吴今培著

SHIYONG
SHIXU
FENXI

实用时序分析

实用时序分析

吴今培 著

湖南科学技术出版社

实用时序分析

吴今培 著

责任编辑：胡海清

*

湖南科学技术出版社出版发行

（长沙市展览馆路8号）

湖南省新华书店经销 湖南省新华印刷二厂印刷

*

1989年12月第1版第1次印刷

开本：850×1168毫米 1/32 印张：9.125 插页：（精）4 字数：236,000

印数：（精）1——400 （平）1——1,600

（精装）ISBN 7—5357—0627—4 / O·70 定价：5.40元

（平装）ISBN 7—5357—0628—2 / O·71 定价：4.10元

地科89—38

序

时间序列分析属于概率统计学科的一个重要分支，近十余年来得到迅速的发展，尤其是实际应用遍及自然科学、社会科学及工程技术的许多领域，越来越显示出旺盛的生命力。

《实用时序分析》是把时间序列分析的理论与方法应用于随机数据处理。作者既综合了时间序列分析的一系列最新文献，又总结了自己在时间序列的实际应用中所取得的某些成果。本书对于那些应用时序分析来建立动态模型和处理实验数据的科技工作者、大学生、研究生，无疑是一本导引性的工程应用技术的著作。

时间序列分析方面的系统著作，我国已有好几本。吴今培同志的这本新著是这一领域的又一产品。值得指出的是本书将稳健统计的思想和投影寻踪的方法溶入时序分析之中，探讨了观测数据因受干扰而存在异常值时的稳健参数估计问题，以及高维动态数据的建模问题。

本书文字流畅，立意明确，在数学处理上力求通俗，以使更多的实际工作者能够成为本书的读者。我很高兴地把这本书介绍给科技界的同志们，希望能对我国时间序列的理论研究与应用研究起到一定的促进作用。

侯振挺

1989年5月于长沙

前　言

在自然科学、社会科学及工程技术的许多领域中，实际工作者和科技人员常常要对一系列观测数据进行分析研究。这些数据一般按时间顺序排列，由于受到多种偶然因素的影响，往往表现出某种随机性，且观测值之间存在着依赖关系。对这种按时间顺序排列的动态数据进行研究，构成了数理统计的一个重要分支——时间序列分析。

自从1970年G.E.P.Box和G.M.JENKINS发表专著“时间序列分析：预测和控制”，对平稳时间序列数据，提出自回归滑动平均模型，以及一套完整的建模、估计、检验、预测和控制方法以来，这一领域吸引了大批科技人员从事理论和方法上的进一步研究，其应用遍及气象、水文、天文、生物、电力、机械、雷达、化工、经济以及交通运输等方面。尤其，随着微计算机的普及，时间序列分析在理论和应用上的研究更是方兴未艾，成为一门充满生机的学科，越来越显示出旺盛的生命力。

然而，不可避免地，在实际的时间序列中也含有一些干扰异常值，并且，数据的观测误差也不服从通常所假定的正态分布，而是具有拖尾特征。因此，常常使得建立在常规估计、检验方法之上的已有时间序列分析方法陷入困境，难以良好地拟合动态数据。当然，给不出准确的预测和控制。

随着稳健统计的发展，人们首先自然地想到了将稳健统计的思想和方法，溶入时间序列分析中，改进现有的时间序列分析中的建模、估计、检验及预测，构造出一类时间序列稳健分析方法，使时间序列分析从稳健统计分析的角度朝着现实世界又迈出了一

大步。

本书共分八章，系统论述了时间序列分析和动态数据处理的一般理论和方法，侧重于时域分析，也涉及了谱分析的有关内容。从工程实际的要求出发，对参数估计的线性递推算法和各种定阶准则作了全面系统的介绍。对时间序列建模的各种原理和手段，以及在计算机上实现的具体步骤也进行了较为完整的讨论。本书对时间序列的稳健分析和多维数据的处理，既收集了较为成熟的内容，也介绍了较新的研究成果。

作者希望本书能对从事时间序列分析应用的实际工作者，在解决工程问题方面有所帮助。因此，在论述时力求深入浅出，既考虑在理论上具有一定程度的完整性，更着重于介绍方法和应用，从而省略了一些复杂的数学证明，读者如有兴趣，可按书末提供的索引查阅有关的文献资料。

本书参考和引用了国内外许多作者的有关论著，从中受到了莫大的教益。我们在此谨向各位学者表示衷心的感谢！

由于作者水平有限，时间又很仓促，我们深知本书还会有缺点和错误，敬请读者批评指正。

作者

1989年4月

目　　录

第一章　理论概要……………………………………………………(1)
§ 1　时间序列的定义…………………………………………(1)
§ 2　时间序列的分析方法……………………………………(3)
§ 3　时间序列的参数模型……………………………………(4)
§ 4　时间序列模型的特性……………………………………(14)
第二章　模式识别………………………………………………………(33)
§ 1　平稳性数据的模式识别…………………………………(33)
§ 2　季节性数据的模式识别…………………………………(40)
§ 3　趋势性数据的模式识别…………………………………(47)
§ 4　异常值数据的模式识别…………………………………(52)
第三章　参数估计………………………………………………………(56)
§ 1　AR 模型的参数估计……………………………………(58)
§ 2　ARMA模型的参数估计…………………………………(83)
第四章　阶次判定………………………………………………………(95)
§ 1　F检验的定阶准则…………………………………………(96)
§ 2　白度检验的定阶准则……………………………………(100)
§ 3　相关熵的定阶准则………………………………………(103)
§ 4　FPE定阶准则……………………………………………(107)
§ 5　AIC定阶准则……………………………………………(111)
§ 6　其它定阶准则……………………………………………(113)
第五章　稳健分析………………………………………………………(117)
§ 1　稳健性概念的描述………………………………………(118)

§ 2　时间序列的稳健分析……………………………………(128)
§ 3　时间序列的稳健估计……………………………………(132)
§ 4　时间序列的稳健化处理…………………………………(149)
第六章　系统建模……………………………………………(158)
§ 1　波克斯—詹金斯的建模方法…………………………(158)
§ 2　潘迪特—吴贤铭的建模方法…………………………(167)
§ 3　蔡-刁的建模方法 ……………………………………(176)
§ 4　时间序列的预测方法……………………………………(187)
第七章　谱估计…………………………………………………(196)
§ 1　经典谱估计…………………………………………………(196)
§ 2　现代谱估计概述……………………………………………(201)
§ 3　AR 参数谱估计 ……………………………………………(207)
§ 4　自适应谱估计………………………………………………(215)
§ 5　稳健谱估计…………………………………………………(222)
第八章　多维时序处理………………………………………(228)
§ 1　多维数据处理的投影寻踪法…………………………(228)
§ 2　多维时序建模的 PP 方法……………………………(237)
§ 3　PP 指标函数的递推计算方法 ………………………(250)
附录……………………………………………………………(269)
附表 1　正态分布表…………………………………………(269)
附表 2　F分布表……………………………………………(272)
附表 3　χ^2分布表……………………………………………(274)
附表 4　r 分布表 ……………………………………………(277)
参考文献………………………………………………………(278)

第一章　理论概要

在工程、经济、自然科学和社会科学等领域的实际工作者和研究人员都要和一系列的观察数据打交道。我们把按时间顺序产生和排列的观察数据序列称为时间序列。比如气象上的月降水量序列，水文上的河流流量序列，海洋上的海浪序列，天文上的太阳黑子序列，地球物理上的地震波序列，医学上的脑电波序列，机械系统的振动序列，雷达系统对活动目标的跟踪测量序列，自动控制系统的输入输出信号序列，自然界某种生物总数的消长序列，商业经济方面的产值和销售额序列等都是时间序列的具体例子。

一般来说，时间序列很难用一个完全确定的数学函数来表述，它们大都具有统计规律性，可以通过概率分布函数对它们的规律性作统计描述。因此，本书讨论的时间序列是指随机的时间序列，亦称随机序列。

§1　时间序列的定义

由一串随机变量…，x_1，x_2，x_3，…构成的序列叫做随机序列，用x_t($t=\cdots$，1，2，3，…）或 $\{x_t\}$ 表示。如果下标t是整数变量，它代表着等间隔的时刻增长量，如第t时、第t天、第t次等，我们就称这种随机序列为时间序列，而整数变量t即认为是指某时刻。

粗知概率论的读者便明白，随机变量x是某个随机现象（或随机试验结果）的总称，做一次试验就能获得x的一个取值 X，

称为x的一个样本值，它是一个普通的数，以一定的概率出现。对于时间序列而言，它的样本值是一串数列…，X_1，X_2，X_3，…，它们分别是随机变量…，x_1，x_2，x_3，…各自的样本值，或称为时间序列$\{x_t\}$的一个现实$\{X_t\}$。在大量的实际问题中，因为随时间t的流逝而不能再重复，往往仅能获得时间序列的一个现实。我们通常称长度为N的动态数据X_1，X_2，X_3，…，X_N为时间序列$\{x_t\}$的长度为N的样本值。熟悉这一概念后，为了简化符号，今后本书将时间序列$\{x_t\}$与它相应的样本值$\{X_t\}$都统一地用$\{x_t\}$表示，而不再用两类字母区别表示了。

对气象工作者，时间序列x_1，x_2，x_3，…，x_N可能是一串旬降雨量数据，而对医师则可能是一串脑电波数据，……。这些数据的特点是明显地承认观测值的排列顺序的重要性，正是数据的顺序与大小反映了数据所包含的信息，反映了数据内部的相互联系。事实上，所获得的数据最为重要和有用的特性，就是观测值之间的依赖关系或相关性，正是这种相关性表征了产生这些数据的现象、过程、系统的“动态”或“记忆”。这种相关性一旦被定量地描述出来，就可以从系统的过去值预测其将来的值。不同系统的相关性或动态也彼此不同，因而用某种数学模型来表达这种相关性是很重要的。

用来分析各种相依有序的离散数据集合的方法称为时间序列分析（简称时序分析）。从表面上看，时序分析撇开了系统变量间内存因果关系和结构关系的影响。但事实上，由于时序中反映了曾经发生过的所有因果关系和结构关系的影响。时序分析是从总的方面进行考察，来说明各种作用力的综合作用。因此，当我们所关心的影响因素错综复杂或有关的数据资料无法得到时，就直接采用时间t作为变量来综合地代替这些因素。时间作为一个明确的自变量进入模型，其意义表面上是表示因变量随时间而自发的变化，而实际上是代表了决定因变量变化的诸因素的联合影响。

§ 2　时间序列的分析方法

时间序列分析所研究的对象，是一串随时间变化而又相互关联的数字序列（动态数据）。人们如何定量地描述数据的相关性呢？

传统的时间序列分析法是在时域上估计观测数据$\{x_t\}$的自相关函数，或在频域上估计它的自谱函数（或称功率谱）。

设$\{x_t\}$是一个平稳的、正态分布的、零均值的时间序列，它的统计特性取决于时域中的自协方差函数

$$r_k = E[x_t x_{t-k}] \tag{1.2.1}$$

或频域中的自谱函数

$$S_x(\omega) = F[r_k] = \sum_{k=-\infty}^{\infty} r_k e^{-i\omega k\Delta} \tag{1.2.2}$$

$F[\cdot]$是指傅里叶变换。式中，ω为圆频率；Δ为采样时间间隔即相邻数据的时间间距；k为延迟或相关步数。

但是，当把时间序列分析方法用于实际现象时，人们容易想到，这时只可能获得被观察的随机现象的有限数据，即观察数据个数N或样本值长度$L=N\Delta$是有限的，不可能得到一个无穷长度的完整的样本值序列。因此，实际上不可能通过观察数据计算出自协方差函数与自谱函数的真值r_k与$S_x(\omega)$。传统的时间序列分析方法是直接利用观察数据算出样本自协方差函数$\hat{r}_k$作为真值r_k的估计值，即

$$\hat{r}_k = \frac{1}{N-k} \sum_{t=k+1}^{N} x_t x_{t-k}。 \tag{1.2.3}$$

为了不使估计值$\hat{r}_k$偏离真值r_k太大，一般$k<N/4$，甚至取为$N/10$。根据公式（1.2.3）得到的$\hat{r}_k$去计算$\hat{S}_x(\omega)$时，k的上、下限就只能限制在$N/4$与$-N/4$的范围内。正因为N或L有限，传统的时间序列分析方法存在着固有的缺陷。第一，按传统方法算出的

$\hat{r}_k$和$\hat{S}_x(\omega)$偏离真值远，即样本自协方差函数是对理论自协方差函数的一种很差的估计。第二，按传统方法算出的$\hat{S}_x(\omega)$会发生谱线泄漏，即观测数据（时间序列）中所包含的谐波成分与幅值受到歪曲。例如，不该有的频率成分却有了，应该具备的幅值改变了。第三，按传统方法所能达到的频率分辨力低，大约等于N/L。这一缺点在分析处理长度较短的观测数据时尤为突出。虽然目前已提出了不少克服以上缺陷的方法，但其效果只能是减少而不能消除。

现代的时间序列分析方法是通过另外一种途径一模型法来实现的。它的主要思路是把时间序列看成是随机系统对于不相关的

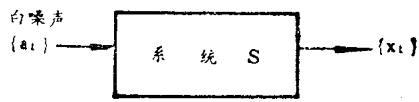

图1.1 动态系统

或独立的“白噪声”输入的反应的一个现实。也就是将时间序列看成是动态系统的输出，而系统的输入是白噪声（图1.1）。这样，动态系统的数学模型就可以把不独立或相关的时间序列输出转化为独立的或不相关的输入。所以现代的时间序列分析方法就归结为寻求这样一种模型，它能实现把不独立数据变成独立数据的转化。然后利用对于独立观测值的统计方法进行估计、预测和控制。

能否找到一个合适的模型去代表或逼近观测数据的相关性呢？换句话说，任何一个随机系统能够建立相应的模型去很好地表示吗？回答是肯定的。理论上已经证明，任何平稳随机系统都能用一个合适的平稳时序模型来逼近到我们所需要的近似程度，这正是建模的任务。

§3 时间序列的参数模型

传统的时间序列分析方法所利用的$\hat{r}_k$和$\hat{S}_x(\omega)$都是非参数模

型，而现代方法是处理动态数据的参数化时域分析方法，是指对观测数据拟合（即给它“配上”）一个参数模型，再利用这个模型对观测数据及产生这一数据的系统进行分析，以便更本质地了解数据的内在结构和系统的动态特征，从而可以利用过去的观测数据对未来值进行预测与控制。

§3.1 建立时序模型的基本思想

时间序列分析的主要想法是认为同一变量在现在时刻的观测值，在时间上同以前的观测值是有联系的。当然在新的时刻会出现未预料的新情况。因此，若记$x_t(t=\cdots, -2, -1, 0, 1, 2, \cdots)$是一个时间上无限伸展的序列，则我们可提出一种描述该序列的模型是

$$x_t=f(x_{t-1}, x_{t-2}, \cdots)+a_t. \tag{1.3.1}$$

这里的函数f把现在的情况同以前的情况联系起来，而a_t表示时刻t出现的新情况，假定它是同t时刻以前的情况无关的随机因素。

模型（1.3.1）的一种常用情况是f取线性形式

$$x_t=\pi_1x_{t-1}+\pi_2x_{t-2}+\cdots+a_t. \tag{1.3.2}$$

通常假定a_t是正态的白噪声序列，其均值为常数，不妨假定为0。这样

$$E[a_t]=0,$$

$$E[a_ta_{t-k}]=\sigma_a^2\delta_k. \tag{1.3.3}$$

式中，δ_k是Kroneker函数，当$k=0$时，$\delta_k=1$，当$k\neq0$时，$\delta_k=0$。显然，σ_a^2就是a_t的方差。式（1.3.3）表明，白噪声序列中的各个随机变量彼此不相关，没有任何统计联系。

用B表示后移算子，即$Bx_t=x_{t-1}$，$B^kx_t=x_{t-k}$，则（1.3.2）式可以写成

$$(1-\pi_1B-\pi_2B^2-\cdots)x_t=a_t \tag{1.3.4}$$

或简记为

$$\pi(B)x_t=a_t.$$

如果把算子$\pi(B)$的逆算子记为

$\psi(B)=\pi^{-1}(B)$，则(1.3.4)式又可写成

$$x_t=\psi(B)a_t. \tag{1.3.5}$$

把$\psi(B)$展成级数

$$\psi(B)=\sum_{j=0}^{\infty}\psi_j B^j, \tag{1.3.6}$$

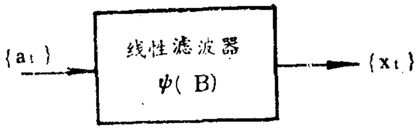

图1.2　线性滤波器

则（1.3.5）式可写成

$$x_t=a_t+\psi_1 a_{t-1}+\psi_2 a_{t-2}+\cdots \tag{1.3.7}$$

这里不妨令$\psi_0\equiv1$。

（1.3.5）式表明一个时间序列$\{x_t\}$可以认为是由一个相互独立的白噪声序列$\{a_t\}$通过一个线性滤波器$\psi(B)$而产生的(图1.2)。从理论上讲，加权序列ψ_1，ψ_2，……可以是有限的也可以是无限的。如果它是有限的，或者虽是无限的，但是收敛的，则这个滤波器就是稳定的，并且时间序列就是平稳的。否则$\{x_t\}$就是非平稳的。

§3.2　自回归(AR)模型

时间序列模型(1.3.2)中只有有限项的模型

$$x_t=\varphi_1 x_{t-1}+\varphi_2 x_{t-2}+\cdots+\varphi_p x_{t-p}+a_t \tag{1.3.8}$$

$$t=1,\ 2,\ 3,\ \cdots$$

称为自回归（Autoregressive）模型，其中p为自回归的阶，φ_1，φ_2，…，φ_p为自回归系数，a_t是均值为零、方差为σ_a^2的正态分布白噪声，即$a_t\sim NID\ (0,\ \sigma_a^2)$，符号NID表示独立正态分布。这样的模型记为AR(p)。

熟悉数理统计的读者都知道，对于回归模型

$$y_t=\beta_1 x_{1t}+\beta_2 x_{2t}+\cdots+\beta_r x_{rt}+\varepsilon_t, \tag{1.3.9}$$

$$t=1,\ 2,\ \cdots,\ N.$$

它表示观测值y_t对于另一组观测值（x_{1t}，x_{2t}，…，x_{rt}）的相关性。模型（1.3.9）将随机变量y_t分解成两部分，一部分是因变量(x_{1t}，x_{2t}，…，x_{rt})。它们所代表的是某些已知的可变化的确定性因素；另一部分，则是残量ε_t，它是由一些不可捉摸的因素

及测量误差产生，通常可假定$\{\varepsilon_t\}$是零均值的独立序列，和前一部分相互独立。因此，观测序列$\{y_t\}$的统计性质由$\{\varepsilon_t\}$来确定，因而$\{y_t\}$是相互独立或不相关的序列，它们都是同一总体y不同次独立随机抽样值。(1.3.9)是一种静态数据模型，它描述的只是一个变量对同一时刻的另一组变量的静态相关。而在模型(1.3.8)中，x_t和x_{t-1}，x_{t-2}，…，x_{t-p}同属于时间序列$\{x_t\}$，是序列中不同时刻的随机变量，它们彼此有一定的相互关系，因而(1.3.8)是一种动态数据模型。该模型所描述的$\{x_t\}$是对其自身的过去的数值进行回归，故称为自回归模型。

AR(p)模型对于a_t有两个假定，一是$\{a_t\}$作为随机序列，在不同时刻互不相关，二是a_t和前时刻的序列观测值$x_k(k<t)$不相关。如果用AR(p)模型去拟合某序列$\{x_t\}$，残差序列$\{a_t\}$不符合上述假定，即$\{a_t\}$不是白噪声，那么用AR(p)描述该序列是不合适的。

用B^p表示p步后移算子，即

$B^p x_t = x_{t-p}$，则（1.3.8）式可写成

$$\varphi(B)x_t = a_t, \tag{1.3.10}$$

其中
$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p. \tag{1.3.11}$$

如果系数多项式$\varphi(B)=0$的根全在单位圆外，即其根的模都大于1，则AR(p)模型是平稳的。

模型(1.3.10)中含有$p+1$个未知参数φ_1，φ_2，…，φ_p和σ_a^2。σ_a^2是白噪声序列$\{a_t\}$的方差。这些未知参数必须利用观测数据来进行估计。只有当以上$p+1$个参数给出之后，AR(p)模型才是完全确定的。

现以最简单的AR（1）模型为例，进一步说明自回归模型的物理意义。

设观测数据x_1，x_2，…，x_N是平稳、零均值的时间序列，它们内部最简单的关系是：x_t只与x_{t-1}线性相关，当然x_t的取值还受其他随机因素的综合影响。即

$$x_t = \varphi_1 x_{t-1} + a_t. \tag{1.3.12}$$

式中φ_1是x_t与x_{t-1}间的线性相关系数；a_t是残差，表示随机因素的综合影响，是白噪声。

一阶自回归模型表明，它将x_t分解为两部分：一部分取决于x_{t-1}，由$\varphi_1 x_{t-1}$表示；另一部分独立于x_{t-1}，由a_t表示。这两者彼此无关。由于在$t-1$时刻，x_t是一未知的随机变量，因此a_t也是一个未知的随机变量，它具有$a_t \sim NID(0, \sigma_a^2)$的特性。一旦在$t$时刻$x_t$被观测得知，则$a_t$不再是随机变量而成为固定的数，它可以由下式计算出来

$$a_t = x_t - \varphi_1 x_{t-1}. \qquad (1.3.13)$$

方程式（1.3.13），简单看来是式（1.3.12）的不同形式，但它有一个很重要的解释：由于x_t是一个相关的序列，而a_t是一个独立的序列，AR(1)模型可以看作是一个把相关数据改造为或变换为独立数据的“装置”，如图1.3所示，这是通过从x_t中将取决于x_{t-1}的部分除去而实现的。

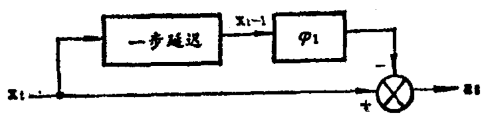

图1.3 AR(1)模型

最后，讨论一下AR(1)模型的二种特殊情况也是很有意义的。如果$\varphi_1 = 0$，则AR(1)模型成为

$$x_t = a_t, \qquad t = 1, 2, \cdots. \qquad (1.3.14)$$

因而x_t实际上是独立的或不相关的序列，则称（1.3.14）式为纯粹随机过程，这是所有时序模型中最简单的模型，它相当于没有“记忆”的过程，即t时刻过程的值和所有过去直到$t-1$时刻的值（实际也包括过程的未来值）都不相关。

AR(1)模型的另一种特殊是$\varphi_1 = 1$时，则（1.3.12）成为

$$x_t = x_{t-1} + a_t, \qquad (1.3.15)$$

即 $$x_t - x_{t-1} = a_t,$$

或 $$\nabla x_t = a_t. \qquad (1.3.16)$$

这里∇表示差分算子。这是当φ_1趋于1时AR(1)模型的一个有趣的极限形式，它表明这一系统具有很大的惯性，即有强的相关或记忆。当x_t从$t-1$时刻移至t时，如果没有一个随机项a_t则它的值（或响应）将保持不变。但就是因为这个a_t，x_t的值才不确定。由于随机项a_t主宰增量∇x_t的大小，所以称（1.3.16）式为随机游动模型。

对于AR(1)模型而言，φ_1的最大值是多少？以后的讨论将告诉我们，如果观测数据或时间序列是平稳的，则

$$|\varphi_1|<1.$$

关于时间序列（或数据）的平稳性，我们做如下的简要说明。如果观测的时间序列被看作是随机过程（它代表产生时间序列的结构方式）的一个可能的现实，它具有以下性质：

1．对所有的时间点具有同样的均值，即

$$E[x_t]=\mu=\text{常数},\qquad \text{对所有}t.$$

2．对所有的时间点具有同样的方差，即

$$E[(x_t-\mu)^2]=\sigma_x^2=\text{常数，对所有}t.$$

3．任何二时间点$(t,t-k)$之间的协方差只取决于时间间隔k，而同这些点在时间轴上的位置无关，即

$$E[x_t,x_{t-k}]=\begin{cases}\sigma_x^2, & k=0;\\ r_k, & k\neq 0.\end{cases}$$

那么，在统计理论中称这样的时间序列$\{x_t\}$为广义平稳过程。从实用观点看，如果时间序列用于谱分析，可以发现功率-频率特性只取决于二阶矩（均值和协方差）特性，因此广义平稳和完全（严格）平稳是差不多的。今后本书提到的平稳时间序列主要是指广义平稳过程。当然，如果过程是高斯分布的，则广义平稳必然意味着严格平稳。

§3.3 滑动（MA）平均模型

时间序列模型（1.3.7）中只有有限项的模型

$$x_t=a_t-\theta_1 a_{t-1}-\cdots-\theta_q a_{t-q},\tag{1.3.17}$$

称为滑动平均（*Moving Average*）模型。这里q为模型的阶次；