文章编号:1000-582X(2008)05-0558-05

动态数据系统的时间序列建模策略

尹光志[1,2],钟　焘[1,2],李德泉[3]
(1. 重庆大学 资源及环境科学学院,重庆 400030;2. 重庆大学 西南资源开发及环境灾害控制工程教育部重点实验室,重庆 400030;3. 重庆市公用事业建设工程公司,重庆 400020)

摘　要:传统的基于相关性分析方法进行建模的局限性和“危险性”主要表现在:估计的样本自相关是非常坏的估计,经常会有大的方差,彼此之间是高度相关的,可能给出原来序列结构一个完全失真的图像,不能较准确和全面地反映系统特性。提出了基于动态数据系统的时间序列建模方法,将时间序列看作是随机系统对不相关的或相互独立的“白噪声”输入响应的一种实现方式。对平稳时间序列,以自回归滑动平均模型为基本模型,并以额值为 1 递增拟合,用 F 检验判断拟合的改善程度,最后用残差分析判断模型的适用性。对非平稳序列,需先分离出确定性趋势,对剩余平稳随机部分建模分析。用该方法对隧道位移监测数据建模分析,预测与实测吻合较好,表明该方法具有适用性好、精度高且便于编制程序实现等优点。

关键词:时间序列;动态数据系统;建模策略;最小二乘估计;线性回归
中图分类号:U458.1　　文献标志码:A

A time series modeling strategy for dynamic data systems

YIN Guang-zhi[1,2], *ZHONG Tao*[1,2], *LI De-quan*[3]
(1. College of Resource and Environmental Sciences, Chongqing University, Chongqing 400030, P. R. China; 2. Key Laboratory for the Exploitation of Southwest Resources and the Environmental Disaster Control Engineering, Ministry of Education, Chongqing University, Chongqing 400030, P. R. China; 3. Chongqing Public Utility Construction Engineering Co., Chongqing 400020, P. R. China)

Abstract: The limitations and “risks” existing in traditional modelling methods based on correlation analysis mainly lies in the following three areas: estimated sample autocorrelation is a poor estimator with regular large variance and high relationships with each other; they may produce a completely distorted image of the original series structure; and they are unable to reflect system characteristics accurately and roundly. A modelling method based on dynamic data systems in time series was presented. The time series was regarded as a realistic way to input response on a stochastic system to uncorrelated white noise. For stationary time series, an autoregressive moving average model was the basic model. The model took 1 as the increasing amplitude and fits model, used the F test to judge the degree to which the fit improved, and used residue analysis to weigh model applicability. For nonstationary time series, it needed to isolate deterministic trends first, and then model and analyze the surplus remaining stochastic portion. This method was used to model and analyze the displacement monitoring data of tunnels. The prediction accorded well with actual measurements. The results show that the method is quite applicable, highly precise, and can be easily implemented through programs.

Key words: time series; dynamic data system; modeling strategy; least squares approximations; linear regression

收稿日期:2007-12-12
基金项目:国家自然科学基金资助项目(50374084)
作者简介:尹光志(1962-),男,重庆大学教授,博士生导师,主要从事岩石力学与工程方向的研究,(E-mail)gzyin@cqu.edu.cn。

时间序列分析是数理统计学的一个重要分支，经过数十年发展，已广泛应用于工农业生产中。这是一种处理随时间变化而又相互关联数据的数学方法，或者说，是一种处理动态数据的参数化时域分析方法。其主要手段是选择恰当的数学模型来近似描述动态数据，通过研究分析，本质地了解数据内在结构和复杂特性，以达到预测和控制的目的[1]。

数据的统计依赖关系通常是用逐次观测值之间的相关函数或自相关函数来表示。现行时间序列分析方法大都是建立在经验的或估计的自相关或其傅里叶变换自谱基础上的，讨论估计的样本自相关在时间序列的建模与估计中的应用，通过对均值预处理，若是非平稳序列则利用差分或者季节性进行处理，利用样本自相关或样本偏相关的截尾性质进行模式识别，利用矩法、最小二乘法、极大似然估计法等进行参数估计，最后对所建立模型进行预测或者控制。

假若样本自相关函数是理论自相关函数的一个良好估计时，则运用样本自相关函数是合适的。但是国外学者 Kendall 指出估计的样本自相关是非常坏的估计，它经常会有大的方差，彼此之间是高度相关的，结果只能给出真实自相关的一种失真形式[2]。由于经验的自相关是理论自相关的一种不良估计，使得基于这些估计的时间序列分析方法变得困难和麻烦，需要强烈依赖某些特设的试凑法。

1　动态数据系统法简介

在进行系统分析时，工程师和科学家使用的数学模型是根据所假设的一个物理结构推导出来的，通常为微分方程。对于复杂的系统，则采用一种试验的方法，如频率分析法。而统计师和经济学家则利用从试验或经验得来的自相关图线和谱图以差分方程的形式来近似表达各种模型。如果把时间序列和系统分析有机结合起来，不仅可以避免大量的试凑工作，而且应用范围会更广。

动态数据系统是一种建立数学模型的方法，它采用一种时间序列形式的动态数据来建立一个具有物理意义的随机差分或微分方程。在动态数据系统方法中，时间序列被看成是随机系统对不相关的或相互独立的“白噪声”输入响应的一种实现方式。连续时间或离散时间的动态系统数学模型把相互依赖或相关的时间序列输出化为相互独立的或不相关的输入（白噪声），把建模作为寻求随机动态系统表达式的过程来处理，该表达式是由时间序列数据推导得出并与之有依赖关系的差分方程或微分方程。这种研究方法可归结为：寻找一个能完成这种相互独立数据转化的模型，对独立的多次观测使用一些标准的统计方法进行估计、预测和控制[3]。

1.1　动态数据系统法的惯用模式

动态数据系统重点研究对象是 ARMA(n,$n-1$)模型，把其它模型作为 ARMA(n,$n-1$)模型的特殊情况来处理[4]。

它通过令某些参数为零，或使某些自回归或滑动平均因子倒置，这实际上包括了许多 $m<n-1$ 与 $m>n-1$ 的其它 ARMA(n,m)模型。

1.2　动态数据系统法的适用性

利用希尔伯特空间线性算子理论可以证明，对于离散的、连续的、标量的以及向量的情况，用一个 ARMA(n,$n-1$)自回归滑动平均模型可以把任一平稳随机系统逼近到所要求的精确程度[5]。

另外，对一个用自回归阶次为 n，滑动平均阶次为任意值的线性微分方程表示的连续自回归滑动平均过程，在均匀间隔上采样时，其综合采样过程是 ARMA(n,$n-1$)模型。

再者，格林函数 G_j 表示系统动态，求取格林函数 $G_j=g_1\lambda_1^j+g_2\lambda_2^j+\cdots+g_n\lambda_n^j$ 与求解具有适当初始条件的齐次差分方程是等价的。由于有 n 个线性相互独立的根 $\lambda_1,\lambda_2,\cdots,\lambda_n$，显然齐次方程是 n 阶的，因此，自回归阶数为 n，且自回归参数由 $(1-\phi_1B-\phi_2B^2-\cdots-\phi_nB^n)G_j=0(j\geqslant n)$ 决定。n 个常数 g_1，g_2，$\cdots$，g_n 必须由 n 个初始条件来确定，其中有一个条件 $G_0=g_1+g_2+\cdots+g_n=1$ 是成立的，这样只有 $n-1$ 个条件还有待确定，一般需要 $n-1$ 个 θ_i，所以滑动平均阶次一般为 $n-1$。

2　动态数据系统法的建模策略

建模时，用 ARMA(n,$n-1$)模型的递增序列去逐步逼近数据的依赖关系，当用残差平方和的减小来作判断依据时，这种建模只停留在拟合的改善程度在统计上毫无意义的这样一个 n 值上[5]。

当模型中不存在滑动平均部分时，最终的方程对模型参数 ϕ_i 和 θ_i 是线性的，可用线性最小二乘法求解模型参数。当模型中存在滑动平均部分时，最终的方程对模型参数 ϕ_i 和 θ_i 是非线性的，必须应用非线性最小二乘法递归求解模型参数。非线性最小二乘法用逐步逼近的方式来实现残差平方和的极小化，先从诸参数的某些初始值（可用逆函数求解）开始，递归计算诸残差并求得其平方和，然后，通过诸

如线性化或最速下降法找到平方和减小所沿的“方向”，用具有较小平方和的参数新值作为初始值，一直继续下去直至获得极小的平方和为止。

2.1 自回归阶次的增额

在拟合模型时，n 以额值为 1 递增进行拟合是可以的。但经验表明，较好的办法是 n 以额值为 2 递增，其理由是[5]：

1)阶数增加 2 相当于特征方程增加 2 个根，可以避免强使模型增加一个实根而造成失误。当根数增加 2 时，它既可以是共轭复根，也可以是 2 个实根(当实际为 1 个实根时，拟合结果中将有 1 个实根接近于零)，因此比较合理。

2)实际物理系统的自由度随着它的复杂程度增加。自由度每增加 1，阶数便增加 2。

拟合 ARMA($2n$,$2n-1$)模型，当 ARMA($2n$,$2n-1$)模型中有参数如 ϕ_{2n}，θ_{2n-1} 的置信区间包括零时，此模型便退化为 ARMA($2n-1$,$2n-2$)，以此类推。

2.2 适用性检验

1)F 检验

F 检验用于模型定阶是比较经典的方法，对 ARMA 模型进行拟合，通过残差平方和的相对变化判断模型是否有升阶的可能和必要[6]。

$$F=\frac{A_1-A_0}{s}\div\frac{A_0}{N-r}\sim F(s,N-r),\qquad(1)$$

式中：A_0 是高阶(不受限)模型的残差平方和；A_1 是低阶(受限)模型的残差平方和；s 是受限参数的个数；r 是高阶(不受限)模型的阶数；$F(s,N-r)$ 表示具有 s 和 $N-r$ 个自由度的 F 分布。

2)残差自相关检验

残差分析是检验过程中不可缺少的环节，可以通过残差自相关较小的特点检验残差的独立性[7]。

a. 判断 $\hat{\rho}_k(a_t)$ 是否在 $\pm 2/\sqrt{N}$ 的范围；

b. 拟合不足的一揽子检验[2]：

$$Q=N\sum_{k=1}^{K}\hat{\rho}_k^2(a_t)\sim\chi^2(K-p-q),\qquad(2)$$

其中：$\hat{\rho}_k(a_t)$ 是残差自相关；$\chi^2(K-p-q)$ 表示具有 $K-p-q$ 个自由度的 χ^2 分布。

2.3 建模步骤

1)拟合 ARMA($2n$,$2n-1$)模型。n 值每增加 1，便利用 F 判据检验残差平方和的改善情况，在从 ARMA($2n$,$2n-1$)模型转到 ARMA($2n+2$,$2n+1$)模型时，当 F 值在预定水平上变得无意义时，就停止向前拟合，并选择 ARMA($2n$,$2n-1$)模型。

2)检验 ϕ_{2n}，θ_{2n-1} 值。观察它们与其最大绝对值 1 相比是否小，它们的置信区间是否包括零。如果否，则 ARMA($2n$,$2n-1$)模型为一个合适的模型。

3)假如 ϕ_{2n}，θ_{2n-1} 的值小，并且它们的置信区间包括零，则拟合一个 ARMA($2n-1$,$2n-2$)模型，用 F 判据检验它与 ARMA($2n$,$2n-1$)模型。假如 F 值不大，则略去小的 MA 参数，拟合一个 $m<2n-2$ 的 ARMA($2n-1$,m)模型，并应用 F 检验，直至得到参数最少的合适模型为止。

4)假如 F 值大，则略去小的滑动平均参数，并且确定 $m<2n-1$ 的一个 ARMA($2n$,m)模型。

5)假如需要时，通过 F 检验拟合模型的理想形式(例如纯 AR 或纯 MA 模型等)，使模型的阶次按额值为 1 递增，直至 F 值小到无意义为止。

6)若选出合适模型，则还应验证残差自相关是否在允许的范围内，作为对 F 检验的补充。

由经验知，当观测次数少时，模型的阶次一般不会高，在实用中，可用 ARMA(n,$n-1$)模型来替代 ARMA($2n$,$2n-1$)模型进行时间序列建模处理。

3 非平稳时间序列建模策略

平稳序列是指时间序列的前两阶矩(均值和协方差)与时间起点无关，即序列拥有固定或恒定的均值，可以通过减去它来化为零均值模型。而非平稳序列是指数据序列的性质与时间起点有关，是时间起点的一个确定性函数[8]。下面主要讨论具有趋向性的一类常见时间序列，季节性序列不在讨论范围中。

基于动态数据系统的时间序列建模策略将模型分解成为 2 部分：确定性部分和随机性部分。对数据中非平稳性部分单独拟合确定性模型，根据数据图形所表现出的实际趋势用最小二乘法作线性或指数逼近，当然也可以是对数函数、双曲函数、多项式或者其它较为接近的函数逼近，对剩下的随机性部分(平稳序列)按照前面的策略进行时间序列建模。在获得确定性函数和平稳序列的 ARMA 模型后，再用这些模型的参数作为初始值，利用非线性最小二乘法迭代并确定各参数的最终值。

线性趋势：$Y_t=\beta_0+\beta_1 t+X_t$，(3)

指数趋势：$Y_t=B_0+B_1 e^{-t/\tau}+X_t$，(4)

式中：t 表示时间；X_t 是平稳随机部分；Y_t 是原数据序列；β_0、β_1、B_0 和 B_1 均是模型参数。

4 建模实例

以重庆市大足县南环二路南山隧道工程现场位

移监测数据为例进行建模分析。数据采取的时间为2007 年 1 月 23 日至 2007 年 3 月 28 日某个断面的收敛位移监测结果,每天 1 次,共 65 个数据,见图 1,Y 为水平位移。

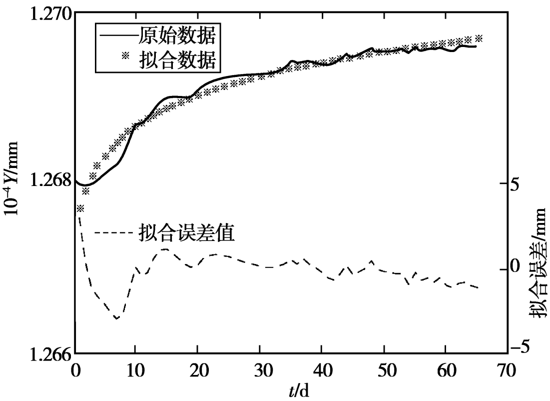

图 1　位移-时间拟合曲线图

1)非线性回归分析

由图 1 可以看出,数据序列是非平稳序列,需要先进行非线性回归处理分离出具有确定性趋势部分,再对剩余的随机部分进行时间序列建模处理。

这里采用隧道中常见的 3 种模型拟合形式:对数函数、指数函数和双曲函数。

其中对数函数拟合误差均值为 0(即是零均值平稳过程),方差为 0.774 7;指数函数拟合误差均值为 0.000 1,方差为 10.616 5;双曲函数拟合误差均值为−0.000 5,方差为 10.608 1。

对数函数拟合误差的均值及方差均为最小,因此选用对数函数作为确定性部分,随机部分即为拟合误差,对数函数拟合效果见图 1。

对数函数拟合结果:

$u=a+b\ln(1+t)=12\ 672.860\ 5+5.693\ 2\ln(1+t)$。

2)时间序列分析

由于任意的 ARMA 模型总可以在一定精度下以适当高阶 AR 模型来逼近[9],那么可以用 AR 模型以额值为 1 递增拟合,直至残差为白噪声。如果模型阶数不高,则认为适用模型为 AR 模型,如果阶数较高,则改用 ARMA 模型拟合。考虑到现场实际需要,可采用 AR 模型进行时间序列建模分析,其计算简单,易于满足现场实时性要求[10]。

a. 参数估计和 F 检验

使用上侧分位数,显著性水平取 0.05。

经计算和统计,求得模型参数为:

AR(1)模型,$\hat{\phi}=[0.758\ 5]$,$A=13.37$。

AR(2)模型,$\hat{\phi}=[1.212\ 7,-0.374\ 3]$,$A=7.76$;

$$F=\frac{13.37-7.76}{7.76}\times\frac{65-2}{1}\approx 45.55>$$
$$F_{0.05}(1,65-2)=4。$$

F 检验显著,表明 AR(2)模型相对 AR(1)模型,对残差的改善明显,所以 AR(2)模型比 AR(1)模型更适合。

AR(3)模型,$\hat{\phi}=[1.054\ 1,-0.100\ 3,-0.115\ 3]$,$A=7.27$;

$$F=\frac{7.76-7.27}{7.27}\times\frac{65-3}{1}\approx 4=$$
$$F_{0.05}(1,65-3)=4。$$

F 检验不显著,表明 AR(3)模型相对 AR(2)模型,对残差的改善不明显,且 AR(3)模型的后 2 个自回归参数较小,接近于零置信区间,综合考虑选用参数较少的 AR(2)模型。

AR(2)模型 $\phi(B)=0$ 的 2 个特征根为 $\lambda_{1,2}=0.606\ 4\pm0.081\ 6i$;它的模 $0.611\ 8<1$,表明系统是稳定的,参数估计是适合的,时间序列预测效果如图 2。

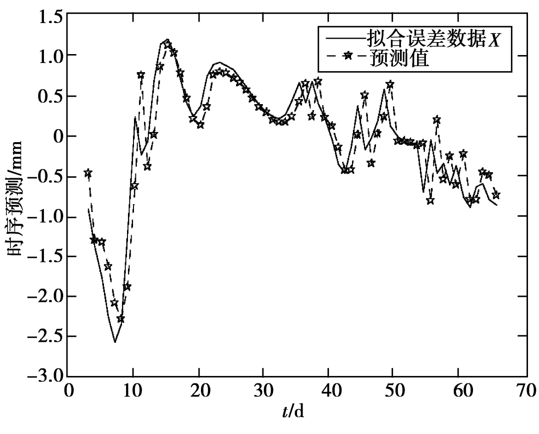

图 2　时间序列预测效果图

b. AR(2)模型残差自相关检验

①残差自相关的 Bartlett 范围($\pm 2/\sqrt{N}$),见图 3。

②拟合不足的一揽子检验:

$$Q=65\sum_{k=1}^{10}\hat{\rho}_k^2(a_t)=10.1<$$
$$\chi^2_{0.05}(10-2-0)=15.5,$$

至此,认为 AR(2)模型是合适的。

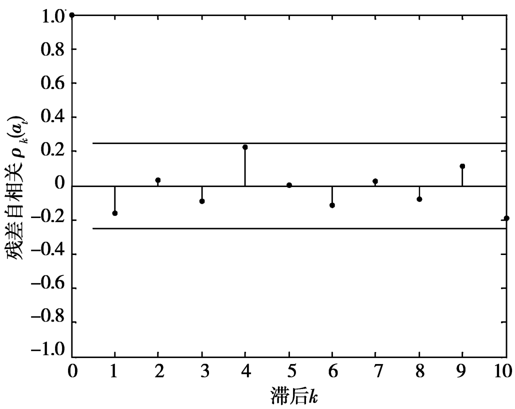

图 3 残差自相关图

5 结 语

1)对于平稳时间序列,可以额值为 1 递增拟合 ARMA($2n$,$2n-1$)模型,通过残差平方和减小是否显著来确定模型。

2)对于非平稳时间序列,需要先分离出确定性部分,然后把剩余部分看成是随机的,用 ARMA($2n$,$2n-1$)模型拟合,最后将两部分用非线形最小二乘法迭代求出最终值,得到较好的模型参数。

3)实际工程应用中,可先用 AR 模型以额值为 1 递增拟合,直至残差为白噪声。如果模型阶数不高,则认为适用模型为 AR 模型,如果阶数较高,则改用 ARMA 模型拟合。

目前,基于动态数据系统的时间序列分析方法的应用范围很广,涉及工业自动化、水文、地质、气象、机械、经济、化工等,但由于计算中大多涉及非线形处理方法,在初值给定和终值求解达到相当高的精度上还存在一定的难度,需要进一步改进和完善;此外,这里着重讨论一维时间序列,对于多维时间序列的推广及其控制和优化还需要进一步探讨。

参考文献:

[1] 安鸿志,陈兆国,杜金观,等.时间序列的分析与应用[M].北京:科学技术出版社,1983.

[2] 博克斯,詹金斯.时间序列分析:预测与控制[M].顾岚,范金城,译.北京:中国统计出版社,1997.

[3] 顾瑞龙,陶志范.动态数据系统及其应用[J].制造技术与机床,1980(12):10-13.
GU RUI-LONG, TAO ZHI-FAN. Dynamic data system and its application [J]. Manufacturing Technology & Machine Tool, 1980(12):10-13.

[4] 杨位钦,张志方.动态数据系统(DDS)方法的应用和建模[J].机器人,1981(3):16-24.
YANG WEI-QIN, ZHANG ZHI-FANG. Dynamic data system application and modeling[J]. Robot, 1981(3):16-24.

[5] 潘迪特,吴宪民.时间序列及系统分析与应用[M].北京:机械工业出版社,1988.

[6] 吕锐.时间序列分析中的模型定阶策略[J].国防科技大学学报,1988,10(4):97-106.
LV RUI. Time series rules of judging models for time series analysis[J]. Journal of National University of Defense Technology, 1988,10(4):97-106.

[7] 贺诗波,刘祥官,郜传厚.高炉硅含量预测控制的时间序列混合建模[J].浙江大学学报,2007,41(10):1739-1742.
HE SHI-BO, LIU XIANG-GUAN, GAO CHUAN-HOU. Hybrid time series predictive control model for silicon content in blast furnace hot metal[J]. Journal of Zhejiang University, 1999,39(4):57-59.

[8] 田铮.动态数据处理的理论与方法[M].西安:西北工业大学出版社,1995.

[9] 钟秋海.动态数据系统(DDS)建模方法的改进[J].北京工业学院学报,1986(4):36-43.
ZHONG QIU-HAI. An improved algorithm for dynamic data system modeling[J]. Journal of Beijing Institute of Technology, 1986(4):36-43.

[10] 徐峰,王志芳,王宝圣.AR 模型应用于振动信号趋势预测的研究[J].清华大学学报,1999,39(4):57-59.
XU FENG, WANG ZHI-FANG, WANG BAO-SHENG. Research on AR model applied to forecast trend of vibration signals[J]. Journal of Tsinghua University, 1999,39(4):57-59.

(编辑 张 苹)