国 防 科 技 大 学 学 报
第 23 卷 第 2 期 JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY Vol.23 No.2 2001

文章编号：1001-2486（2001）02-0115-04

基于正交最小二乘估计的非线性时间序列的预测*

沈 辉，胡德文

（国防科技大学机电工程与自动化学院，湖南 长沙 410073）

摘 要：在对非线性时间序列的短期预测中经常采用局部线性化的预测算法，原有的算法使用普通最小二乘法（LS）估计近似线性模型的参数。对于存在噪声的数据，该算法的数值稳定性较差。本文在对非线性空间进行局部线性化的基础上，采用正交最小二乘方法（OLS）对线性模型同时进行结构选择与参数辨识，改善了数值的病态特性，增强了算法的稳定性。

关键词：非线性时间序列；预测；局部线性化；正交最小二乘估计

中图分类号：TP13 文献标识码：A

Nonlinear Time Series Prediction Based on Orthogonal Least Squares Algorithm

SHEN Hui，HU De-wen

（College of Mechatronics Engineering and Automation，National Univ. of Defense Technology，Changsha 410073，China）

Abstract：Local linear prediction is often applied to predict nonlinear time series，which uses the ordinary least square（LS）method to estimate the parameters in the approximated linear models. If there exists noise in the process，the computational stability of the method is rather poor. This paper presents an improved method that uses the orthogonal least square（OLS）algorithm to estimate both the structure and the parameters in the linear models from linearizing locally the whole nonlinear space. The proposed method can solve the ill－posed numerical problem to some extent and increase the stability of prediction algorithm.

Key words：nonlinear time series；prediction；local linearization；orthogonal least square（OLS）

在非线性时间序列的预测中经常采用局部线性化方法，基本思想是通过相空间重构[1]，根据时间序列将目标点周围的邻域线性化，由构造的线性模型得到下一个预测点。其优点是算法简单，不需要了解整个系统的动态行为，只需要根据待测点附近的样本值对下一时间进行预测。

然而，除了将非线性局部空间近似为线性空间所带来的误差，基于普通最小二乘估计（LS）的局部线性化方法[2]的固有弊病是估计的方差得不到控制。对于病态的数据矩阵，估计的方差可能变得很大，这样得到的预测值是不可信的。同时，随着相空间重构维数的增加，预测的误差迅速增大，算法将变得不稳定。而正交最小二乘法估计（OLS），不仅对模型中各个项的系数进行估计，而且也对模型中的结构进行分析，舍弃其中对输出影响很小的项，使构造的模型能更精确地反映实际系统的主要部分，因此在许多领域用来取代 LS 进行模型的估计。基于以上分析，本文研究了基于 OLS 估计的非线性时间序列的局部预测算法，数值仿真表明，采用 OLS 方法确实增强了算法的数值稳定性，从而扩大了算法的应用范围。

1 局部线性化算法的基本原理[3]

假设一非线性时间序列 $x_i = x\ (i\tau_s)$，$i = 1$，…，N。其中 τ_s 为采样时间，取 $\tau_s = 1$。由时间序列重构相空间[4]：

$$\widetilde{x}_i = [x_{i-(m-1)\tau}, x_{i-(m-2)\tau}, \cdots, x_t]^T, i = 1 + (m-1)\tau, \cdots, N \tag{1}$$

* 收稿日期：2000-09-02
基金项目：高等学校骨干教师基金；湖南省自然科学基金（00JJY2060）；模式识别国家重点实验室开放课题基金
作者简介：沈辉（1975-），男，博士生。

其中 m 为嵌入空间的维数，τ 为延迟时间。设训练段时间序列 Γ，长度为 n。目标点 $\tilde{x}_t \in \Gamma$，待预测点 x_{t+T} 可通过以下局部线性化方法进行预测：

找到距离目标点 $\tilde{x}_t$ 最近的 k 个向量：$\tilde{x}_{t(1)}$，$\tilde{x}_{t(2)}$，…，$\tilde{x}_{t(k)} \in \Gamma$，令：

$$\tilde{X} = [\tilde{x}_{t(1)}, \cdots, \tilde{x}_{t(k)}]^{\mathrm{T}} \in \mathrm{R}^{k \times m}, \tilde{y} = [x_{t(1)+T}, \cdots, x_{t(k)+T}]^{\mathrm{T}} \in \mathrm{R}^{k} \tag{2}$$

T 是需要预测的步长。局部线性化模型的建立是基于 $\tilde{X}$ 和 $\tilde{y}$ 的中心化：令 $\bar{x}$、$\bar{y}$ 分别是 $\tilde{X}$、$\tilde{y}$ 列向量的平均值，并且：

$$\begin{aligned} X &= \tilde{X} - \mathbf{1}\bar{x}^T, y = \tilde{y} - \mathbf{1}\bar{y}, \\ x_t &= \tilde{x}_t - \bar{x}, y_t = x_{t+T} - \bar{y} \end{aligned} \tag{3}$$

这里，1 表示元素均为 1 的向量。在目标点的局部邻域内建立以下线性模型：

$$y = X\mathbf{b} + \varepsilon \quad \mathrm{E}\{\varepsilon\} = 0, \mathrm{Var}\{\varepsilon\} = \sigma^2 I \tag{4}$$

这是一个线性回归方程，利用最小二乘法可以得到回归系数 b 的估计值 $\hat{b}$：

$$\hat{b}_{LS} = (X^T X)^{-1} X^T y \tag{5}$$

设 $\hat{y}_t$ 是 y_t 的预测值，则由 $\hat{y}_t = x_t^T \hat{b}$ 及（3）式可以得到预测值 $\hat{x}_{t+T}$。

2 基于正交最小二乘估计的预测算法

对于（4）式给出的模型，假设 $X^T X$ 为正定矩阵，于是，存在单位正交矩阵 Q，使得：$Q^T X^T X Q = \sum = \mathrm{diag}\{\lambda_1, \lambda_2, \Lambda\lambda_m\}$，。其中，$\lambda_1 \geqslant \lambda_2 \geqslant \Lambda \geqslant \lambda_m \geqslant 0$ 是 $X^T X$ 的奇异值。

忽略 X 阵的噪声影响，预测值 $\hat{y}_t$ 的均方误差为：

$$\mathrm{MSE}(\hat{y}_t) = \mathrm{E}\{y_t - \hat{y}_t\}^2 = (\mathrm{E}\{y_t\} - y_t)^2 + \mathrm{E}\{y_t - E\{y_t\}\}^2 \tag{6}$$

均方误差分解为偏差的平方项和方差项：

$$\mathrm{bias}(\hat{y}_t) = \mathrm{E}\{y_t\} - y_t = 0 \tag{7}$$

$$\mathrm{Var}(\hat{y}_t) = \mathrm{E}\{y_t - \mathrm{E}\{y_t\}\}^2 = \sigma^2 \sum_{i=1}^{m} \frac{1}{\lambda_i^2} \{(Q_i^T x_t)^2 + \lambda_i^2 (Q_i^T b)^2 + \sigma^2\} \tag{8}$$

可见，当 σ^2 确定时，$\mathrm{E}\| \hat{y}_t \|$ 与 $\| y_t \|^2$ 的偏差主要是由 $1/\lambda_i$ 引起的，当 $1/\lambda_i$ 很大时，最小二乘估计变得不可信。这在对非线性时间序列进行重构时经常遇到。

以下运用 OLS 方法对局部线性预测的算法进行改进：

将局部线性空间正交化，用等价的正交模型代替原来的线性模型，将 X 正交分解：

$$X = WA \tag{9}$$

其中 W 为正交矩阵。等价的辅助模型为：

$$y = \mathbb{W} \cdot g + \varepsilon, \quad g \in R^m, \mathbf{A}b = g \tag{10}$$

则 g_i 的估计值

$$\hat{g}_i = \frac{y^T w_i}{\| w_i \|^2}, \quad 1 \leqslant i \leqslant m \tag{11}$$

因此，正交向量 $\{w_i, i = 1, 2, \cdots, m\}$ 形成了由原始模型中的向量 $\{x_i, i = 1, 2, \cdots, m\}$ 构成的线性空间的 m 个正交基，即

$$\mathrm{span}\{w_1, w_2, \cdots, w_m\} = \mathrm{span}\{x_1, x_2, \cdots, x_m\} \tag{12}$$

从几何角度看，$\hat{g}_i \| w_i \|$ 可以作为观测向量 y_i 向新的正交空间的第 i 个坐标正交投影的幅值。任何观测向量 y_i 从物理角度都可以解释为新的向量空间中的互不相关的向量 $\{g_i w_i\}_{i=1}^m$ 与误差向量 ε_i 的组合。这 m 个向量的每一个都独立地贡献一部分能量 $\hat{g}_i \| w_i \|^2$ 到观测向量 $\| y_i \|^2$ 的能量中去。如果有的贡献的能量很少，甚至小于噪声的能量，其对观测向量的影响就可以忽略不计。通过相空间重构构造的线性模型中，一般只有某些回归因子起着关键的作用，今提取其中关键的回归因子，选择的标准采用误差下降率：

$$ERR_i \triangleq \frac{\hat{g}_i \| \boldsymbol{w}_i \|^2}{\| \boldsymbol{y}_i \|^2} \times 100\% \tag{13}$$

对于大的 ERR_i，表示该回归因子对观测向量有重要影响，因此应该包括在辅助模型中，相反，对于小的 ERR_i，则可以忽略。

为了由大到小地选择需要的回归因子，本文采用正交前向回归（OFR）方法[4]对 OLS 方法加以改进，使回归因子按照对观测向量的影响程度由大到小排列，可以从开始依次选择回归因子，直到满足需要的精度需要。通过回归因子的重新选取与排列，系数 g 成为 g'，矩阵 W，A 也相应地转化成为 W'和 A'。

假设提取的关键因子个数为 j，辅助模型成为：

$$\boldsymbol{y} = \boldsymbol{W}'\boldsymbol{g}' + \boldsymbol{\varepsilon}_i \tag{14}$$

目标点在新的正交基下的坐标为

$$\tilde{\boldsymbol{x}}'_t = \tilde{\boldsymbol{x}}_t \cdot (\boldsymbol{A}')^{-1} \tag{15}$$

相应的预测值：

$$\hat{y}_t = \tilde{\boldsymbol{x}}'_t \hat{\boldsymbol{g}}' \tag{16}$$

结合（3）与（16）式可以得到预测值 $\hat{x}_{t+T}$。

3 数值仿真

考虑 Lorenz 混沌系统[6]产生的时间序列，加入信噪比为 5% 的随机噪声。取 $N = 2000$，$n = 1500$，最后 500 个为考察点，第 T 步的预测误差用归一化的均方根误差衡量：

$$\mathrm{NRMSE} = \sqrt{\frac{(1/(N - T - n)) \sum_{t=n+1}^{N-T} (x_{t+T} - \hat{x}_{t+T})^2}{(1/N) \sum_{i=1}^{N} (x_i - \bar{x})^2}} \tag{17}$$

其中 N 为整个时间序列的长度，n 为训练段的点的数目，最后用于考察的点的数目为 $N - n$，$\bar{x}$是整个数据的样本均值。

图 1 显示的是在不同嵌入维数下单步预测的误差。图 1（a）与图 1（b）分别是 $k = 12$ 与 $k = 8$ 的情况。

图 1（a）由于 m 始终距离 k 较远，误差曲线变化缓慢，采用 OLS 方法的改进效果不大明显。图 1（b）的误差曲线开始变化缓慢，但当 $m > 5$ 时误差迅速上升，尤其当 m 接近 k 时，预测误差随着嵌入维数的增加迅速增大，反映了 LS 算法在高维情况下的不稳定性。而使用 OLS 方法的误差曲线变化始终比较平缓，表明该算法的稳定性大大增强。这在实际中有重要的意义。因为实际中由于有限的观测数据，或者某些混沌系统具有较大的嵌入维数，使得不可能要求 k 远大于 m，对算法改进后即使严格限制 k 的大小，也能通过选择适当的嵌入维数在合理的精度下对混沌数据进行预测。

图 2 显示在不同嵌入维数下的直接多步预测误差。图 2（a）和图 2（b）分别是在较低的嵌入维数和较高嵌入维数的情况。由于较低的嵌入维数对系统进行了较好的重构，采用两种方法的误差曲线十分接近，但当嵌入维数进一步增加，LS 方法的误差迅速增大，相比之下，OLS 方法具有更好的稳定性。

4 结论

本文针对非线性时序的预测提出一种基于正交最小二乘方法的局部预测算法。该方法采用正交最小二乘估计辨识目标点附近的近似线性空间，与局部线性化预测方法相比具有更好的算法稳定性。

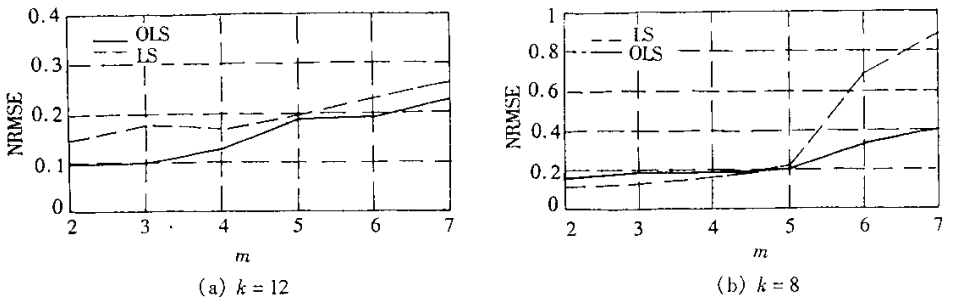

图1 Lorenz混沌时间序列的单步预测误差曲线

Fig.1 One-step predict for data generated from the Lorenz system

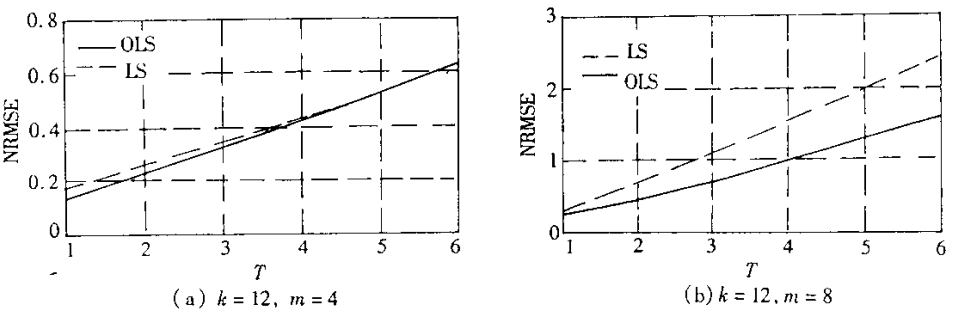

图2 Lorenz混沌时间序列的多步预测误差曲线

Fig.2 Multi-step predict for data generated from the Lorenz system

参考文献：

[1] Kugiumtzis D. State Space Reconstruction Parameters in the Analysis of Chaotic Time Series – the Role of the Time Window Length [J]. Physica D. 1996, 95: 13 – 28.

[2] Farmer J D, Sidorowich J J. Predicting Chaotic Time Series [J]. Phys. Rev. Lett. 1987, 59: 845 – 848.

[3] Kugiumtzis D, Lingjarde O C, Christophersen N. Regularized Local Linear Prediction of Chaotic Time Series [J]. Physica D. 1998, 112: 344 – 360.

[4] Casdagli M. Nonlinear Prediction of Chaotic Time Series [J]. Physica D. 1989, 35: 335 – 356.

[5] Navone H D, Ceccatto H A. Forecasting Chaos From Small Data Set: A Comparison of Different Nonlinear Algorithms [J]. J. Phys. A. 1995, 28 (12): 3381 – 3388.

[6] Lorenz E N. Deterministic Nonperiodic Flow [J]. J. Atmos. Sci. 1991, 20: 579 – 616.