利用自回归模型的平稳时序数据快速辨识算法*

黄雄波[1]，胡永健[2]
（1. 佛山职业技术学院 电子信息系，广东 佛山 528000；2. 华南理工大学 电子与信息学院，广州 510641）

摘 要：由于自回归模型的参数估计可归结为求解一个线性方程组的问题，所以其在平稳时序数据的辨识过程中具有广泛的应用场合。提出了一种基于自回归模型的快速辨识算法，以递推的方式对平稳时序数据自相关函数矩阵的秩的下界值进行估计，再以该估计值作为自回归模型的起始阶数对系统进行依次的递阶辨识。最后，基于 F 检验对相邻阶次的拟合误差的变化趋势进行显著性检验，并以检验结果作为算法的结束条件。新算法在保证较高辨识精度的条件下，其计算效能及辨识精度的稳定性均优于现有的自回归模型辨识算法，实验结果验证了新算法的有效性和先进性。

关键词：平稳时序数据；自回归模型；递阶辨识；自相关函数
中图分类号：TP311.11；TP301.6 **文献标志码**：A **文章编号**：1001-3695(2018)09-2643-05
doi:10.3969/j.issn.1001-3695.2018.09.019

Fast identification algorithm for stationary time series data using auto-regressive model

Huang Xiongbo[1], Hu Yongjian[2]
(1. *Dept. of Electronic & Information Engineering, Foshan Professional Technical College, Foshan Guangdong* 528000, *China*; 2. *School of Electronic & Information Engineering, South China University of Technology, Guangzhou* 510641, *China*)

Abstract: Because the parameter estimation of auto-regressive model can be reduced to the solution of a linear equation group, it has a wide range of applications in the identification of stationary time series data. This paper proposed a fast identification algorithm based on auto-regression model firstly, bound by recursion on the stationary time series data auto-correlation matrix rank estimation, then, estimated value as the starting order regression model in turn hierarchical identification, finally, the fitting error of adjacent F checked the order of significance test based on the test results, and as the end condition of algorithm. With the auto-regressive model identification algorithm compared with the existing, because the new algorithm avoided in the hierarchical identification from the first order, and thus it obtained better computing performance. The experimental results verify the effectiveness and superiority of the new algorithm.

Key words: stationary time series data; auto-regressive model; hierarchical identification; auto-correlation function

0 引言

对自然界各种物理现象、社会经济行为以及机械设备等事物的发展和运行情况进行观测，往往可以得到一系列伴随着时间变化而变化的时序数据，通过对这些数据进行适当的加工处理，便可透过事物的自身现象，掌握其内在规律和有关成因，进而对其可能的发展趋势进行预测。将时序数据[1]视为某一随机过程，若该过程的均值和方差都是与时间 t 无关的常数，且其自协方差函数也仅与时间差有关，则称该时序数据为平稳（广义平稳或弱平稳）时序数据，否则，称其为非平稳时序数据。目前，关于平稳时序数据的辨识建模的研究成果已十分丰富，相对地，非平稳时序数据至今还没有形成系统的分析方法[2]。事实上，在实际的应用中，非平稳时序数据更多的是通过差分或分段等方法转换为平稳时序数据后再进行辨识处理。例如，邹柏贤等人[3]利用方差分析法对网络通信系统的非单播数据包序列进行趋势和周期成分的剔除，并基于自回归滑动平均模型对余下的平稳时序数据进行辨识，从而得到一种高精度的网络流量过载的预警算法；程浩等人[4]利用牛顿迭代法对非平稳时序数据的自相关函数逼近模型进行优化求解，并设计实现了一种分段平稳时序数据的划分算法，该算法在模糊图像的处理中得到了令人满意的恢复效果；黄雄波[5]从平稳时序数据的定义出发，构造了具有递推机制的均值、方差及自相关函数的突变点的检测算法，在此基础上，实现了一种高效的局部平稳时序数据的析出算法。

对于平稳时序数据而言，其线性辨识模型主要有三种，即自回归（auto-regressive，AR）模型、滑动平均（moving average，MA）模型和自回归滑动平均（auto-regressive moving average，ARMA）模型，这些模型可以相互转换，又由于自回归模型的参数估计可转换为求解一个线性方程组的问题，故其在实际中有着更广泛的应用场合。系统辨识的主要任务有两个，即辨识模型阶次的确定和辨识模型参数的估计，以自回归模型为例，系统辨识的过程就是寻找最佳的阶次 $p(1\leqslant p, p\in \mathrm{N})$ 及计算相应的参数，使得在整体辨识误差为最小的情况下，时序数据的各期数据均可由 p 期过去的数据按照相应的参数线性组合而成。自回归模型的经典求解算法有两种[6]：Levinson-Durbin 算法和 Burg 算法。由于这两种算法均为递推算法，故它们都具有良

收稿日期：2017-05-29；**修回日期**：2017-07-09 **基金项目**：广东省自然科学基金团队项目(9351064101000003)；广东省应用型科技研发专项基金资助项目(2015B010130003)；广东省科技计划工业攻关项目(2011B010200031)；佛山职业技术学院校级重点科研项目(2015KY006)；佛山职业技术学院横向重点资助项目(H201813)

作者简介：黄雄波(1975-)，男，广东南海人，副教授，博士研究生，主要研究方向为时间序列分析及数字图像处理(xiongbo_h75@126.com)；胡永健(1962-)，男，湖北武汉人，教授，博导，主要研究方向为多媒体信息安全、信息隐藏和数字图像处理.

好的计算效能,但在不同的样本数量或高阶模型中,上述算法的辨识精度却不具鲁棒性[7]。例如,Levinson-Durbin 算法在小样本序列或高阶模型的场合中,其辨识精度通常会出现明显的下滑[8,9];而 Burg 算法的辨识精度也随着模型阶数的增加而下降[10]。为改善这些经典求解算法的鲁棒辨识精度,周毅等人[11]借助数据乘积矩阵的分块求逆原理,给出了一种计算等价 AR 模型参数估计和相应准则函数的依阶次递增的递推算法,然而,该算法在提升鲁棒辨识精度的同时,却需要消耗相当的计算耗时。针对上述问题,本文提出了一种自回归模型的快速鲁棒辨识算法,实验表明,新算法在辨识精度和辨识耗时方面较现有算法而言具有较强的鲁棒性。

1 问题描述

1.1 自回归模型及其参数估计

设平稳时间数据为 $x_t(t=1,2,\cdots,n)$,其 p 阶自回归模型的数学描述如式(1)所示。

$$x_t=\varphi_1x_{t-1}+\varphi_2x_{t-2}+\cdots+\varphi_px_{t-p}+u_n \tag{1}$$

其中:$\varphi_1,\varphi_2,\cdots,\varphi_p$ 为自回归参数;u_n 是均值为零、方差为 σ^2 的正态分布白噪声,即 $u_n\sim\mathrm{NID}(0,\sigma^2)$。

对于具有长度为 n 的平稳时间数据 x_t 而言,其自相关函数 $R_x(m)(m=0,1,2,\cdots,n-1)$ 可用式(2)进行估算。

$$\hat{R}_x(m)=\frac{1}{n-m}\sum_{i=0}^{n-m}x_ix_{i+m} \tag{2}$$

在 p 阶的自回归模型中,根据 Yule-Walker 方程,有

$$R'=R\varphi \tag{3}$$

其中:
$$R=\begin{bmatrix} R_x(\hat{0}) & R_x(\hat{1}) & \cdots & R_x(\hat{p-1}) \\ R_x(\hat{1}) & R_x(\hat{0}) & \cdots & R_x(\hat{p-2}) \\ \vdots & \vdots & & \vdots \\ R_x(\hat{p-1}) & R_x(\hat{p-2}) & \cdots & R_x(\hat{0}) \end{bmatrix}$$

$$\varphi=\begin{bmatrix}\varphi_1\\ \varphi_2\\ \vdots\\ \varphi_p\end{bmatrix},R'=\begin{bmatrix}R_x(\hat{0})\\ R_x(\hat{1})\\ \vdots\\ R_x(\hat{p})\end{bmatrix}$$

当式(3)的系数矩阵 R 为非奇异矩阵时,联合式(2)和(3),便可以求出 p 阶自回归模型的参数向量$[\varphi_1,\varphi_2,\cdots,\varphi_p]^{\mathrm{T}}$。

1.2 自回归模型的定阶问题

由式(3)可知,在自回归模型的辨识过程中,其首要问题是按某一准则来确定待辨识序列的阶次 p。下面以最小误差平方和的准则为例,分析自回归模型的定阶问题。

使用 p 阶的自回归模型来辨识平稳时间数据 x_t,不失一般性,忽略白噪声 u_n 的影响,则有

$$\hat{x}_t=\varphi_1x_{t-1}+\varphi_2x_{t-2}+\cdots+\varphi_px_{t-p}=\sum_{i=1}^{p}\varphi_ix_{t-i} \tag{4}$$

计算对应的整体辨识误差平方和 R_{ss} 有

$$R_{ss}=\sum_{t=p+1}^{n}[x_t-\hat{x}_t]^2=\sum_{t=p+1}^{n}[x_t-\sum_{i=1}^{p}\varphi_ix_{t-i}]^2 \tag{5}$$

以最小辨识误差平方和的准则对自回归模型进行定阶,其实质就是选取某一正整数 $p(1\leqslant p<n-1,p\in\mathrm{N})$ 作为系统阶次,并使得式(5)中的 R_{ss} 最小。

由于 R_{ss} 是阶次 p 的递减函数,即 R_{ss} 是随着阶次 p 的增加而减少[12~14]。据此,确定阶次的有效方法就是依次地对 $p=1,2,\cdots,n-1$ 的一系列模型进行参数估计并计算出它们各自对应的 R_{ss},若相邻阶次其 R_{ss} 的变化趋势不显著,则可把此时对应的 p 值作为模型的最佳阶次。上述的定阶方法从一阶开始依次递阶穷举,具有很大的盲目性,当最佳阶次 $p>>1$ 时,则此时需要消耗很大的计算开销。

2 自回归快速辨识算法的设计与实现

从 Yule-Walker 方程可知,最佳的阶次 p 总能使式(3)的线性方程组有唯一解,根据线性代数的理论得知,式(3)的线性方程组有唯一解的充要条件是,自相关函数矩阵 R 及其增广矩阵($R\quad R'$)的秩均等于 p,即 $\mathrm{rank}(R)=\mathrm{rank}(R\quad R')$ 成立。据此,可得到一种改进的自回归模型的快速辨识算法:通过估计出平稳时序数据自相关函数矩阵 R 的秩的下界值,并以该值作为自回归模型的起始阶数对系统进行依次的递阶辨识,当新近迭代所得的 R_{ss} 变化趋势不显著时,便结束算法的递阶迭代过程。易知,新算法由于避免了从一阶依次穷举辨识,故其具有较好的计算效能。

2.1 矩阵秩的下界值估计

设 $A=(a_{ij})_{n\times n}$,则矩阵 $A=(a_{ij})_{n\times n}$ 的秩的下界值可用式(6)进行估算[15~17]。

$$\mathrm{rank}(A)\geqslant\frac{|\mathrm{tr}(A)|^2}{\sqrt{\|A\|_F^4-\frac{1}{2}\|AA^{\mathrm{T}}-AA^{\mathrm{T}}\|_F^2}} \tag{6}$$

其中:$\mathrm{rank}(A)$ 为矩阵 A 的秩;$\mathrm{tr}(A)=\sum_{i=1}^{n}a_{ii}$ 为矩阵 A 的迹;$\|A\|_F=\{\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2\}^{\frac{1}{2}}$ 为矩阵 A 的 Frobenius 范数;A^{T} 为矩阵 A 的转置矩阵。

对于长度为 n 的平稳时间数据 x_t 而言,其自回归模型阶次的理论最大值为 $n-1$,故自相关函数矩阵 R 的维数的最大值也为 $n-1$,在模型阶次尚未确定的时候,若运用式(6)对模型的起始阶数进行估算,则需要计算出 $\hat{R}_x(0),\hat{R}_x(1),\cdots,\hat{R}_x(n-1)$ 等全部自相关函数。易知,当模型的最佳阶次 $p<<n$ 时,由于 $\mathrm{rank}(R=(r_{ij})_{p\times p})=\mathrm{rank}(R=(r_{ij})_{(p+1)\times(p+1)})=\cdots=\mathrm{rank}(R=(r_{ij})_{(n-1)\times(n-1)})$ 成立,故计算自相关函数 $\hat{R}_x(p+1),\hat{R}_x(p+2),\cdots,\hat{R}_x(n-1)$ 以及由此引起自相关函数矩阵 R 的元素增加所导致的计算开销均应避免,据此,有必要对式(6)进行递推的计算改进。

注意到式(6)的估算式为一非线性表达式,故这里将式(7)分为 U、V、W 三部分对自相关函数矩阵 R 的秩的下界值进行递推估算。

$$\mathrm{rank}(A^{(1)})\geqslant\frac{|U^{(1)}|^2}{\sqrt{(V^{(1)})^2-\frac{1}{2}W^{(1)}}} \tag{7}$$

具体的推导过程如下:

设矩阵 $A^{(0)}=(a_{ij})_{n\times n}$,$A^{(1)}=\begin{bmatrix}A^{(0)} & B\\ C & D\end{bmatrix}$,且 $B=(b_{ij})_{n\times 1}$,$C=(c_{ij})_{1\times n}$,$D=(d_{ij})_{1\times 1}$ 成立,若

$$U^{(0)}=\mathrm{tr}(A^{(0)}) \tag{8}$$

$$V^{(0)}=\|A^{(0)}\|_F^2 \tag{9}$$

$$W^{(0)}=\|A^{(0)}(A^{(0)})^{\mathrm{T}}-(A^{(0)})^{\mathrm{T}}A^{(0)}\|_F^2 \tag{10}$$

则有
$$U^{(1)}=\mathrm{tr}(A^{(1)})=\mathrm{tr}(A^{(0)})+\mathrm{tr}(D) \tag{11}$$

即
$$U^{(1)}=\begin{bmatrix}U^{(0)} & 0\\ 0 & 0\end{bmatrix}+\begin{bmatrix}0 & 0\\ 0 & d\end{bmatrix} \tag{12}$$

而
$$V^{(1)}=\|A^{(1)}\|_F^2=[(\sum_{i=1}^{n+1}\sum_{j=1}^{n+1}|a_{ij}|^2)^{\frac{1}{2}}]^2=\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2+\sum_{k=1}^{n}(|b_{k,n+1}|^2+|c_{n+1,k}|^2)+|d|^2 \tag{13}$$

即　　$$V^{(1)} = V^{(0)} + \sum_{k=1}^{n}(|b_{k,n+1}|^2 + |c_{n+1,k}|^2) + |d|^2 \tag{14}$$

又

$$W^{(1)} = \| A^{(1)}(A^{(1)})^{\mathrm{T}} - (A^{(1)})^{\mathrm{T}}A^{(1)} \|_F^2 = \left\| \begin{bmatrix} A^{(0)} & B \\ C & D \end{bmatrix}\begin{bmatrix} (A^{(0)})^{\mathrm{T}} & C^{\mathrm{T}} \\ B^{\mathrm{T}} & D \end{bmatrix} - \begin{bmatrix} (A^{(0)})^{\mathrm{T}} & C^{\mathrm{T}} \\ B^{\mathrm{T}} & D \end{bmatrix}\begin{bmatrix} A^{(0)} & B \\ C & D \end{bmatrix} \right\|_F^2 = \left\| \begin{bmatrix} A^{(0)}(A^{(0)})^{\mathrm{T}} + BB^{\mathrm{T}} - (A^{(0)})^{\mathrm{T}}A^{(0)} - C^{\mathrm{T}}C & A^{(0)}C^{\mathrm{T}} + BD - (A^{(0)})^{\mathrm{T}}B - C^{\mathrm{T}}D \\ C(A^{(0)})^{\mathrm{T}} + DB^{\mathrm{T}} - B^{\mathrm{T}}A^{(0)} - DC & CC^{\mathrm{T}} + DD - B^{\mathrm{T}}B - DD \end{bmatrix} \right\|_F^2 \tag{15}$$

令 $S = BB^{\mathrm{T}} - C^{\mathrm{T}}C, Z = CC^{\mathrm{T}} - B^{\mathrm{T}}B, X = A^{(0)}C^{\mathrm{T}} + BD - (A^{(0)})^{\mathrm{T}}B - C^{\mathrm{T}}D, Y = C(A^{(0)})^{\mathrm{T}} + DB^{\mathrm{T}} - B^{\mathrm{T}}A^{(0)} - DC$。则

$$W^{(1)} = W^{(0)} + \sum_{i=1}^{n}\sum_{j=1}^{n}|s_{ij}|^2 + \sum_{k=1}^{n}(|x_{k,n+1}|^2 + |x_{n+1,k}|^2) + |d^2| \tag{16}$$

综合式(7)~(16)便可得到自相关函数矩阵 R 的秩的下界的递推估算值。

2.2　自回归模型的参数校验

为了判定当模型的阶次发生改变时,其对应的 R_{ss} 的变化趋势是否显著,可引入如下的校验统计量:

$$t = \frac{(R_{ss2} - R_{ss1})(n-p)}{R_{ss1}} \tag{17}$$

其中:R_{ss1} 和 R_{ss2} 分别为 p 和 $p+1$ 阶所对应的整体拟合误差平方和;n 为平稳时序数据的长度。

大量的数理统计表明,当平稳时序数据的长度 n 足够大时,式(17)中的统计量 t 渐进地服从自由度为$(1, n-p)$的 $F(1, n-p)$分布[18,19],给定显著性水平 α(如 $\alpha = 0.05$),查 F 分布表可得到对应的临界值 F_α。于是,便得到如式(18)所示的自回归模型阶次的递推结束条件。

$$\begin{cases} \text{if } t \geqslant F_\alpha \text{ then} & p \text{ 是不合适的阶次} \\ \text{if } t < F_\alpha \text{ then} & p \text{ 是合适的阶次} \end{cases} \tag{18}$$

2.3　自回归快速辨识算法的实现

对上述的分析进行归纳和总结,便可以设计实现对应的平稳时序数据的自回归快速辨识算法。

算法 1　平稳时序数据的自回归快速辨识算法

输入:长度为 n 的平稳时间数据 x_t,显著性水平 α。

输出:自回归模型的最佳阶次 p 及其对应的模型参数向量$[\varphi_1, \varphi_2, \cdots, \varphi_p]^{\mathrm{T}}$。

a)初始化:根据输入的显著性水平 α 从 F 校验分布表得到对应的临界值 F_α,令 $m=0, p_1=1$,并按照式(2)估算平稳时间数据的自相关函数 $R(m)$。

b)$m = m+1$,并参照步骤 a)的方法重新估算 $R(m)$。

c)构建如式(3)所示的自相关函数的系数矩阵 $\boldsymbol{R} = (r_{ij})_{m\times m}$;在此基础上,参照式(7)的方法,利用式(12)(14)和(16)估算出系数矩阵 $\boldsymbol{R}$ 的秩的下界值→p_2。

d)若 p_1 与 p_2 不相等,则 $p_1 = p_2$,并跳转步骤 b);否则,$p = p_1$,跳转步骤 e)。

e)求解式(3)所示的线性方程组$(r_{ij})_{p\times p}(\varphi_k)_{p\times 1} = (r_l)_{p\times l}$,得到参数向量解$[\varphi_1, \varphi_2, \cdots, \varphi_p]^{\mathrm{T}}$,计算该参数向量解对应的整体拟合误差平方和→$R_{ss1}$。

f)$p = p+1$,并参照步骤 e)的方法求解参数向量解$[\varphi_1, \varphi_2, \cdots, \varphi_p]^{\mathrm{T}}$ 及其对应的整体拟合误差平方和→R_{ss2}。

g)把步骤 e)f)中的 R_{ss1} 和 R_{ss2} 代入式(17)计算校验统计量 t,按照式(18)对整体拟合误差平方和的变化显著性进行判别,若 $t \geqslant F_\alpha$ 则 $R_{ss1} = R_{ss2}$,并跳转步骤 f);否则,跳转步骤 h)。

h)输出最佳阶次 p 及其对应的模型参数向量$[\varphi_1, \varphi_2, \cdots, \varphi_p]^{\mathrm{T}}$,算法结束。

3　实验及结果分析

为了验证本文算法的有效性及先进性,这里将选取不同阶数和不同样本长度的平稳时序数据模型来进行自回归辨识算法的对比实验。实验在 PC 机上进行,其硬件配置为,Intel 酷睿 i5 4570四核 CPU、Kingmax DDR3 16GB RAM、Western Digital 500G Hard Disk;操作系统与开发环境为,Microsoft Windows 10、Microsoft Visual Studio 2010 集成开发环境中的 C++语言。在实验过程中,着重对比现有算法与本文算法的辨识精度和计算开销等技术指标,并对相关实验结果加以详细的分析和讨论。

3.1　实验过程与方法

在实验的过程中,使用了阶数分别为 5、10 及 15 共三个平稳时序数据模型,它们的具体参数如表 1 所示。

表 1　实验中所用的平稳时序数据模型的具体参数

模型	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8	φ_9	φ_{10}	φ_{11}	φ_{12}	φ_{13}	φ_{14}	φ_{15}
AR(5)	−0.81	0.21	0.19	−0.17	0.33	0	0	0	0	0	0	0	0	0	0
AR(10)	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.15	−0.12	0.18	−0.11	0	0	0	0	0
AR(15)	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.12	0.12	0.11	−0.12	−0.13	−0.16

基于 MATLAB 软件中,利用 randn( )随机数发生函数和 filter( )数字滤波器函数分别为上述模型各生成长度为 100,200,300,…,1000 的样本序列,这样,根据不同模型和不同样本长度便可以组建不同的实验组合。在这些不同的实验组合中,分别运行 Levinson-Durbin、Burg、文献[6]的算法及本文的快速算法,并从辨识精度和辨识耗时方面对各种算法进行评价。

3.2　实验的结果与分析

为了全面准确地评价各种算法的辨识精度,这里以式(19)的平均绝对百分误差(mean absolute percentage error,MAPE)和式(20)均方根误差(root mean square error,RMSE)作为评价指标,对上述算法在不同实验组合中的辨识精度进行评价。

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|x(t) - \hat{x}(t)|}{x(t)} \tag{19}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}[x(t) - \hat{x}(t)]^2} \tag{20}$$

其中:$x(t)$为样本序列的原值;$\hat{x}(t)$则为辨识序列的数值。

各种算法在不同阶数的实验模型组合中的辨识参数如表 2~4 所示(限于篇幅,这里仅列出样本长度 $n = 100, 500, 1000$ 三种情况),而它们对应的平均绝对百分误差 MAPE 和均方根误差 RMSE 数值如表 5 所示。

表 2　各种算法在基于 AR(5)模型实验组合的辨识参数

实验组合		φ_1	φ_2	φ_3	φ_4	φ_5
Levinson-Durbin 算法	n=100	−0.81	0.21	0.18	−0.11	0.32
	n=500	−0.81	0.21	0.20	−0.17	0.31
	n=1000	−0.81	0.21	0.19	−0.19	0.32
Burg 算法	n=100	−0.81	0.21	0.18	−0.19	0.34
	n=500	−0.81	0.21	0.20	−0.17	0.33
	n=1000	−0.81	0.21	0.19	−0.19	0.33
文献[6] 算法	n=100	−0.81	0.21	0.19	−0.18	0.32
	n=500	−0.81	0.21	0.19	−0.18	0.30
	n=1000	−0.81	0.21	0.19	−0.18	0.31
本文算法	n=100	−0.81	0.21	0.19	−0.23	0.35
	n=500	−0.81	0.21	0.20	−0.16	0.29
	n=1000	−0.81	0.21	0.20	−0.15	0.32

通过对比表 1~4 中各实验组合的辨识参数,结合表 5 的 MAPE 和 RMSE 数值,对各种算法进行总结,便可得出如下的结论:

a)Levinson-Durbin 算法的辨识精度与样本长度 n 有着一定的关联性,且随着模型阶数的增加,算法的辨识精度出现了明显的下滑,特别是在小样本序列($n = 100$)的组合实验中其下滑的情况尤为突出。

b)文献[6]算法在同一阶数下的不同样本长度的组合实验中具有稳定的辨识精度,与 Levinson-Durbin 算法类似,随着模型阶数的增加,该算法的辨识精度在整体上也出现了一定的下降。

c)在小样本序列的实验组合中,Burg 算法具有良好的辨识精度,而本文算法则与 Levinson-Durbin 和 Burg 算法一样,其辨识精度均随着模型阶数的增加而降低,三种算法的下降速度以 Levinson-Durbin 算法最为明显、Burg 算法其次、本文算法最为缓慢;在样本长度 $n \geqslant 500$ 的组合实验中,本文算法具有良好的辨识精度,且辨识精度的稳定性远优于其他两种算法。

事实上,由于 Levinson-Durbin 算法和本文算法均需要计算序列样本的自相关函数,这样,在高阶模型和小样本序列的实验组合中,用式(2)所得到的自相关函数估算值便出现较大的偏差,从而也就导致了上述两种算法的辨识精度变差;此外,在高阶模型的参数估计时,Levinson-Durbin、Burg 及文献[6]算法因其递推机制固有的误差积累也使得它们的辨识精度随着模型阶数的增加而降低。

表 3 各种算法在基于 AR(10)模型实验组合的辨识参数

实验组合		φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8	φ_9	φ_{10}
Levinson-Durbin 算法	n=100	0.21	0.14	−0.27	−0.19	0.22	−0.12	0.17	−0.17	0.24	−0.10
	n=500	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.16	−0.12	0.24	−0.15
	n=1000	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.16	−0.13	0.26	−0.12
Burg 算法	n=100	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.15	−0.11	0.19	−0.13
	n=500	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.16	−0.14	0.26	−0.12
	n=1000	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.16	−0.13	0.29	−0.10
文献[6]算法	n=100	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.17	−0.14	0.21	−0.16
	n=500	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.17	−0.14	0.21	−0.13
	n=1000	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.17	−0.14	0.21	−0.17
本文算法	n=100	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.14	−0.14	0.27	−0.15
	n=500	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.15	−0.12	0.20	−0.13
	n=1000	0.21	0.14	−0.27	−0.19	0.24	−0.14	0.15	−0.12	0.17	−0.10

表 4 各种算法在基于 AR(15)模型实验组合的辨识参数

实验组合		φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8	φ_9	φ_{10}	φ_{11}	φ_{12}	φ_{13}	φ_{14}	φ_{15}
Levinson-Durbin 算法	n=100	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.17	−0.19	0.10	0.13	−0.09	−0.08	−0.25
	n=500	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.14	0.13	0.12	−0.14	−0.19	−0.10
	n=1000	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.15	0.12	0.13	−0.13	−0.18	−0.16
Burg 算法	n=100	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.14	0.15	0.10	−0.15	−0.11	−0.22
	n=500	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.15	0.14	0.09	−0.14	−0.17	−0.13
	n=1000	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.15	0.12	0.13	−0.16	−0.11	−0.15
文献[6] 算法	n=100	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.14	0.13	0.12	−0.15	−0.18	−0.11
	n=500	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.14	0.13	0.12	−0.14	−0.19	−0.10
	n=1000	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.14	0.13	0.12	−0.14	0.17	−0.12
本文算法	n=100	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.11	−0.15	0.14	0.10	−0.16	−0.11	−0.21
	n=500	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.12	0.12	0.11	−0.14	−0.12	−0.20
	n=1000	0.13	−0.45	−0.18	0.21	−0.17	−0.16	0.22	0.16	0.13	−0.12	0.12	0.11	−0.14	−0.11	−0.19

表 5 各次实验组合的 MAPE 和 RMSE 数值

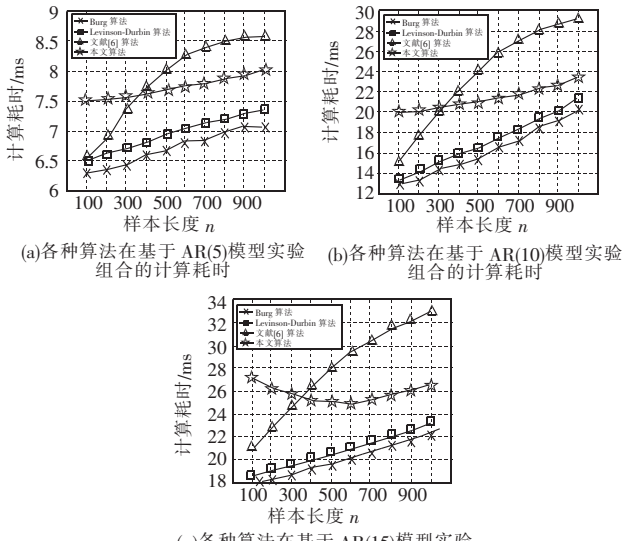
实验组合		AR(5)		AR(10)		AR(15)	
		MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Levinson-Durbin 算法	n=100	6.86%	53.96	14.56%	153.27	36.21%	292.66
	n=500	6.04%	62.72	9.77%	73.98	13.29%	155.45
	n=1000	6.12%	40.28	9.45%	86.41	13.11%	147.52
Burg 算法	n=100	6.65%	51.18	10.06%	106.27	16.45%	189.08
	n=500	5.89%	36.07	8.36%	55.17	12.98%	139.77
	n=1000	6.02%	59.44	9.05%	82.09	12.50%	130.39
文献[6] 算法	n=100	6.47%	48.53	8.96%	61.53	12.27%	122.58
	n=500	6.42%	73.84	8.48%	59.35	12.08%	111.81
	n=1000	6.45%	49.39	8.31%	69.86	12.16%	117.33
本文算法	n=100	7.78%	97.04	11.33%	123.22	17.43%	219.56
	n=500	6.29%	67.81	6.95%	41.32	6.88%	52.55
	n=1000	6.22%	44.25	6.91%	54.44	6.74%	48.76

各种算法在不同阶数的实验组合中的计算耗时如图 1 所示,从图 1 也可以得出如下的结论:

a)随着模型阶数的增加,各种算法的计算耗时也有了一定数量级的增加。从整体上进行比较,各种算法的计算耗时的排序是,Burg 算法最小、Levinson-Durbin 算法次之、本文算法排第三,而文献[6]算法则是最多的。

b)Levinson-Durbin 算法、Burg 算法与样本长度 n 存在着一定的线性关系,且该线性关系在不同阶数的模型中具有一定的稳定性;而文献[6]算法的计算耗时则与样本长度 n 有着抛物线关系;相对地,本文算法的计算耗时与样本长度 n 有着较平滑的对应关系。

通过对各种算法的运行原理分析便可得知,由于 Levinson-Durbin 算法的计算耗时主要花费在模型参数的递推计算过程中,而该计算过程与模型阶数和样本长度有着线性相依关系,这也与实验结论相吻合。文献[6]算法通过样本数据的矩阵分块求逆的方法进行递阶的参数辨识,故其计算耗时与样本长度 n 有着紧密的关系。此外,Burg 算法与文献[6]算法由于避开了自相关函数的计算,故它们的辨识精度不受样本长度 n 的影响。本文算法对各实验组合的自相关函数矩阵的秩的下界估计值如表 4 所示,由于设计了对应的递推算法,故本文算法的计算耗时主要是花费在线性方程组的求解过程,特别是在图 1(c)中,当样本长度 $n=100$ 时,由于此时的模型阶数较高,而少量的样本所估算的自相关函数便出现了较大偏差,从而导致对应秩的下界估计值偏离真实的模型阶数较多,于是,求解线性方程组的次数较其他实验组合多了,故其花费的计算耗时较样本长度为 500 和 1000 的时候还要多。



(a)各种算法在基于 AR(5)模型实验组合的计算耗时 (b)各种算法在基于 AR(10)模型实验组合的计算耗时

(c)各种算法在基于 AR(15)模型实验组合的计算耗时

图 1 各算法在不同模型实验组合的计算耗时

表 6　本文算法估算的自相关函数矩阵秩的下界值

n 值	AR(5)	AR(10)	AR(15)
$n=100$	4	8	10
$n=500$	4	8	13
$n=1000$	4	8	13

综上所述,在满足一定样本长度的要求下,本文算法的辨识精度和计算性能的鲁棒性能均优于现有算法,据此,本文算法是有效和可行的。

4　结束语

自回归模型是平稳时序数据中最常用和最广泛的辨识模型,基于现有算法的基础上,提出了一种改进的快速辨识算法,该算法较现有算法而言具有更稳定的辨识精度和更良好的计算效能。下一步的主要工作有,研究基于小样本场合的自相关函数的高精度估算方法,同时,也需要研究更为精确的矩阵的秩的下界值估计方法,以便进一步提升算法的适用范围和计算性能。

参考文献:

[1] 姜婷婷,肖卫东,张翀,等. 基于桑基图的时间序列文本可视化方法[J]. 计算机应用研究,2016,33(9):2683-2687.

[2] 王宏禹,邱天爽. 确定性信号分解与平稳随机信号分解的统一研究[J]. 通信学报,2016,37(10):1891-1898.

[3] 邹柏贤,刘强. 基于 ARMA 模型的网络流量预测[J]. 计算机研究与发展,2002,39(12):1645-1652.

[4] 程浩,刘国庆,成孝刚. 一种分段平稳随机过程自相关函数逼近模型[J]. 计算机应用,2012,32(2):589-591.

[5] 黄雄波. 非平稳时序数据的分段辨识及其递推算法[J]. 计算机系统应用,2017,26(5):180-185.

[6] Deng Feng, Bao Changchun. Speech enhancement based on AR model parameters estimation[J]. Speech Communication, 2016, 79(5): 30-46.

[7] 丁锋. 系统辨识算法的复杂性、收敛性及计算效率研究[J]. 控制与决策,2016,31(10):1729-1741.

[8] Boshnakov G N, Lambert-Lacroix S. A periodic Levinson-Durbin algorithm for entropy maximization[J]. Computational Statistics and Data Analysis, 2012, 56(1): 15-24.

[9] 胡明慧,王永山,邵惠鹤. 基于改进的莱文森算法对电机转速特性的研究[J]. 中国电机工程学报,2007,27(30):77-80.

[10] Matsuura M. On a recursive method including both CG and Burg's algorithms[J]. Applied Mathematics and Computation, 2012, 219(10): 773-780.

[11] 周毅,丁锋. 依等价 AR 模型阶次递增的自回归滑动平均模型辨识[J]. 华东理工大学学报:自然科学版,2008,34(3):425-431.

[12] 张仪萍,王士金,张土乔. 沉降预测的多层递阶时间序列模型研究[J]. 浙江大学学报:工学版,2005,39(7):983-986.

[13] Liu Yanjun, Ding Feng, Shi Yang. An efficient hierarchical identification method for general dual-rate sampled-data systems[J]. Automatica, 2014, 50(3): 962-970.

[14] 陈茹雯,湛时时. 基于非线性自回归时序模型的振动系统辨识[J]. 计算机应用研究,2016,33(10):3021-3025.

[15] 胡兴凯,伍俊良. 矩阵秩的下界和特征值估计[J]. 山东大学学报:理学版,2009,44(8):46-50.

[16] 黄廷祝. 矩阵秩的下界估计与 Schur 不等式的改进[J]. 电子科技大学学报,1993,22(5):537-541.

[17] Koltchinskii V, Lounici K, Tsybakov A B. Estimation of low-rank covariance function[J]. Stochastic Processes and Their Applications, 2016, 126(12): 3952-3967.

[18] 马铁丰,王松桂. 线性混合模型方差分量的检验[J]. 高校应用数学学报:A 辑,2007,22(4):433-440.

[19] 刘晓鹏, 刘坤会. F 分布密度函数之性质[J]. 应用概率统计, 2005,21(3):304-314.

(上接第 2642 页)先进的相似性度量算法。在 l_1 趋势滤波的基础上如何更好地对时间序列进行分段,在去噪的同时也能保留时间序列的特征点或关键点是以后要深入研究的方向。

参考文献:

[1] 杨一鸣,潘嵘,潘嘉林,等. 时间序列分类问题的算法比较[J]. 计算机学报,2007,30(8):1259-1266.

[2] Yu Fusheng, Dong Keqiang, Chen Fei, *et al*. Clustering time series with granular dynamic time warping method[C]//Proc of IEEE International Conference on Granular Computing. Washington DC: IEEE Computer Society, 2007: 393.

[3] Wan Yuqing, Gong Xueyuan, Si Y W. Effect of segmentation on financial time series pattern matching[J]. Applied Soft Computing, 2016, 38(1): 346-359.

[4] Rasheed F, Alhajj R. A framework for periodic outlier pattern detection in time-series sequences[J]. IEEE Trans on Cybernetics, 2014, 44(5): 569-582.

[5] Frawley C. Fast subsequence matching in time-series database[J]. Proc Sigmod, 2008, 23(2): 419-429.

[6] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series[C]//Proc of Working Notes of the Knowledge Discovery in Databases Workshop. Palo Alto, CA: AAAI Press, 1994: 359-370.

[7] Chen Lei, Ng R. On the marriage of Lp-norms and edit distance[C]// Proc of the 13th International Conference on Very Large Data Bases. Toronto: VLDB Endowment, 2004: 792-803.

[8] Vlachos M, Kollios G, Gunopulos D. Discovering similar multidimensional trajectories[C]//Proc of the 18th International Conference on Data Engineering. Piscataway, NJ: IEEE Press, 2002: 673-684.

[9] Sun Youqiang, Li Jiayong, Liu Jixue, *et al*. An improvement of symbolic aggregate approximation distance measure for time series[J]. Neurocomputing, 2014, 138(8): 189-198.

[10] Wang Xiaoyue, Mueen A, Ding Hui, *et al*. Experimental comparison of representation methods and distance measures for time series data[J]. Data Mining and Knowledge Discovery, 2013, 26(2): 275-309.

[11] 肖瑞,刘国华. 基于趋势的时间序列相似性度量和聚类研究[J]. 计算机应用研究,2014,31(9):2600-2605.

[12] 李海林,郭崇慧. 基于形态特征的时间序列符号聚合近似方法[J]. 模式识别与人工智能,2011,24(5):665-672.

[13] 李桂玲. 时间序列的分割及不一致发现研究[D]. 武汉:华中科技大学,2012.

[14] 丁永伟,杨小虎,陈根才,等. 基于弧度距离的时间序列相似度量[J]. 电子与信息学报,2011,33(1):122-128.

[15] 刘博宁,张建业,张鹏,等. 基于曲率距离的时间序列相似性搜索方法[J]. 电子与信息学报,2012,34(9):2200-2207.

[16] 秦磊,谢邦昌. L1 和 L2 规则化趋势滤波的稳健集成方法[J]. 统计研究,2013,30(11):99-102.

[17] Efron B, Hastie T, Johnstone I, *et al*. Least angle regression[J]. Annals of Statistics, 2004, 32(2): 407-451.

[18] Keogh E, Xi X, Wei L, *et al*. The UCR time series classification/clustering homepage [EB//OL]. 2006. http://www.cs.ucr.edu/~eamonn/time_series_data/.

[19] Derryberry D W R. Appendix A: using datamarket[M]//Basic Data Analysis for Time Series with R. Hoboken: Wiley, 2014.