9}$$

$\tilde{\beta}$ 作为 β^* 的估计.

对上述算法, 不仅要证明集收敛, 即对充分大的数据长度 N 有 $A_N = A^*$, 还要证明参数收敛, 即 $\tilde{\beta} \to \beta^*$.

定理 2.1 [39] 假设 $f(\cdot)$ 在 x^0 二次连续可微, $\{x(k)\}_{k \geqslant 1}$ 为独立同分布随机序列且在 x^0 处有正的概率密度函数, 选择带宽常数 h 满足当数据长度 $N \to \infty$ 时, $h \to 0$ 并且 $Nh^{p+6} \to \infty$, 选取加权系数 $\lambda_M = cM^{2/3}$, 其中常数 $c > 0$, 则当 N 趋于无穷时有如下结论成立:

$$A_N = A^*,$$
$$\tilde{\beta}_j \to \beta_j^*, \quad j \in A^*.$$

本节针对非线性系统作用变量 “全局稠密” 和 “局部稀疏” 的特点, 从系统的任意固定点处着眼, 给出了变量选择算法, 并证明了所得估计值同时具有 “集收敛” 和 “参数收敛”. 从算法构造上看, 本节属于 “自上而下” 的算法, 从全维系统出发, 进而剔除冗余变量, 保留作用变量. 下一节将介绍 “自下而上” 的变量选择算法, 具体地, 算法从一维系统出发, 首先判断最重要的变量, 进而研究二维系统, 判断次重要的变量, 依次类推, 给出所有变量重要性的刻画.

3 全局变量的选择算法

3.1 前向/后向变量选择算法

考虑系统 (1.1), 设数据集为 $\{x(k), y(k)\}_{k=1}^{N}$. 与前一节不同, 本节考虑全局作用变量的辨识: 系统 (1.1) 的维数为 p, 给定 n ($n \leqslant p$), 构造算法判定 n 个对系统最为重要的变量并依重要性程度排序.

算法主要依据以下思路: 假若 $x_{i_1}(k)$ 是系统最重要的变量, 那么用 $x_{i_1}(k)$ 描述系统所得的误差比用其他变量描述系统的误差要小, 即存在函数 $g(\cdot)$, 使得

$$\begin{aligned} &|f(x_1(k), \ldots, x_p(k)) - g(x_{i_1}(k))| \leqslant |f(x_1(k), \ldots, x_p(k)) - h(x_j(k))|, \\ &\quad \forall\, h(\cdot) : \mathbb{R} \to \mathbb{R}, \quad \forall\, j = 1, \ldots, p. \end{aligned} \tag{3.1}$$

上述思路依次推广到 2 维、3 维直至 n 维系统, 从而确定系统最为重要的前 n 个变量.

下面给出具体算法.

首先介绍低维邻域的概念: 任意固定 i ($1 \leqslant i \leqslant p$) 和数据 $x(k) = [x_1(k) \cdots x_p(k)]^{\mathrm{T}}$ ($1 \leqslant k \leqslant N$), 若数据 $x(j)$ ($1 \leqslant j \leqslant N$) 满足

$$|x_i(k) - x_i(j)| \leqslant h, \tag{3.2}$$

则称数据 $x(j)$ 落于 $x_i(k)$ 的 "一维邻域", 其中 $h > 0$ 为带宽参数, h 的取值决定了一维邻域的大小.

记落于 $x_i(k)$ 一维邻域的数据集合为 $\{x(k_1^i), x(k_2^i), \ldots, x(k_{l_i}^i)\}$, 相应的系统输出为 $\{y(k_1^i), y(k_2^i), \ldots, y(k_{l_i}^i)\}$, 基于一维邻域内的数据构造点 $x_i(k)$ 处的一维核估计 (参见文献 [38])

$$\hat{f}_i(x(k)) = \frac{\sum_{j=1}^{l_i} K_1((\frac{x(k_j^i) - x(k)}{h})_i) y(k_j^i)}{\sum_{j=1}^{l_i} K_1((\frac{x(k_j^i) - x(k)}{h})_i)}, \tag{3.3}$$

其中

$$\left(\frac{x(k_j^i) - x(k)}{h}\right)_i = \sqrt{\left(\frac{x_i(k_j^i) - x_i(k)}{h}\right)^2},$$

$K_1(\cdot)$ 为单变量核函数, 除满足第 2 节的假设条件外, 还满足有界支撑, 即若 $|x| > 1$, 则 $K_1(x) = 0$.

依核估计 (3.3), 计算指标 i 的残差 (residual)

$$\mathrm{RSS}(i) = \frac{1}{N} \sum_{k=1}^{N} (y(k) - \hat{f}_i(x(k)))^2. \tag{3.4}$$

显然, 若 $x(k) = [x_1(k) \cdots x_p(k)]^{\mathrm{T}}$ 的第 i 个分量对系统最为重要, 则第 i 个分量对应的残差应为最小, 因而定义如下估计作为系统 (1.1) 最为重要的变量:

$$i_1^* = \underset{1 \leqslant i \leqslant p}{\operatorname{argmin}}\, \mathrm{RSS}(i). \tag{3.5}$$

依上面思路, 可从 $\{1, 2, \ldots, p\}/i_1^*$ 中选取系统第 2 重要的变量, 具体算法如下:

任意固定 $i \in \{1, 2, \ldots, p\}/i_1^*$ 和 $k \in \{1, \ldots, N\}$, 基于带宽 h 构造 $(x_{i_1^*}(k), x_i(k))$ 的 "二维邻域"

$$\{x(j) \mid (x(k) - x(j))_{i_1^*, i} = \sqrt{(x_{i_1^*}(k) - x_{i_1^*}(j))^2 + (x_i(k) - x_i(j))^2} \leqslant h, i \neq i_1^*\}. \tag{3.6}$$

相应的输入输出集合分别记为 $\{x(k_1^{i_1^*,i}),x(k_2^{i_1^*,i}),\ldots,x(k_{l_{i_1^*,i}}^{i_1^*,i})\}$ 和 $\{y(k_1^{i_1^*,i}),y(k_2^{i_1^*,i}),\ldots,y(k_{l_{i_1^*,i}}^{i_1^*,i})\}$; 基于二维邻域内的数据构造核估计

$$\hat{f}_{i_1^*,i}(x(k))=\frac{\sum_{j=1}^{l_{i_1^*,i}}K_2((\frac{x(k_j^{i_1^*,i})-x(k)}{h})_{i_1^*,i})y(k_j^{i_1^*,i})}{\sum_{j=1}^{l_{i_1^*,i}}K_2((\frac{x(k_j^{i_1^*,i})-x(k)}{h})_{i_1^*,i})},\tag{3.7}$$

其中 $K_2(\cdot)$ 为有界支撑的二维核函数,

$$\left(\frac{x(k_j^{i_1^*,i})-x(k)}{h}\right)_{i_1^*,i}=\sqrt{\left(\frac{x_{i_1^*}(k_j^{i_1^*,i})-x_{i_1^*}(k)}{h}\right)^2+\left(\frac{x_i(k_j^{i_1^*,i})-x_i(k)}{h}\right)^2};$$

进一步, 构造残差

$$\mathrm{RSS}(i_1^*,i)=\frac{1}{N}\sum_{k=1}^{N}(\hat{f}_{i_1^*,i}(x(k))-y(k))^2,\tag{3.8}$$

并定义

$$i_2^*=\underset{i\in\{1,2,\ldots,p\}/i_1^*}{\operatorname{argmin}}\ \mathrm{RSS}(i_1^*,i),\tag{3.9}$$

i_2^* 即是系统的第 2 重要变量.

依次类推, 可逐项确定系统最为重要的前 n 个变量 $x_{i_1^*}(k),x_{i_2^*}(k),\ldots,x_{i_n^*}(k)$, 至此完成了变量的前向选择. 为避免前向选择可能导致的错误和疏漏, 进一步利用后向变量选择算法来对前向算法的结果进行检验. 直观上讲, 检验集合 $\{1,2,\ldots,p\}/\{i_1^*,\ldots,i_n^*\}$ 中的变量是否能进一步减小残差, 若不能, 则保留前向算法所得结果; 若进一步减小残差, 则后向算法辨识出更为重要的变量, 更新辨识结果. 具体算法如下:

选取 $i\in\{1,2,\ldots,p\}/\{i_1^*,i_2^*,\ldots,i_n^*\}$, 固定 $k\in\{1,\ldots,N\}$ 和带宽 h, 计算 $(x_i(k),x_{i_2^*}(k),\ldots,x_{i_n^*}(k))$ 的 n 维邻域, 相应的输入输出集记为 $\{x(k_j^{i,i_2^*,\ldots,i_n^*}),j=1,\ldots,l_{i,i_2^*,\ldots,i_n^*}\}$ 和 $\{y(k_j^{i,i_2^*,\ldots,i_n^*}),\ j=1,\ldots,l_{i,i_2^*,\ldots,i_n^*}\}$, 计算核估计

$$\hat{f}_{i,i_2^*,\ldots,i_n^*}(x(k))=\frac{\sum_{j=1}^{l_{i,i_2^*,\ldots,i_n^*}}K_n((\frac{x(k_j^{i,i_2^*,\ldots,i_n^*})-x(k)}{h})_{i,i_2^*,\ldots,i_n^*})\cdot y(k_j^{i,i_2^*,\ldots,i_n^*})}{\sum_{j=1}^{l_{i,i_2^*,\ldots,i_n^*}}K_n((\frac{x(k_j^{i,i_2^*,\ldots,i_n^*})-x(k)}{h})_{i,i_2^*,\ldots,i_n^*})},\tag{3.10}$$

其中 $K_n(\cdot)$ 为 n 维核函数, 残差

$$\mathrm{RSS}(i,i_2^*,\ldots,i_n^*)=\frac{1}{N}\sum_{k=1}^{N}(y(k)-\hat{f}_{i,i_2^*,\ldots,i_n^*}(x(k)))^2,\tag{3.11}$$

并计算残差的极小值点

$$i^*=\underset{i\in(1,2,\ldots,p)/(i_1^*,i_2^*,\ldots,i_n^*)}{\operatorname{argmin}}\ \mathrm{RSS}(i,i_2^*,\ldots,i_n^*).\tag{3.12}$$

若 $\mathrm{RSS}(i^*,i_2^*,\ldots,i_n^*)<\mathrm{RSS}(i_1^*,i_2^*,\ldots,i_n^*)$, 即系统的残差可进一步减小, $x_{i^*}(k)$ 相对于 $x_{i_1^*}(k)$ 更为重要, 系统前 n 个重要变量更新为 $(x_{i^*}(k),x_{i_2^*}(k),\ldots,x_{i_n^*}(k))$; 若 $\mathrm{RSS}(i^*,i_2^*,\ldots,i_n^*)\geqslant\mathrm{RSS}(i_1^*,i_2^*,\ldots,i_n^*)$, 即 $i\in\{1,2,\ldots,p\}/\{i_1^*,i_2^*,\ldots,i_n^*\}$ 中不存在相对于 $x_{i_1^*}(k)$ 更为重要的变量, 前向选择算法得到的辨识结果得到保留. 不失一般性, 记执行一步后向选择算法得到的变量仍为 $(x_{i_1^*}(k),x_{i_2^*}(k),\ldots,x_{i_n^*}(k))$, 依上述过程, 依次对 $i_2^*,\ldots,i_n^*$ 执行后向选择算法, 最终结果即作为系统最为重要的前 n 个变量.

前向选择算法如下:

(III.A1) 选定变量 $x_i(k)$, $1 \leqslant i \leqslant p$, 计算一维邻域 (3.2)、核估计 (3.3) 和残差 (3.4);

(III.A2) 依 (3.5) 计算残差的极小值点 i_1^* 及对应变量 $x_{i_1^*}(k)$;

(III.A3) 选定变量 $x_i(k)$, $i \in (1,2,\ldots,p)/i_1^*$, 计算二维邻域 (3.6)、核估计 (3.7) 和残差 (3.8);

(III.A4) 依 (3.9) 计算残差的极小值点 i_2^* 及对应变量 $x_{i_2^*}(k)$;

(III.A5) 依上述思路, 计算 $i_3^*,\ldots,i_n^*$.

后向选择算法如下:

(III.B1) 选取 $i \in \{1,2,\ldots,p\}/\{i_1^*,i_2^*,\ldots,i_n^*\}$, 计算 $(x_i(k),x_{i_2^*}(k),\ldots,x_{i_n^*}(k))$ 的 n 维邻域, 核估计 (3.10) 和残差 (3.11);

(III.B2) 依 (3.12) 计算残差的极小值点 i^* 及对应变量 $x_{i^*}(k)$, 若 $\mathrm{RSS}(i^*,i_2^*,\ldots,i_n^*) < \mathrm{RSS}(i_1^*,i_2^*,\ldots,i_n^*)$, 系统前 n 个重要变量更新为 $(x_{i^*}(k),x_{i_2^*}(k),\ldots,x_{i_n^*}(k))$; 若 $\mathrm{RSS}(i^*,i_2^*,\ldots,i_n^*) \geqslant \mathrm{RSS}(i_1^*,i_2^*,\ldots,i_n^*)$, 前向选择算法辨识结果得以保留;

(III.B3) 依上述过程, 依次对 $i_2^*,\ldots,i_n^*$ 执行后向选择算法, 最终结果即作为系统最为重要的前 n 个变量.

从上可见, 前向/后向算法的关键在于核估计, 对核估计有如下结论成立:

定理 3.1 [39] 考虑系统 (1.1), 假设作用变量的真实个数为 n^*, $n^* \leqslant p$, $n^* \leqslant n$. 假设 $f(\cdot)$ 连续可微, 系统噪音为独立同分布随机序列且二阶矩有限, 选取带宽参数满足 $h \to 0$, $h^n N \to \infty$, 若数据集 $\{x(k)\}$ 为 α 混合相依且混合系数满足 $0 < \alpha(k) \leqslant c\rho^k$, 其中常数 $c > 0$, $0 < \rho < 1$, 则有 $\hat{f}_{i_1^*,i_2^*,\ldots,i_n^*}(x^0) - f(x^0)$ 依概率收敛到零, 其中 $x^0 \in \mathbb{R}^p$ 为任意给定点.

所谓 "混合相依" 是指系统数据随时间间隔的增加渐近独立. 一般来讲, 若系统具有一定的 "稳定性"、输入信号和噪音具有一定的 "激励性", 则 "混合相依" 必然成立, 相关结论可参见文献 [38, 40].

3.2 变量个数 n 的辨识

上节在系统的变量个数为 n 的条件下, 给出了变量选择算法, 如何确定变量个数 n 成为应用变量选择算法的前提. 本节将探讨这个问题并给出具体算法.

假设系统作用变量的真实个数为 n^0, 则对 n ($n^0 \leqslant n \leqslant p$), 存在函数 $g(\cdot)$ 使得

$$f(x(k)) - g_{x_{i_1^*},\ldots,x_{i_n^*}}(x(k)) = 0, \tag{3.13}$$

$$y(k) - g_{x_{i_1^*},\ldots,x_{i_n^*}}(x(k)) = v(k), \tag{3.14}$$

其中 $x_{i_1^*},\ldots,x_{i_n^*}$ 包括了系统所有的真实变量. 从 (3.14) 可知, 若 $n \geqslant n^0$, 则 $y(k) - g_{x_{i_1^*},\ldots,x_{i_n^*}}(x(k))$ 为白噪音, 而当 $n < n^0$ 时, $y(k) - g_{x_{i_1^*},\ldots,x_{i_n^*}}(x(k))$ 非白噪音. 因此, 检验 (3.14) 是否为白噪音为变量个数的辨识提供了一种思路.

定义

$$r(k) = y(k) - g_{x_{i_1^*},\ldots,x_{i_n^*}}(x(k)), \tag{3.15}$$

$$\gamma(j) = E(r(k) - Er(k))(r(k-j) - Er(k)), \tag{3.16}$$

$$\rho(j) = \frac{\gamma(j)}{\gamma(0)}. \tag{3.17}$$

进一步定义噪音估计值、样本均值、样本自协方差函数和样本自相关函数如下:

$$\hat{r}(k) = y(k) - \hat{f}_{i_1^*,\ldots,i_n^*}(x(k)), \tag{3.18}$$

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N}\hat{r}(k), \tag{3.19}$$

$$\hat{\gamma}(j) = \frac{1}{N-j}\sum_{k=j+1}^{N}(\hat{r}(k)-\hat{\mu})(\hat{r}(k-j)-\hat{\mu}), \tag{3.20}$$

$$\hat{\rho}(j) = \frac{\hat{\gamma}(j)}{\hat{\gamma}(0)}, \tag{3.21}$$

其中 $\hat{f}_{i_1^*,\ldots,i_n^*}(x(k))$ 为前向/后向算法所得的核估计.

若 $\hat{r}(k)$ 近似为白噪音, 由 Box-Pierce 检验可知, $N\sum_{j=1}^{p-1}\hat{\rho}(j)^2$ 近似服从自由度 $p-1$ 的 χ^2 分布. 因而构造基于假设检验的系统变量个数 n 的辨识算法如下:

(1) 原假设 H_0: $r(k)$ 为白噪音;

(2) 备择假设 H_1: $r(k)$ 非白噪音.

给定显著性水平 α $(0<\alpha<1)$ 及相关的阈值 d, 计算 $N\sum_{j=1}^{p-1}\hat{\rho}(j)^2$: 若 $N\sum_{j=1}^{p-1}\hat{\rho}(j)^2 < d$, 则接受 H_0; 否则拒绝 H_0, 系统真实变量个数应比假设检验中设定的 n 更大. 对 $n=1,\ldots,p$ 执行上面的假设检验, 使 $r(k)$ 为白噪音的最小正整数, 即作为系统真实变量个数的估计.

从上述算法可见, 检验 $r(k)$ 是否为白噪音转化为检验 $r(k)$ 与 $r(k-j)$ 不相关, 进而应用 Box-Pierce 检验. 对线性系统来讲, 这种算法十分有效, 但对非线性系统有时可能得到与实际不一致的结果 (参见文献 [41, 42]). 为保证算法的稳健性以及对非线性系统的有效性, 可考虑如下推广的 Box-Pierce 检验: 根据文献 [41, 42], 当 $N\to\infty$ 时,

$$Q_{p-1} = N[\rho(1),\ldots,\rho(p-1)]V^{-1}\begin{bmatrix}\rho(1)\\ \vdots\\ \rho(p-1)\end{bmatrix} \tag{3.22}$$

服从自由度为 $p-1$ 的 χ^2 分布, 其中

$$V = \frac{C}{\gamma(0)^2} = \begin{bmatrix}c_{11} & \cdots & c_{1,p-1}\\ \vdots & \ddots & \vdots\\ c_{p-1,1} & \cdots & c_{p-1,p-1}\end{bmatrix}/\gamma(0)^2, \quad C = \begin{bmatrix}c_{11} & \cdots & c_{1p}\\ \vdots & \ddots & \vdots\\ c_{p1} & \cdots & c_{pp}\end{bmatrix},$$

$$c_{ij} = \sum_{l=-\infty}^{\infty}E(r(k)-\mu)(r(k-i)-\mu)(r(k+l)-\mu)(r(k+l-j)-\mu), \quad i,j=1,\ldots,p-1.$$

由 (3.22) 可见, 当矩阵 V 为单位阵时, Box-Pierce 检验实为 (3.22) 的特殊情形.

利用系统的测量数据估计 (3.22) 中的 $\rho(i)$ $(i=1,\ldots,p-1)$ 和矩阵 V, 指定显著性水平 α 和阈值 d, 作假设检验可辨识系统的真实变量个数 n.

3.3 残差函数全局极小值点的辨识—分枝定界算法

前一节给出的前向/后向算法中, 前向算法每步辨识出一个重要变量, 后向算法每步检验一个辨识出的变量是否正确, 由于系统 (1.1) 本身的非线性, 若算法每步只寻找/检验一个变量, 则可能辨识结

果会落于残差函数的局部极小值点而非全局极小值点. 因此, 有必要在前向/后向算法辨识结果的基础上进一步设计算法, 寻找残差函数的全局极小值点.

注意到系统维数为 p, 真实作用变量的个数为 n, 若用穷举算法, 需计算 $\frac{p(p-1)\cdots(p-n+1)}{n(n-1)\cdots1}$ 种可能. 对维数 p 很大的系统, 穷举法显然难以具体操作. 另一方面, 尽管前向/后向算法所得结果可能并非全局极小值点, 但它为进一步搜索全局极小值点提供了一个好的初始值. 基于这种思路, 我们引入如下分枝定界算法.

算法具体思路如下: 将变量集 $\{x_1,\ldots,x_q\}$ 分成互不包含的若干个子集 (可能有交集), 计算每个子集的残差, 若某个子集的残差大于前向/后向算法所得的残差, 则可断言此变量集必不含全局极小值点, 不必在此子集中进一步搜索; 反之, 若此子集的残差小于前向/后向算法所得到的残差, 意味着此子集包含更为重要的系统变量, 将此变量集划分成若干子集进一步寻优直至算法收敛.

通过图 1 所给的例子进一步解释上述算法: 假设系统维数 $p=8$, 系统变量为 $(x_1,\ldots,x_8)$, 真实变量个数 $n=2$, 通过前向/后向算法所得估计值为 (x_3,x_5), 以 (x_3,x_5) 为初值, 可将系统变量进一步划分为以下集合:

$$A=\{(x_1,x_2,x_4,x_6,x_7,x_8)\}, \tag{3.23}$$

$$B=\{(x_3,x_5)\}, \tag{3.24}$$

$$C=\{(x_3,x_1),(x_3,x_2),(x_3,x_4),(x_3,x_6),(x_3,x_7),(x_3,x_8)\}, \tag{3.25}$$

$$D=\{(x_5,x_1),(x_5,x_2),(x_5,x_4),(x_5,x_6),(x_5,x_7),(x_5,x_8)\}. \tag{3.26}$$

根据前向/后向算法的后向选择过程, 可知变量 (x_3,x_5) 对应的残差一定小于集 C 与 D 中的变量对应的残差, 因此可直接忽略计算集 C 与 D 残差的过程; 对集 A, 若 $\mathrm{RSS}(1,2,4,6,7,8)\geqslant\mathrm{RSS}(3,5)$, 可断言 (x_3,x_5) 即是残差的全局极小值点; 若 $\mathrm{RSS}(1,2,4,6,7,8)<\mathrm{RSS}(3,5)$, 则全局极小值点落于变量集合 $(x_1,x_2,x_4,x_6,x_7,x_8)$, 利用前向/后向算法寻找变量集 $(x_1,x_2,x_4,x_6,x_7,x_8)$ 的残差极小值点, 然后利用估计值进一步将 $(x_1,x_2,x_4,x_6,x_7,x_8)$ 划分成互不包含的子集, 不断重复上述过程直至算法收敛.

由上述寻优过程可见, 算法避免了穷举法的不可操作性, 所得估计值必为残差的全局极小值点. 注意到上述算法实质上形成了 “树状” 的寻优过程, 并且寻优过程限定在树的一个有界分枝内, 因而称为分枝定界 (branch and bound) 算法.

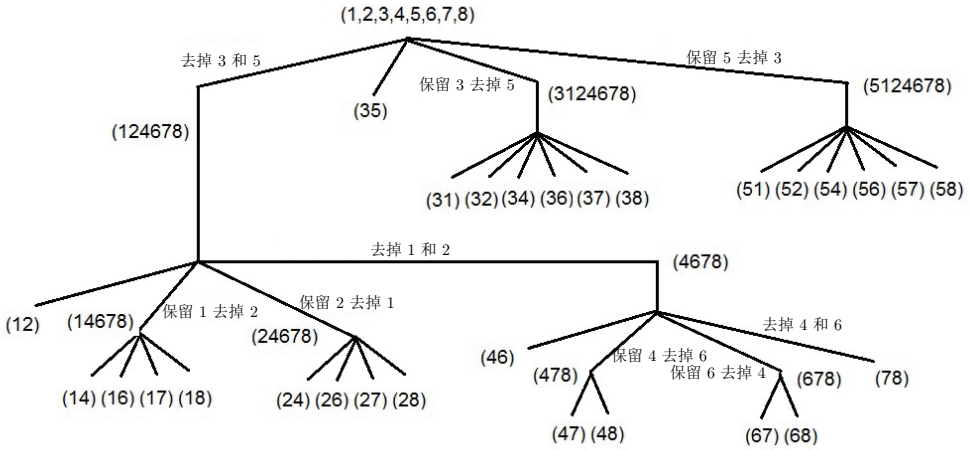

图 1　分枝定界算法

4 可加非线性模型的变量选择

4.1 变量选择算法

前两节针对系统 (1.1), 分别给出了局部作用变量和全局作用变量的辨识算法. 系统 (1.1) 涵盖了一大类动态系统, 如伴有外部输入的自回归系统 (autoregressive systems with exogeneous input, ARX) 和 Hammerstein 系统等, 因而, 变量选择算法可应用到这些系统的辨识上. 另一方面, 针对特殊结构的非线性系统, 可以有针对性地设计变量选择算法, 本节考虑单变量可加非线性系统的相关问题, 这类系统在工程技术领域有广泛应用 (参见文献 [38, 43–45]).

考虑单变量可加非线性系统

$$y(k)=f_0+\sum_{j=1}^{d}f_j(x(kj))+v(k),\quad 1\leqslant k\leqslant n, \tag{4.1}$$

其中 $\{y(k),x(k1),\dots,x(kd),k=1,\dots,n\}$ 为测量数据, f_0 为未知常数, $\{f_j(\cdot)\}_{j=1}^{d}$ 为单变量未知函数, $v(k)$ 为系统噪音.

定义 $\mathcal{I}=\{j=1,\dots,d\mid f_j(\cdot)\neq 0\}$, $\mathcal{I}^c=\{1,\dots,d\}\backslash\mathcal{I}$. 系统 (4.1) 的变量选择就是要判断 $\{f_j(\cdot)\}_{j=1}^{d}$ 中哪些为零、哪些非零, 即是要辨识集合 $\mathcal{I}$ 与 $\mathcal{I}^c$.

定义 $Y=[y(1),\dots,y(n)]^{\mathrm{T}}$, $V=[v(1),\dots,v(n)]^{\mathrm{T}}$, $f_j=[f_j(x(1j)),f_j(x(2j)),\dots,f_j(x(nj))]^{\mathrm{T}}$, $\mathbf{1}_n=[1,\dots,1]^{\mathrm{T}}$, 系统 (4.1) 可表示为

$$Y=f_0\mathbf{1}_n+\sum_{j=1}^{d}f_j+V. \tag{4.2}$$

辨识算法可分为两步, 由于 $\{f_0,f_j(\cdot),j=1,\dots,d\}$ 未知, 首先构造算法估计 $\{f_0,f_j(\cdot),j=1,\dots,d\}$, 在此基础上, 构造准则函数并以此来判断 $\{f_j(\cdot)\}_{j=1}^{d}$ 的零元和非零元.

假设 $\{f_0,f_j(\cdot),j=1,\dots,d\}$ 的估计为 $\{\hat{f}_0,\hat{f}_j(\cdot),j=1,\dots,d\}$ (具体算法下节给出), 系统 (4.1) 的变量选择算法如下:

$$\hat{c}=[\hat{c}_1,\dots,\hat{c}_d]^{\mathrm{T}}=\underset{c_i\geqslant 0}{\operatorname{argmin}}\left\{\frac{1}{2}\left\|Y-\hat{f}_0-\sum_{j=1}^{d}c_j\hat{f}_j\right\|^2+\lambda_n\sum_{j=1}^{d}c_j\right\}. \tag{4.3}$$

针对系统 (4.1) 和算法 (4.3), 引入如下假设:

(A3.1) $\|(f_{\mathcal{I}}^{\mathrm{T}}f_{\mathcal{I}}/n)^{-1}\|<\infty$, 其中 $f_{\mathcal{I}}$ 是由 f_j ($j\in\mathcal{I}$) 构成的矩阵, 并且 $\|f_j\|/\sqrt{n}<\infty$, $j=1,\dots,d$;

(A3.2) 存在趋于零的正序列 $\{\delta_n\}_{n\geqslant 1}$ 使得 $\|\hat{f}_j-f_j\|^2/n=O_P(\delta_n^2)$, $j=1,\dots,d$, 其中 $\hat{f}_j$ 是 f_j 的估计;

(A3.3) 选取权系数 λ_n 满足 $\lambda_n/n\to 0$, 并且 $\delta_n=o(\lambda_n/n)$.

假设 (A3.1) 是系统 (4.1) 的 "持续激励条件", 假设 (A3.2) 要求函数估计具有一定的收敛速度. 在上述假设下, 可证明以下结论:

定理 4.1 [46] 考虑可加非线性系统 (4.1). 假设 (A3.1)–(A3.3) 成立, 则有

$$\mathrm{P}(\hat{c}_j=0)\to 1,\quad \forall\, j\in\mathcal{I}^c, \tag{4.4}$$

$$\mathrm{P}(\hat{c}_j>0)\to 1,\quad \forall\, j\in\mathcal{I}. \tag{4.5}$$

由定理 4.1 可见, 通过判断准则函数 (4.3) 的极小值点为零或非零, 就可判断对应函数 $f_j(\cdot)$ 是否为零, 从而实现变量选择的目的. 准则函数 (4.3) 需要 $f_j(\cdot)$ 的估计, 下一节给出辨识算法.

4.2 f_0 和 f_j $(j=1,\ldots,d)$ 的辨识算法

考虑系统 (4.1) 的开环辨识, 则可以适当地选择输入信号使系统具有平稳遍历性 (参见文献 [38, 40]), 不妨假设 $Ef_j(x(kj))=0$, 否则系统 (4.1) 可等价改写为

$$y(k)=\bar{f}_0+\sum_{j=1}^{d}\bar{f}_j(x(kj))+v(k),\quad 1\leqslant k\leqslant n, \tag{4.6}$$

其中 $\bar{f}_0=f_0+\sum_{j=1}^{d}Ef_j(x(kj))$, $\bar{f}_j(x(kj))=\sum_{j=1}^{d}(f_j(x(kj))-Ef_j(x(kj)))$.

首先考虑 $f_j(\cdot)$ 的核估计

$$\hat{f}_j(v_j)=\frac{\sum_{i=1}^{n}K_{h_j}(x(ij)-v_j)y_i}{\sum_{i=1}^{n}K_{h_j}(x(ij)-v_j)}, \tag{4.7}$$

其中 h_j 为带宽参数, $K(\cdot)$ 为一维核函数. 可以证明, 若对任意 $i\neq j$ 有 $\{x(kj)\}_{k\geqslant 1}$ 与 $\{x(ki)\}_{k\geqslant 1}$ 独立, 则核估计 (4.7) 是 $f_j(v_j)$ 的强一致估计 (参见文献 [38]). 但对系统 (4.1), $\{x(kj)\}_{k\geqslant 1}$ 与 $\{x(ki)\}_{k\geqslant 1}$ 独立的假设过于苛刻, 这种假设排除了工程中常见的许多系统 (如可加 NARX 系统等), 因而有必要改进算法、放宽假设条件. 为此, 我们首先引入以下算法:

$$\hat{p}_j(v_j)=\frac{1}{n}\sum_{i=1}^{n}K_{h_j}(x(ij)-v_j), \tag{4.8}$$

$$\hat{p}_{jl}(v_j,v_l)=\frac{1}{n}\sum_{i=1}^{n}K_{h_j}(x(ij)-v_j)K_{h_l}(x(ij)-v_l). \tag{4.9}$$

构造准则函数

$$\min_{\bar{f}_0,\bar{f}_1,\ldots,\bar{f}_d}\frac{1}{2}\int\sum_{i=1}^{n}\Big(y_i-\bar{f}_0-\sum_{j=1}^{d}\bar{f}_j(v_j)\Big)^2\prod_{r=1}^{d}K_{h_r}(x_{ir}-v_r)dv, \tag{4.10}$$

$$\text{s.t.}\ \int\bar{f}_j(v_j)\hat{p}_j(v_j)dv_j=0,\quad j=1,\ldots,d. \tag{4.11}$$

准则函数 (4.10) 和 (4.11) 是针对 $\bar{f}_0,\bar{f}_1,\ldots,\bar{f}_d$ 的泛函优化, 其极小值点记为 $\tilde{f}_0,\tilde{f}_1,\ldots,\tilde{f}_d$. 利用 Lagrange 乘子和变分法, 可得极小值点 $\tilde{f}_0,\tilde{f}_1,\ldots,\tilde{f}_d$ 满足方程 (参见文献 [47])

$$\widetilde{f}_0=\frac{1}{n}\sum_{i=1}^{n}y_i-\sum_{j=1}^{d}\int\hat{p}_j(v_j)\widetilde{f}_j(v_j)dv_j, \tag{4.12}$$

$$\widetilde{f}_j(v_j)=\hat{f}_j(v_j)-\sum_{l\neq j}\int\widetilde{f}_l(v_l)\frac{\hat{p}_{jl}(v_j,v_l)}{\hat{p}_j(v_j)}dv_l-\overline{Y},\quad j=1,\ldots,d, \tag{4.13}$$

其中 $\hat{f}_j(v_j)$ 是 (4.7) 给出的核估计, $\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

任意给定初值 $\{\widetilde{f}_j^{(0)}(\cdot),j=1,\ldots,d\}$, (4.12) 和 (4.13) 可由以下算法迭代求解:

$$\widetilde{f}_j^{(k)}(v_j)=\hat{f}_j(v_j)-\sum_{l<j}\int\widetilde{f}_l^{(k)}(v_l)\frac{\hat{p}_{jl}(v_j,v_l)}{\hat{p}_j(v_j)}dv_l-\sum_{l>j}\int\widetilde{f}_l^{(k-1)}(v_l)\frac{\hat{p}_{jl}(v_j,v_l)}{\hat{p}_j(v_j)}dv_l-\overline{Y}, \tag{4.14}$$

$$\widetilde{f}_0^{(k)}=\overline{Y}-\sum_{j=1}^{d}\int\hat{p}_j(v_j)\widetilde{f}_j^{(k)}(v_j)dv_j. \tag{4.15}$$

注意 (4.14) 和 (4.15) 包含函数 $\widetilde{f}_j^{(k)}(\cdot)$ 在整个实数域上的积分. 在实际计算中, 我们首先给出初值 $\widetilde{f}_j^{(0)}(\cdot)$ 在有限个格点上的值以及相应的密度估计 (4.7), 据此以有限和代替无穷积分, 迭代计算 (4.14) 和 (4.15).

算法总结如下:

(IV.1) 设置格点并依 (4.7)–(4.9) 计算 $\hat{p}_j$ 和 $\hat{p}_{jl}$, $j \neq l$;

(IV.2) 设置迭代算法的初始值 $\widetilde{f}_j^{(0)} = \hat{f}_j$, $j = 1, \ldots, d$;

(IV.3) 迭代计算 (4.14) 和 (4.15) 直至满足算法的终止条件;

(IV.4) 利用格点上的函数估计值作内插以得到大范围的函数逼近.

可以证明, 在系统具有混合相依性的条件下, 迭代算法产生的估计序列具有收敛性和渐近正态性. 综合算法 (4.3) 和 (4.10), 对可加非线性系统 (4.1) 同时实现了未知函数的估计和变量选择 (参见文献 [46, 47]).

5 仿真例子

本节主要考察第 3 节算法的数值模拟, 第 2 和 5 节算法的数值模拟可参见文献 [39, 46].

例 5.1 考虑如下数值例子 (参见文献 [23]):

$$\begin{aligned} y(k) = {} & 10\sin(x_1(k)x_2(k)) + 20(x_3(k) - 0.5)^2 + 10x_4(k) + 5x_5(k) + x_6(k)x_7(k) \\ & + x_7(k)^2 + 5\cos(x_6(k)x_8(k)) + \exp(-|x_8(k)|) + 0.5\eta(k), \end{aligned} \tag{5.1}$$

其中 $\eta(k)$ 是独立同分布的随机变量, $\eta(k) \sim \mathcal{N}(0,1)$, $x_3(k)$ 和 $x_5(k)$ 是区间 $[-1,1]$ 上均匀分布的独立同分布随机序列, 系统的其他变量满足

$$\begin{aligned} x_4(k) &= x_3(k) \cdot x_5(k) + 0.1 \cdot \eta(k), \\ x_1(k) &= x_3(k)^2 \cdot x_5(k) + 0.1 \cdot \eta(k), \\ x_2(k) &= x_3(k) \cdot x_5(k)^2 + 0.1 \cdot \eta(k), \\ x_6(k) &= x_1(k) - x_4(k) + 0.1 \cdot \eta(k), \\ x_7(k) &= x_3(k)^3 \cdot x_5(k) + 0.1 \cdot \eta(k), \\ x_8(k) &= x_2(k) \cdot x_5(k) + 0.1 \cdot \eta(k). \end{aligned} \tag{5.2}$$

由上可见, 系统 (5.1) 虽为 8 维系统, 真正的作用变量只有 $x_3(k)$ 和 $x_5(k)$, 变量个数 $n = 2$.

设数据长度为 500, 设定带宽参数 $h = 0.2$. 利用第 3 节所给算法, 考察变量个数和作用变量的辨识.

第 1 步 利用前向/后向算法, 当 $n = 1, 2, \ldots, 8$ 时, 依次计算系统残差. 从图 2 可见, 当 $n \geqslant 2$ 时, 系统残差基本保持不变, 因而初步判定系统的变量个数 $n = 2$.

第 2 步 利用推广的 Box-Pierce 检验辨识变量个数: 计算 (3.22), 可得 $Q_{p-1} = Q_7 = 8.2273$, 给定显著性水平 $\alpha = 0.05$, 对应 $\chi^2_{0.05}(7) = 14.07$, $\chi^2_{0.05}(7) > Q_{p-1}$, 因此接受 $n = 2$ 作为系统真实的变量个数.

第 3 步 固定 $n = 2$, 利用前向/后向算法辨识系统的作用变量, 重复 100 次数值模拟, 每次试验的数据长度均为 500, 每次试验的辨识结果均为 $x_3(k)$ 和 $x_5(k)$.

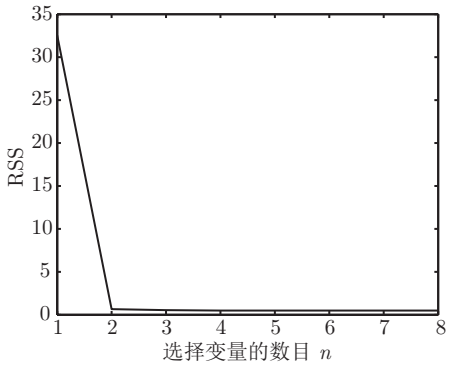

图 2　残差与变量个数的关系

第 4 步 利用分枝定界算法寻找残差函数的全局极小值点: 依图 3 所示, 将变量集分成若干子集, 计算可得

$$\mathrm{RSS}(x_3,x_5)\leqslant\min\{\mathrm{RSS}(x_j,x_5),\mathrm{RSS}(x_3,x_j)\},\quad j=(1,2,4,6,7,8),$$
$$\mathrm{RSS}(x_3,x_5)=0.6484<\mathrm{RSS}(x_1,x_2,x_4,x_6,x_7,x_8)=23.04.$$

从而可知, $x_3(k)$ 和 $x_5(k)$ 是残差函数的全局极小值点.

第 5 步 辨识结果的验证: 由前四步的结果可知, 存在函数 $g(\cdot)$ 使得

$$y(k)=f(x_1(k),x_2(k),x_3(k),x_4(k),x_5(k),x_6(k),x_7(k),x_8(k))=g(x_3(k),x_5(k)).$$

选取一组新的输入信号

$$x_3(k)=0.9*\sin\left(\frac{2\pi k}{20}\right),\quad x_5(k)=0.9*\cos\left(\frac{2\pi k}{20}\right),\quad k=1,\ldots,40.$$

利用前 500 组数据作系统辨识, 后 40 组数据作验证. 具体地讲, 分别假设系统的变量个数为 8 和 2, 利用前 500 组数据分别估计函数 f 和 g, 设估计值为 $\hat{f}$ 和 $\hat{g}$, 然后利用后 40 组数据计算系统输出的估计值

$$\hat{y}_1(k)=\hat{f}(x_1(k),x_2(k),x_3(k),x_4(k),x_5(k),x_6(k),x_7(k),x_8(k)),\quad k=1,\ldots,40,$$
$$\hat{y}_2(k)=\hat{g}(x_3(k),x_5(k)),\quad k=1,\ldots,40,$$

并计算估计值的拟合度 (Goodness-of-Fit, GoF)

$$\mathrm{GoF}_i=\left(1-\sqrt{\frac{\sum(y(k)-\hat{y}_i(k))^2}{\sum(y(k)-\frac{1}{40}\sum y(k))^2}}\right)\times100\%,\quad i=1,2.$$

图 4 中实线是系统真实输出, 折线是假设变量个数 $n=8$ 时系统输出的估计值 (GoF = 0.6940), 点状折线是系统变量个数 $n=2$ 时输出的估计值 (GoF = 0.9205). 显然, 当精确获知系统的真实变量时, 系统的输出估计更为精确.

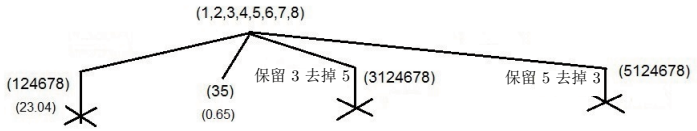

图 3　分枝定界算法寻找残差函数的全局极小值点

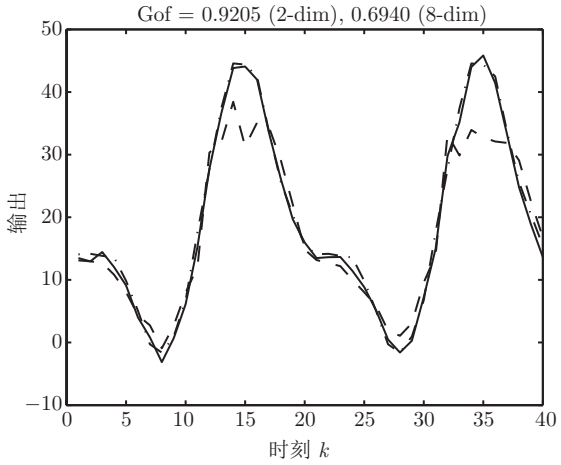

图 4 变量个数分别为 2 和 8 时估计模型的输出与真实输出的对比

例 5.2 在系统 (5.1) 基础上, 考虑如下数值例子:

$$y(k)=0.1\sin(x_1(k)x_2(k))+20(x_3(k)-0.5)^2+10x_4(k)+5x_5(k)+0.1x_6(k)x_7(k)\\+0.1x_7(k)^2+0.1\cos(x_6(k)x_8(k))+0.1\exp(-|x_8(k)|)+0.5\eta(k),\tag{5.3}$$

其中 $\eta(k)$ 是独立同分布的随机变量, $\eta(k)\sim\mathcal{N}(0,1)$, $x_i(k)$ $(i=1,\ldots,8)$ 是区间 $[-1,1]$ 上均匀分布的独立同分布随机序列. 可见, $x_i(k)$ $(i=1,\ldots,8)$ 均是系统的作用变量, 变量个数 $n=8$. 另一方面, 从系统 (5.3) 参数值的大小可见, 变量 x_3、x_4 和 x_5 起主要作用, 相比之下其他变量的作用很小.

仍然利用第 3 节的算法, 第 1 步, 利用前向/后向算法, 当 $n=1,2,\ldots,8$ 时, 依次计算系统残差. 从图 5 可见, 当 $n\geqslant 3$ 时, 系统残差基本保持不变, 因而初步判定系统的变量个数 $n=3$; 第 2 步, 利用推广的 Box-Pierce 检验辨识变量个数: 计算 (3.22), 可得 $Q_{p-1}=Q_7=6.9591$, 给定显著性水平 $\alpha=0.05$, 对应 $\chi^2_{0.05}(7)=14.07$, $\chi^2_{0.05}(7)>Q_{p-1}$, 因此接受 $n=3$ 作为系统真实的变量个数; 第 3 步, 固定 $n=3$, 利用前向/后向算法辨识系统的作用变量, 重复 100 次数值模拟, 每次试验的数据长度均为 500, 每次试验的辨识结果均为 $x_3(k)$、$x_4(k)$ 和 $x_5(k)$; 第 4 步, 辨识结果的验证: 由前四步的结果可知, 存在函数 $g(\cdot)$ 使得 $y(k)=f(x_1(k),x_2(k),x_3(k),x_4(k),x_5(k),x_6(k),x_7(k),x_8(k))\approx g(x_3(k),x_4(k),x_5(k))$. 选取一组新的输入信号

$$x_3(k)=0.8*\sin\left(\frac{2\pi k}{20}\right),\quad x_4(k)=0.8*\sin\left(\frac{2\pi k}{20}\right)\cdot\cos\left(\frac{2\pi k}{20}\right),\\x_5(k)=0.8*\cos\left(\frac{2\pi k}{20}\right),\quad k=1,\ldots,40.$$

利用前 500 组数据作系统辨识, 后 40 组数据作验证. 具体地讲, 分别假设系统的变量个数为 8 和 3, 利用前 500 组数据分别估计函数 f 和 g, 设估计值为 $\hat{f}$ 和 $\hat{g}$, 然后利用后 40 组数据计算系统输出的估计值

$$\hat{y}_1(k)=\hat{f}(x_1(k),x_2(k),x_3(k),x_4(k),x_5(k),x_6(k),x_7(k),x_8(k)),\quad k=1,\ldots,40,\\\hat{y}_2(k)=\hat{g}(x_3(k),x_4(k),x_5(k)),\quad k=1,\ldots,40,$$

并计算估计值的拟合度 GoF (GoF $=0.9220$, $n=3$; GoF $=0.5189$, $n=8$). 从图 6 可见, 舍去次要变量、利用最主要的作用变量 $x_3(k)$、$x_4(k)$ 和 $x_5(k)$ 可以使系统辨识的精度更高.

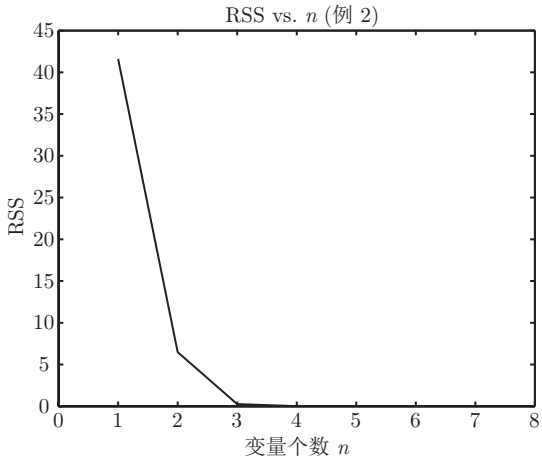

图 5　残差与变量个数的关系

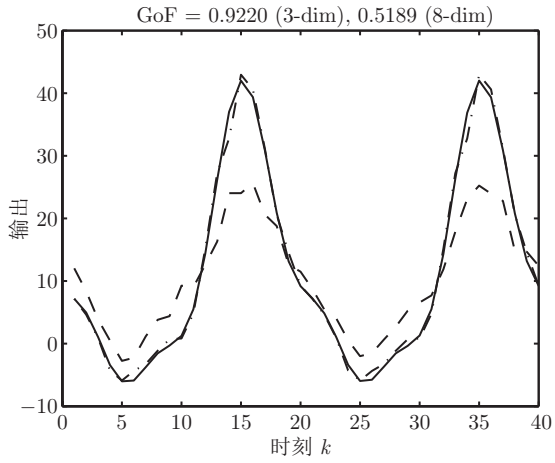

图 6　变量个数估计模型的输出与真实输出的对比

6　小结与讨论

本文针对非参数非线性系统, 从局部变量选择和全局变量选择等不同角度给出了算法和主要结论. 第 2、3.1 和 3.2、4 节的详细理论推导可参见文献 [39, 46, 47], 第 3.3 小节的内容不见于其他文献. 非线性系统变量选择的研究还处于起步阶段, 有许多问题值得深入探讨, 如闭环情形下的相关问题等.

综上, 系统的变量选择对于模型简化和提高辨识精度等有着重要作用, 与机器学习和稀疏表示等交叉学科联系密切. 本文介绍了我们近期在这个问题上的研究进展, 由于学识所限, 难免存在疏漏与偏颇, 恳请专家同行不吝赐教.

参考文献

1　Soderstrom T, Stoica P. System Identification. New York: Prentice Hall, 1989

2　Su S, Yang F. On the dynamical modeling with neural fuzzy networks. IEEE Trans Neural Netw, 2002, 13: 1548–1553

3　Bai E W, Liu Y. Recursive direct weight optimization in nonlinear system identification: A minimal probability approach. IEEE Trans Automat Control, 2007, 52: 1218–1231

4　Kocijan J, Girard A, Banko B, et al. Dynamic systems identification with Gaussian process. Math Comput Model Dyn Syst, 2003, 11: 411–424

5　Li K, Peng J. Neuro input selection—a fast model based approach. Neurocomputing, 2007, 70: 762–769

6　Li K, Peng J, Bai E W. A two-stage algorithm for identification of nonlinear dynamic systems. Automatica, 2006, 42: 1187–1196

7　Ohlsson H, Roll J, Glad T, et al. Using manifold learning for nonlinear system identification. In: Proc of the 7th IFAC Symposium on Nonlinear Control Systems. Pretoria: IFAC, 2007, 170–175

8　Pillonetto G, Quang M, Chiuso A. A new kernel-based approach for nonlinear system identification. IEEE Trans Automat Control, 2011, 56: 2825–2840

9　Roll J, Nazin A, Ljung L. Nonlinear system identification via direct weight optimization. Automatica, 2005, 41: 475–490

10　Arribas-Gil A, Bertin K, Meza C, et al. LASSO-type estimators for semiparametric nonlinear mixed-effects model estimation. Stat Comput, 2013, 24: 443–460

11　Bonin M, Seghezza V, Piroddi L. LASSO-enhanced simulation error minimization method for NARX model selection. In: Proceedings of 2010 American Control Conference. Baltimore: IEEE, 2010, 4522–4527

12　Candes E, Tao T. Near optimal signal recovery from random projections: Universal encoding strategies. IEEE Trans Inform Theory, 2006, 52: 5406–5425

13 Chiuso A, Pillonetto G. Bayesian and nonparametric methods for system identification and model selection. In: Proceedings of European Control Conference. Strasbourg: EUCA, 2014, 2376–2381
14 Cox T, Cox M A. Multidimensional Scaling, 2nd ed. London: Chapman and Hall, 2000
15 Debruyne M, Hubert M, Suykens J A K. Model selection in kernel based regressioin using the influence function. J Mach Learn Res, 2008, 9: 2377–2400
16 Delvecchio D, Piroddi L. A nonlinear active noise control scheme with online model structure selection. In: Proceedings of 50th IEEE Conference on Decision and Control. Orlando: IEEE, 2011, 8014–8019
17 Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control, 1974, 19: 716–723
18 Bomerger J, Seborg D. Determination of model order for NARX models directly from input-output data. J Process Control, 1998, 8: 459–568
19 Cao L. Practical method for determining the minimum embedding dimension input-output models for nonlinear dynamic systems. In: Proceedings of 1997 American Control Conference. San Francisco: IEEE, 1997, 2520–2523
20 Chen H F, Guo L. Identification and Stochastic Adaptive Control. Boston: Birkhäuser, 1991
21 He X, Asada H. A new method for identify orders of input-output models for nonlinear dynamic systems. In: Proceedings of American Control Conference. San Francisco: IEEE, 2003, 2520–2523
22 Knight K, Fu W. Asymptotics for Lasso-type estimators. Ann Statist, 2000, 28: 1356–1378
23 Mao K, Billings S A. Variable selection in nonlinear system modeling. Mech Syst Signal Process, 2006, 13: 351–366
24 Morici S, Spiriti E, Piroddi L. An indirect model selection algorithm for nonlinear active noise control. In: Proceedings of 2013 European Control Conference. Zürich: EUCA, 2013, 2910–2915
25 Ojeda F, Falck T, De Moor B, et al. Polynomial componentwise LS-SVM: Fast variable selection using low rank updates. In: The 2010 International Joint Conference on Neural Networks. Barcelona: IEEE, 2010, 1–7
26 Pillonetto G, Chen T, Ljung L. Kernel-based model order selection for identification and prediction of linear dynamic systems. In: Proceedings of 52nd IEEE Conference on Decision and Control. Florence: IEEE, 2013, 5174–5179
27 Zou H. The adaptive Lasso and its oracle properties. J Amer Statist Assoc, 2006, 101: 1418–1429
28 Gatu C, Kontoghiorghes E. Branch and bound algorithms for computing the best subset regression models. J Comput Graph Statist, 2006, 15: 139–156
29 Hand D. Branch and bound in statistical data analysis. J R Stat Soc D, 1981, 30: 1–13
30 Hong X, Mitchell S, Chen C, et al. Model selection approaches for nonlinear system identification: A review. Internat J System Sci, 2008, 39: 925–949
31 Kennel M, Brown R, Abarbanel H. Determining embedding dimension for phase-space reconstruction using geometrical construction. Phys Rev A (3), 1992, 45: 3403–3411
32 Peduzzi P. A stepwise variable selection procedure for nonlinear regression methods. Biometrics, 1980, 36: 510–516
33 Peng J, Ferguson S, Rafferty K, et al. An efficient feature selection method for mobile devices with application to activity recognition. Neurocomputing, 2011, 74: 3543–3552
34 Roweis S, Saul L. Nonlinear dimensionality reduction by local linear embedding. Science, 2000, 290: 2323–2326
35 Mosci R, Rosasco L, Santoro M, et al. Is there sparsity beyond additive models? IFAC Proc Volumes, 2012, 45: 971–976
36 Box G E P, Pierce D. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Amer Statist Assoc, 1970, 65: 1509–1526
37 Wei H L, Billings S A. Feature subset selection and ranking for data dimensionality reduction. IEEE Trans Pattern Anal Mach Intell, 2007, 29: 162–166
38 Bai E W. Non-parametric nonlinear system identification: An asymptotic minimum mean squared error estimator. IEEE Trans Automat Control, 2010, 55: 1615–1626
39 Bai E W, Li K, Zhao W, et al. Kernel based approaches to local nonlinear non-parametric variable selection. Automatica, 2014, 50: 100–113
40 Zhao W X, Chen H F, Bai E W, et al. Kernel-based local order estimation of nonlinear non-parametric systems. Automatica, 2015, 51: 243–254
41 Lobato I N, Nankervis J, Savin N. Testing for zero autocorrelation in the presence of statistical dependence. Econometric Theory, 2002, 18: 730–743
42 Velasco C, Lobato I. A simple and general test for white noise. In: Econometric Society 2004 Latin American Meetings. Santiago: Econometric Society, 2004, 112–113
43 Breiman L, Friedman J. Estimating optimal transformation for multiple regression and correlation. J Amer Statist Assoc, 1985, 80: 580–619
44 Deaton A, Muellabaer J. Economics and Consumer Behavior. Cambridge: Cambridge University Press, 1980

45 Sockett E B, Daneman D, Clarson C, et al. Factors affecting and patterns of residual insulin secretion during the first year of type 1 (insulin-dependent) diabetes mellitus in children. Diabetologia, 1987, 30: 453–459
46 Mu B Q, Zheng W X, Bai E W. Variable selection and identification of high-dimensional nonparametric additive nonlinear systems. IEEE Trans Automat Control, doi: 10.1109/TAC.2016.2605741, 2016
47 Mammen E, Linton O, Nielsen J. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Ann Statist, 1999, 27: 1443–1490

Variable selection in nonlinear non-parametric system identification

BAI ErWei, LI Kang, ZHAO WenXiao, MU BiQiang & ZHENG WeiXing

Abstract This paper considers a problem of variable selection for a high dimensional nonlinear non-parametric system. Different algorithms are proposed for cases when the number of observed data increasing to infinity, being finite, and the nonlinear system being additive. The theoretical properties of the proposed algorithms are obtained, which are validated by simulation examples. The algorithms find the relationship between the input and output variables, and further the inter-dependence of input variables so that the importance of the input variables can be established.

Keywords **variable selection, nonlinear nonparametric system, additive nonlinear system, local contributing variable, global contributing variable**

MSC(2010) **93B30, 93E12, 93C10**

doi: 10.1360/N012016-00002