Article ID: 1000-8152(2001)06-0949-05

Convergence of Hierarchical Stochastic Gradient Identification for Transfer Function Matrix Model*

DING Feng, YANG Jiaben and XU Yongmao

(Department of Automation, Tsinghua University · Beijing, 100084, P.R.China)

Abstract: The hierarchical identification principle is stated, and the hierarchical stochastic gradient (HSG) algorithm for the transfer function matrix (TFM) model for multivariable systems is presented. In the hierarchical identification, the system parameters are divided into the parameter vector, which includes the coefficients of the characteristic polynomial of the system, and the parameter matrix, which includes the coefficients of the numerators of the TFM polynomials, respectively. The convergence analysis, using martingale hyperconvergence theorem, shows that the parameter estimation error (PEE) given by the HSG algorithm is consistently bounded, and that PEE consistently converges to zero under the persistent excitation condition. Hierarchical identification has a small amount of calculation and is easy to be realized.

Key words: identification; hierarchical identification; multivariable system; parameter estimation

Document code: A

传递函数阵递阶随机梯度辨识方法的收敛性分析

丁 锋 杨家本 徐用懋

(清华大学自动化系·北京,100084)

摘要: 阐述了递阶辨识原理,提出了传递函数阵模型参数的递阶随机梯度(HSG)辨识方法.在递阶辨识中,系统参数被分解为参数向量和参数矩阵.前者是由系统的特征多项式的系数构成的,后者是由传递函数矩阵分子多项式的系数构成的.借助于鞅超收敛定理的收敛性分析表明,HSG算法的参数估计误差一致有界;当持续激励条件成立时,参数估计误差一致收敛于零.递阶辨识方法具有计算量小和容易实现等特点.

关键词: 辨识; 递阶辨识; 多变量系统; 参数估计

1 Introduction

Reducing the large computational effort required by previous identification algorithms for multivariable systems is one of the most difficult projects to be solved in identification area. One scheme is to develop identification algorithms which require less computation[1]. For example, the combined identification methods simultaneously to estimate all the parameters of the whole multivariable system[2,3] rather than to estimate the parameters of each subsystem of a multivariable system[4,5], the multi-innovation identification algorithm which does not require matrix inversion[6], the hierarchical identification algorithm for large-scale systems[7], the hierarchical least squares algorithm for the transfer function matrix model, and the hierarchical stochastic gradient algorithm in this paper.

The basic principle of hierarchical identification is that, at first, a system is decomposed into some subsystems with smaller dimension and fewer variables, then the parameters of each subsystem are estimated respectively. However, there exist associated items between the sub-systems, i.e., the ith subsystem includes the unknown parameters of other subsystems. So, this involves very difficult iterative calculations. In order to solve this problem, when computing the parameter estimates of the ith subsystem at time t, the unknown parameters of other subsystems are replaced with their estimates at time $(t-1)$.

The hierarchical identification for TFM is that the parameters of the system are divided into a parameter ve-

* Foundation item: supported by the National Natural Science Foundation of China (NSFC) (60074029, 69934010), and the Foundation of Information College, Tsinghua University, and CIMS Project for Fujian Petrochemical Corporation Ltd.

Received date: 1999-07-19; Revised date: 2000-12-28.

ctor and a parameter matrix, and then they are estimated, respectively. The parameter vector consists of the coefficients of the characteristic polynomial of the system, and the parameter matrix consists of the coefficients of the numerators of the TFM polynomials.

The hierarchical identification algorithms require less computational burden than Sen and Sinha's algorithm[8], but its convergence analysis is more difficult. In this paper, the convergence of the HSG algorithm is studied by using martingale hyperconvergence theorem, but the convergence of the hierarchical least squares algorithm in Ref. [9] will still be difficult to prove.

2 Hierarchical identification for the TFM model

Consider the multi-input multi-output stochastic system described by the TFM model[9]

$$A(z)y(t) = \begin{bmatrix} B_{11}(z) & B_{12}(z) & \cdots & B_{1r}(z) \\ B_{21}(z) & B_{22}(z) & \cdots & B_{2r}(z) \\ \vdots & \vdots & \vdots & \vdots \\ B_{m1}(z) & B_{m2}(z) & \cdots & B_{mr}(z) \end{bmatrix} u(t) + w(t), \tag{1}$$

where $u(t) = [u_1(t), u_2(t), \cdots, u_r(t)]^{\mathrm{T}} \in \mathbb{R}^r$ is the system input vector, $y(t) = [y_1(t), y_2(t), \cdots, y_m(t)]^{\mathrm{T}} \in \mathbb{R}^m$ is the system output vector, z^{-1} represents the unit delay operator, i.e., $z^{-1}y(t) = y(t-1)$, $zy(t) = y(t+1)$. $w(t) \in \mathbb{R}^m$ is a stochastic noise vector with zero mean, $A(z)$ is the monic characteristic polynomial of the system (of degree n) defined as the least common denominator of all entries of the transfer function matrix of the system, and

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n},$$

$$B_{ij}(z) = \beta_{ij}(1) z^{-1} + \beta_{ij}(2) z^{-2} + \cdots + \beta_{ij}(n) z^{-n}.$$

The number of the parameters $(a_i, \beta_{ij}(k))$ to be identified in model (1) is equal to $S_1 = n(mr+1)$.

The sequence $\{w(t)\}$ is assumed to be a martingale difference sequence defined on a probability space (Ω, F, P) and adapted to the sequence of nondecreasing sub-sigma algebra $\{F_t, t \in \mathbb{N}\}$ where $\{F_t\}$ is generated by the observations up to and including time t, i.e. $F_t = \sigma(y(t), u(t), y(t-1) \cdots, u(0))$ and F_0 is assumed to contain all initial condition information. The sequence $\{w(t)\}$ satisfies the following noise assumptions:

A1) $E[w(t) \mid F_{t-1}] = 0$, a.s..

A2) $E[\| w(t) \|^2 \mid F_{t-1}] = \sigma_w^2(t) \leqslant \sigma_w^2 < \infty$, a.s..

A3) $\limsup_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \| w(i) \|^2 \leqslant \sigma_w^2 < \infty$, a.s..

where the norm of the matrix X is defined by $\| X \|^2 = \mathrm{tr}[X X^{\mathrm{T}}]$.

Eq.(1) can be expressed as

$$A(z)y(t) = B(z)u(t) + w(t), \tag{2}$$

where

$$B(z) = B_1 z^{-1} + B_2 z^{-2} + \cdots + B_n z^{-n},\ B_i \in \mathbb{R}^{m\times r}.$$

In vector form, Eq. (2) may be written as

$$y(t) + \psi(t) a = \theta^{\mathrm{T}} \varphi(t) + w(t), \tag{3}$$

where

$$\psi(t) = [y(t-1), y(t-2), \cdots, y(t-n)] \in \mathbb{R}^{m\times n},$$

$$\theta^{\mathrm{T}} = [B_1, B_2, \cdots, B_n] \in \mathbb{R}^{m\times(nr)},$$

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^n,\ \varphi(t) = \begin{bmatrix} u(t-1) \\ u(t-2) \\ \vdots \\ u(t-n) \end{bmatrix} \in \mathbb{R}^{nr}.$$

Let $Y(t) \triangleq y(t) - \theta^{\mathrm{T}}\varphi(t)$ and $Z(t) \triangleq y(t) + \psi(t)a$, then system (3) may be decomposed into the following two imaginary subsystems

$$\text{S1} \qquad Y(t) = -\psi(t)a + w(t), \tag{4}$$

$$\text{S2} \qquad Z(t) = \theta^{\mathrm{T}}\varphi(t) + w(t), \tag{5}$$

$Y(t) \in \mathbb{R}^m$, $\psi(t) \in \mathbb{R}^{m\times n}$ and $a \in \mathbb{R}^n$ in Eq. (4) may be regarded as the output vector, information matrix and parameter vector of system S1. In the same way $Z(t) \in \mathbb{R}^m$, $\varphi(t) \in \mathbb{R}^{nr}$ and $\theta^{\mathrm{T}} \in \mathbb{R}^{m\times(nr)}$ in Eq.(5) may be regarded as the output vector, information vector and parameter matrix of subsystem S2.

According to the least squares principle, the least squares estimates of the parameter vector a and the parameter matrix θ may be obtained from

$$\hat{a}(t) = \hat{a}(t-1) + L_1(t)[Y(t) + \psi(t)\hat{a}(t-1)],$$

$$\hat{\theta}(t) = \hat{\theta}(t-1) + L_2(t)[Z^{\mathrm{T}}(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1)],$$

where $\hat{a}(t)$ and $\hat{\theta}(t)$ are the estimates of a and θ at time t, and $L_1(t)$ and $L_2(t)$ are a gain matrix and a gain vector.

Since $Y(t)$ and $Z(t)$ contain the unknown parameter matrix θ and unknown parameter vector a, it is impossible to realize the algorithm. The problem can be solved using the hierarchical identification/control principle for large-scale system[7,10], and these unknown variables may be replaced with their corresponding estimates $\hat{\theta}$ and $\hat{a}$ at time $(t-1)$. The result is the hierarchical stochastic gradient identification algorithm of estimating the parameters for the TFM model:

$$\hat{a}(t) = \hat{a}(t-1) - \frac{\psi^{\mathrm{T}}(t)}{r(t)}[y(t) + \psi(t)\hat{a}(t-1) - \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t)], \quad (6)$$

$$\hat{\theta}(t) = \hat{\theta}(t-1) - \frac{\psi(t)}{r(t)}[y^{\mathrm{T}}(t) + (\psi(t)\hat{a}(t-1))^{\mathrm{T}} - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1)], \quad (7)$$

$$r(t) = r(t-1) + \|\psi(t)\|^2 + \|\varphi(t)\|^2,\ r(0) = 1, \quad (8)$$

where I_m represents an $m \times m$ identity matrix. The initial values of the HSG algorithm may be chosen as $\hat{a}(0)$ = a small real vector (10^{-4}), $\hat{\theta}(0)$ = a small real matrix (10^{-4}).

3 Convergence of the HSG algorithm

Lemma 1 Assume that the vector $x(t) \in \mathbb{R}^n$ and the vector $\phi(t) \in \mathbb{R}^n$ satisfy the following equations:

$$\phi^{\mathrm{T}}(t)x(t) = 0, \text{ for } t \to \infty$$

and

$$\lim_{t\to\infty}[x(t) - x(t-k)] = 0, \text{ for any } 0 < k < \infty,$$

and that the vector $\phi(t)$ is sufficiently rich (persistently excited), i.e. there exist constants $0 < \alpha \leqslant \beta < \infty$ and an integer $N \geqslant n$ such that for any $t > 0$, the following inequalities hold:

$$\text{(A4)}\ \alpha I \leqslant \frac{1}{N}\sum_{i=1}^{N}\phi(t+i)\phi^{\mathrm{T}}(t+i) \leqslant \beta I, \text{ a.s.},$$

then

$$\lim_{t\to\infty} x(t) = 0.$$

Proof Let $\epsilon(t+k) = x(t+k) - x(t)$ or $x(t+k) = x(t) + \epsilon(t+k)$, it is obvious that $\epsilon(t+k)$ converges to zero, i.e. $\lim_{t\to\infty}\epsilon(t) = 0$. In the same way, let $\epsilon_1(t) = \phi^{\mathrm{T}}(t)x(t)$, we have $\lim_{t\to\infty}\epsilon_1(t) = 0$. So

$$\phi^{\mathrm{T}}(t+i)x(t+i) = \epsilon_1(t+i),$$

or

$$\phi^{\mathrm{T}}(t+i)x(t) = -\phi^{\mathrm{T}}(t+i)\epsilon(t+i) + \epsilon_1(t+i).$$

After taking the norm $\| * \|^2$ of both sides of the above equation, the summation from $i = 1$ to $i = N$ is

$$x^{\mathrm{T}}(t)\left[\sum_{i=1}^{i=N}\phi(t+i)\phi^{\mathrm{T}}(t+i)\right]x(t) = \sum_{i=1}^{i=N}\| -\phi^{\mathrm{T}}(t+i)\epsilon(t+i) + \epsilon_1(t+i)\|^2 \leqslant 2\sum_{i=1}^{i=N}[\|\phi(t+i)\|^2\epsilon^2(t+i) + \epsilon_1^2(t+i)].$$

Taking the trace of Condition (A4) will lead to $\|\phi(t)\|^2 \leqslant M \triangleq nN\beta < \infty$, and using Condition (A4), we have

$$0 \leqslant N\alpha\|x(t)\|^2 \leqslant 2\sum_{i=1}^{i=N}[M\epsilon^2(t+i) + \epsilon_1^2(t+i)].$$

Taking the limit of both sides of the above inequality will obtain the conclusion of Lemma 1 according to limited existence criterion.

Theorem 1 For the multivariable system (3) and the HSG algorithm (6) ~ (8), if Assumptions (A1) ~ (A3) hold, and $\sum_{t=1}^{\infty} r^{-1}(t) = \infty$, then the parameter estimation error given by the HSG algorihtm is consistently bounded, i.e.

$$\lim_{t\to\infty}\|\hat{a}(t) - a\|^2 + \|\hat{\theta}(t) - \theta\|^2 < \infty, \text{a.s.}.$$

Proof Define the parameter estimation error vector $\tilde{a}(t)$ and the parameter estimation error matrix $\tilde{\theta}(t)$ as

$$\tilde{a}(t) \triangleq \hat{a}(t) - a, \quad (9)$$

$$\tilde{\theta}(t) \triangleq \hat{\theta}(t) - \theta, \quad (10)$$

Substituting Eqs. (3) and (6) into Eq. (9) yields

$$\tilde{a}(t) = \tilde{a}(t-1) - \frac{\psi^{\mathrm{T}}(t)}{r(t)}[\xi(t) - \eta(t) + w(t)], \quad (11)$$

where

$$\xi(t) = \psi(t)\hat{a}(t-1) - \psi(t)a = \psi(t)\tilde{a}(t-1), \quad (12)$$

$$\eta(t) = \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t) - \theta^{\mathrm{T}}\varphi(t) = \tilde{\theta}^{\mathrm{T}}(t-1)\varphi(t). \quad (13)$$

Substituting Eqs. (3) and (7) into Eq. (10) yields

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) + \frac{\varphi(t)}{r(t)}[\xi(t) - \eta(t) + w(t)]^{\mathrm{T}}. \quad (14)$$

Define the stochastic Lyapunov function as

$$T(t) \triangleq \|\tilde{a}(t)\|^2 + \|\tilde{\theta}(t)\|^2. \quad (15)$$

Substituting Eqs. (11) and (14) into Eq. (15), we have

$$T(t) = T(t-1) - \frac{2}{r(t)}[\|\xi(t) - \eta(t)\|^2 + (\xi(t) - \eta(t))^{\mathrm{T}}w(t)] + [\xi(t) - \eta(t) + w(t)]^{\mathrm{T}} \cdot \frac{\psi(t)\psi^{\mathrm{T}}(t) + \|\varphi(t)\|^2 I_m}{r^2(t)}[\xi(t) - \eta(t) + w(t)] \leqslant$$

$$T(t-1) - \frac{2}{r(t)}\|\xi(t) - \eta(t)\|^2 - \frac{2}{r(t)}(\xi(t) - \eta(t))^{\mathrm{T}}w(t) + \frac{\|\psi(t)\|^2 + \|\psi(t)\|^2}{r^2(t)}.$$

$$[\|\xi(t)-\eta(t)\|^2+\|w(t)\|^2]+$$
$$2[\xi(t)-\eta(t)]^{\mathrm{T}}\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r^2(t)}w(t). \tag{16}$$

Since $\xi(t)-\eta(t), \psi(t), \varphi(t)$, and $r(t)$ are uncorrelated to $w(t)$ and are F_{t-1}-measurable, taking the conditional expectation of both sides of Eq. (16) with respect to F_{t-1} and using Assumptions A1) ~ A3) gives

$$\begin{aligned}
&\mathrm{E}[T(t)\mid F_{t-1}]\leqslant\\
&T(t-1)-\frac{2}{r(t)}\|\xi(t)-\eta(t)\|^2+\\
&\frac{\|\psi(t)\|^2+\|\varphi(t)\|^2}{r^2(t)}\|\xi(t)-\eta(t)\|^2+\\
&\frac{\|\psi(t)\|^2+\|\varphi(t)\|^2}{r^2(t)}\sigma_w^2(t)\leqslant\\
&T(t-1)-\frac{r(t)+r(t-1)}{r^2(t)}\|\xi(t)-\\
&\eta(t)\|^2+\frac{\|\psi(t)\|^2+\|\varphi(t)\|^2}{r^2(t)}\sigma_w^2,
\end{aligned}$$

or

$$\begin{aligned}
&E[T(t)\mid F_{t-1}]-T(t-1)\leqslant\\
&-\frac{1}{r(t)}\|\xi(t)-\eta(t)\|^2+\\
&\frac{\|\psi(t)\|^2+\|\varphi(t)\|^2}{r^2(t)}\sigma_w^2\triangleq -b(t).
\end{aligned}\tag{17}$$

Consider the set

$$R_t=[(\tilde{a}(t),\tilde{\theta}(t)):\|\xi(t)-\eta(t)\|^2\leqslant \frac{\|\psi(t)\|^2+\|\varphi(t)\|^2}{r^2(t)}\sigma_w^2,\mathrm{a.s.}].$$

A similar derivation to Ref. [11] and applying martingale hyperconvergence theorem [12] to (17) show that $T(t)$ converges to a bounded random variable T_0 a.s., and $(\tilde{a}(t),\tilde{\theta}(t))\in R_t$ for large t.

If $r(t)\to\infty$ and $\|\psi(t)\|^2+\|\varphi(t)\|^2<\infty$, then the following relationship holds:

$$\lim_{t\to\infty}(\tilde{a}(t),\tilde{\theta}(t))\in R_\infty=$$
$$\lim_{t\to\infty}[(\tilde{a}(t),\tilde{\theta}(t)):\|\xi(t)-\eta(t)\|^2=0,\mathrm{a.s.}]. \tag{18}$$

This completes the proof of Theorem 1.

Theorem 2 For the multivariable system (3) and the HSG algorithm (6) ~ (8), if the conditions of Theorem 1 hold, and the vector $\phi_i(t)\triangleq\begin{bmatrix}\psi_i^{\mathrm{T}}(t)\\ \varphi(t)\end{bmatrix}(i=1,2,\cdots,m)$ is sufficiently rich, $\psi_i(t)$ is the ith row of $\psi(t)$; then the parameter estimation error given by the HSG algorithm consistently converges to zero, i.e.

$$\lim_{t\to\infty}\|\hat{a}(t)-a\|^2+\|\hat{\theta}(t)-\theta\|^2=0,\ \mathrm{a.s.}.$$

Proof Since $\phi_i(t)(i=1,2,\cdots,m)$ is sufficiently rich, then

$$\lim_{t\to\infty}r(t)=\lim_{t\to\infty}O(t)=\infty. \tag{19}$$

From (18), we have

$$\xi(t)=\eta(t),\ \text{for } t\to\infty,$$

or

$$\psi(t)\tilde{a}(t-1)=\tilde{\theta}^{\mathrm{T}}(t-1)\varphi(t),\ \text{for } t\to\infty. \tag{20}$$

Let $\tilde{\theta}_i^{\mathrm{T}}(t-1)$ represent the ith row of $\tilde{\theta}^{\mathrm{T}}(t-1)$, and

$$x_i(t)\triangleq\begin{bmatrix}\tilde{a}(t-1)\\ -\tilde{\theta}_i(t-1)\end{bmatrix},$$

then (20) may be decomposed into the following m equations:

$$\phi_i^{\mathrm{T}}(t)x_i(t)=0,\ i=1,2,\cdots,m,\ \text{for } t\to\infty. \tag{21}$$

From (A3), (18), (19), (11) and (14), we may obtain

$$\lim_{t\to\infty}[x_i(t)-x_i(t-k)]=0,\ \text{for any } 0<k<\infty. \tag{22}$$

From (21) and (22), it is not difficult to reach the conclusions of Theorem 2 by using Lemma 1.

References

[1] Ding Feng and Xie Xinmin. Recursive estimation: for transfer function matrix subsubmodels auxiliary model method [J]. Control and Decision, 1991, 6(6):447 - 452 (in Chinese)

[2] Ding Feng and Xie Xinmin. Combined identification algorithm for linear multivariable system [J]. Control Theory and Applications, 1992,9(5):545 - 550(in Chinese)

[3] Ding Feng. Convergence analysis of auxiliary model identification for multivariable system [J]. Control Theory and Applications, 1997,14(2):192 - 200 (in Chinese)

[4] El-Sherief H. Parametric identification of a state-space model of multivariable system using the extended least-squares method [J]. IEEE Trans. SMC, 1981,11(3):223 - 227

[5] Sinha N K and Kwong Y H. Recursive estimation of the parameters of linear multivariable system [J]. Automatica, 1979,15(4):471 - 475

[6] Ding Feng, Xie Xinmin and Fang Chongzhi. Mutlti-innovation identification method for time-varying systems [J]. Acta Automatica, Sinica, 1996,22(1):85 - 91

[7] Ding Feng and Yang Jiaben. Hierarchical identification for large scale systems [J]. Acta Automatica, Sinica, 1999,25(5):647 - 654 (in Chinese)

[8] Sen A and Sinha N K. On-line estimation of the parameters of a multivariable system using matrix pseudoinverse [J]. Int. J. Syst. Sci., 1976,7(4):461 – 471

[9] Wang Zhixiang and Ding Feng. Parameter estimation algorithms of the main models and submodels for transfer function matrices [J]. Control and Decision, 1995,10(4):311 – 316 (in Chinese)

[10] Xi Yugeng. Introduction of Dynamical Large Scale Systems [M]. Beijing: National Defense Industry press, 1988 (in Chinese)

[11] Ding Feng, Yang Jiaben and Xu Yongmao. Least squares identification of generalized time-varying systems [J]. Journal of Tsinghua University, 2000,40(3):86 – 89 (in Chinese)

[12] Ding Feng and Yang Jiaben. Remarks on martingale hyperconvergence theorem and the convergence analysis of the forgetting factor least squares algorithms [J]. Control Theory and Applications, 1999,16(4):569 – 572 (in Chinese)

本文作者简介

丁　锋　见本刊2001年第3期第437页.

杨家本　见本刊2001年第3期第437页.

徐用懋　女.1935年生.1958年毕业于清华大学动力系.现任清华大学自动化系教授,博士生导师.学术方向为控制理论与控制工程,尤其是工业过程的建模、控制与优化.

(Continued from page 948)

backpropagation updating law can be used to train the weights of the proposed neural networks. Simulation results of different systems have demonstrated the feasibility of the proposed methods.

References

[1] Elman J L. Finding structure in time [J]. Cognitive Science, 1990, 14(2):179 – 211

[2] Scott G M and Ray W H. Creating efficient nonlinear network process models that allow model interpretation [J]. J. Process Control, 1993,3(3):163 – 178

[3] Pham D T and Oh S J. A recurrent backpropagation neural network for dynamical system identification [J]. Journal of Systems Engineering, 1992,2(4):213 – 223

[4] Funahashi K. On the approximate realization of continuous mappings by neural networks [J]. Neural Networks,1989,2(3):183 – 192

[5] Hirsch M W and Smale S. Differential equations, dynamical systems and linear algebra [M]. San Diego: Academic Press,1974

[6] Sales K R and Billings S A. Self-tuning control of nonlinear ARMAX models [J]. Int. J. Control, 1990,51(4):753 – 769

本文作者简介

任雪梅　1967年生.副教授.1989年毕业于山东大学自动控制专业,获学士学位.1992年在北京航空航天大学控制理论与应用专业获硕士学位,1995年获该专业的博士学位.同年分配到北京理工大学自动控制系工作.主要研究兴趣为神经网络,学习控制,火控系统.