《现代应用数学方法》丛书（4）

科学出版社

21

数学天元基金

非线性系统中数据处理的统计方法

项静恬 史久恩 著

《现代应用数学方法》丛书(4)

非线性系统中数据处理的统计方法

项静恬　史久恩 著

科学出版社

内 容 简 介

本书提供了非线性复杂系统中数据处理的分段、分解、综合、降维四类技术，给出了实现上述技术的变点分析、季节调整、组合预测、综合评价及投影跟踪等统计新方法. 这些方法是科研和应用人员榨取数据信息、描述复杂系统的有效工具. 书中深入浅出地介绍了每种方法的思想、原理和程序步骤，通过实例讨论并比较各种方法的特点与应用效果.

本书可供科研人员、大专院校数理统计系师生阅读.

图书在版编目(CIP)数据

非线性系统中数据处理的统计方法/项静恬，史久恩著. 北京：科学出版社，1997

(现代应用数学方法/方开泰主编)

ISBN 7-03-005640-X

Ⅰ. 非… Ⅱ. ①项… ②史… Ⅲ. 非线性系统(自动化)-数据处理-数理统计-方法 Ⅳ. 0212.1

中国版本图书馆 CIP 数据核字 (96) 第 19731 号

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

中国科学院印刷厂印刷

新华书店北京发行所发行 · 各地新华书店经售

*

1997 年 8 月第 一 版	开本：850×1168 1/32
1997 年 8 月第一次印刷	印张：7 7/8
印数：1—1 600	字数：201 000

定 价：12.00 元

《现代应用数学方法》丛书

总　序

应用数学的发展与自然科学和社会科学有着广泛的交叉和渗透. 一方面，它为形形色色的物理、化学、生物、社会等现象提供描述和分析的数学工具. 另一方面，这些实际问题的解决又为数学学科的发展提供了动力和永不枯竭的源泉. 许多成功的应用数学方法，如解非线性方程的牛顿-高斯法、曲线拟合的最小二乘法、线性规划的单纯形法等，成了当今应用数学工作者手中不可缺少的工具. 它们之所以有如此强大的生命力，原因在于方法本身有坚实的理论基础，同时又有鲜明的应用背景，能用于不同的领域. 因此，成功的应用数学方法是理论联系实际的桥梁和纽带.

我国的数学要达到世界先进水平，要对人类有所贡献，一个重要的方面是要有一批独创的应用数学方法.《现代应用数学方法》丛书的出版，希望能为鼓励和促进我国的数学工作者创造或介绍更多的现代应用数学方法增加一个舞台.

这套丛书的宗旨是介绍现代应用数学方法. 这些方法应该是目前世界上最先进的，或者是我国独创的，或者是国外已经普遍使用但国内知之甚少的方法. 丛书着重阐明所介绍方法的应用背景和思想，避开深奥的数学论证，力求深入浅出、图文并茂. 有数值及应用性的例子，使读者易于理解和使用. 丛书要求短小精练，突出新的方法，不求齐全，一般 10 万字左右. 书末所附的文献将指出方法的理论背景以及最近的进展. 以便读者进一步深入研究.

本丛书的出版得到国家"天元"项目的资助，得到科学出版社的大力支持，是全体编委努力的成果. 我们要特别感谢许多作者在百忙中为丛书撰写文稿，付出了辛勤的劳动. 我们希望这套丛书的

出版对我国应用数学的发展起到促进作用，衷心地希望丛书成为广大读者的良师益友.

胡国定(南开大学)
方开泰(中国科学院应用数学研究所)

前　言

近年来接触资源、环境、运输、经济等方面的实际课题，对象全都是内部结构复杂、指标和因素较多的非线性系统. 对于此类系统结构及输入、输出的模拟、预测和调控，经典的数理统计工具往往不能满足需要. 在完成众多科研及应用课题的过程中，我们研究并积累了不少模拟复杂序列、描述系统非线性结构的数理统计与时间序列方法，并将之归纳为数据处理的分段、分解、综合和降维四类技术. 现将这些方法和技术整理提炼，集成本书以飨读者.

全书共分五章. 第一章介绍判断数据结构突变的变点分析方法，通过寻找系统的模型变点及系统输入、输出的均值变点或概率变点，来实现描述数据突变的“分段”处理技术；第二章讨论的季节调整技术，能将复杂的非线性序列“分解”为仅含趋势、周期或随机成分的子序列，并分别实现模拟与预测；第三章提供了描述系统复杂结构及众多变量的“综合”技术，通过构造组合模型、集团因子和综合指标来合理地模拟系统，提高分析、预测和评价的科学性；第四章涉及的投影跟踪方法，是对多变量系统中高维输入或输出实现“降维”技术的有效工具. 作为以上四类技术的全面运用，本书最后一章结合实例给出了对多指标复杂系统建立调控、预测、评价、综合体系的方法与过程.

之所以从近代数理统计工具中选择以上四类技术介绍给读者，是因为本书作者在近十多年应用实践中感到这些技术应用最广泛、最有效，也较容易被广大应用工作者理解和运用. 书中提供的变点分析、季节调整、组合预测、综合评价及投影跟踪等方法，都是近期国内外统计研究的新分支新成果，国内尚缺或少有专著系统地介绍与提供. 希望本书的问世能对广大从事数据处理的应用工作者有所帮助.

为了帮助各行各业数据处理的应用人员尽快掌握并使用上述各种方法和技术，本书力求做到深入浅出. 不仅提供每种方法的基本思想和实现步骤，同时还结合实例讨论方法的应用. 读者只需具备数理统计的基础，即可读通方法并能用于本专业的数据分析. 书中尽量避免数学理论的证明和复杂公式的推导. 为了满足部分读者和院校教学的需要，书中对一些仅涉及基础数学工具的理论给予证明，且对涉及较深知识和较广思路的内容提供了相应的参考文献. 以应用为目的的读者，完全可以跳过这些节段，丝毫不会影响阅读的完整性和方法的运用. 总之，作者希望本书能适应尽多读者的需要与兴趣，并成为广大应用工作者处理数据榨取信息的有效工具.

作者感谢方开泰研究员，是他约请作者撰写本书；感谢陈希孺、张建中、程侃教授，他们在百忙中审阅本书，提供了具体的修改建议；感谢潘一民、顾岚、陈忠琏等教授，他们的研究成果丰富了书中的方法和实例；感谢气象、地理、天文、经济、运输等众多课题协作人员，他们各专业的数据处理难题促进了统计方法的研究，又在实际课题中验证了方法的应用效果. 最后，作者感谢所有关心本书运用本书的读者，希望对书中的错误和不足提出宝贵的意见.

目　录

第一章　系统结构变化的判断和检验

系统的非线性结构，常表现为系统模型或输出的某个(些)量突然发生了变化. 在自然界、社会、经济等领域内，突变现象很常见很重要且反映系统的本质. 欲掌握灾害规律，要检查工序质量，都归结于系统的灾变研究.

设 $\{x_t, t = 1,2,\cdots,N\}$ 为非线性系统的输出，其系统模型或输出序列在某未知时刻起了突然变化，该时刻即称为变点. 变点统计分析的目的是判断和检验变点的存在、位置、个数，并估计出变化的跃度.

近十年来，国内外统计学家研究出若干方法，能有效地判断分析均值变点、概率变点及模型参数变点[1,2]，并在历史气候突变、灾异性地球物理现象等分析预测中获得有成效的应用.

§1.1　变点探索分析

非线性系统的复杂性往往使变点的存在与特性难以分辨，因此在变点分析初阶，有必要采用探索分析方法对之进行直观判断或分类.

探索数据分析法是 70 年代末产生和兴起的统计分析新方法. 其不同于经典统计方法之处，是它无需对数据的统计分布做先验性假定，直接通过运算和图形来观察数据规律及特性，从而为进一步的统计分析提供初步的结论与依据.

我们采用"滚动考察"的办法对非线性系统的输出序列$\{x_t\}$，$t=1,2,\cdots,N$进行变点探索分析. 具体做法是将总序列按长度 n 为一期，以 n_0 为滚动间隔分成若干子序列($n_0 \ll n \ll N$)，其中$\{x_1, \cdots, x_n\}$为第 1 期，$\{x_{n_0+1}, x_{n_0+2}, \cdots, x_{n_0+n}\}$为第 2 期，$\cdots$，$\{x_{(l-1)n_0+1},$

$\cdots, x_{(l-1)n_0+n}\}$ 为第 l 期，对分序列作“滚动”探索分析.

§1.1.1 Tukey 箱线图法

该方法按如下步骤“滚动”实行变点探索分析.

(1) 确定对总序列进行分期的样本长度 n 及“滚动”考察的时间间隔 n_0(例如 $N=1000$ 时取 $n=100, n_0=10$).

(2) 对长度为 n 间隔为 n_0 的各期子序列按数据从小到大升秩排列，并计算下列五种稳健统计量：最小值 I，1/4 分位数 H，中位数 M，3/4 分位数 Q，最大值 A.

(3) 对每期子序列画 Tukey 箱线图(见图 1.1).

(4) 以时间 t 为横轴，数据值为纵轴，在各期起点时间位置上画出箱线图进行比较(图 1.2).

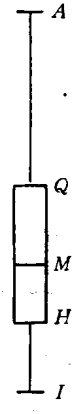

图 1.1 Tukey 箱线图

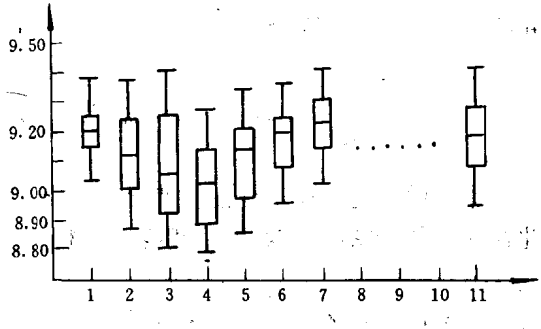

图 1.2 箱线图的滚动比较(示意图)

(5) 调整 n 和 n_0，重复(2)—(4)以探索总序列变点.

由图 1.1 可见，Tukey 箱线图能集中而形象地描述各期子序列的主要统计特性. 用图 1.2 对序列滚动考察，即可对系统模型或输出序列的均值等特征参数的变化及变点位置提供初步判断.

§1.1.2 频数分布图法

§1.1.1 的探索分析法适用于数据值连续的情形. 本节的分段频数"滚动"考察方法,对连续数据与等级数据都可采用,具体的实现步骤为:

(1) 确定"滚动"考察的子序列长度 n 及间隔 n_0.

(2) 将连续型数据按大小分为 J 个区组(等级).

(3) 统计出第 l 期子序列中第 j 级数据的个数 n_{lj},并计算频数 p_{lj}: $p_{lj}=n_{lj}/n, j=1,2,\cdots,J, l=1,2,\cdots,L$.

(4) 以期号 l 为横轴,p_{lj} 值为纵轴作出 J 张同级数据在不同期中频数的分布图(图 1.3),考察其频数分布的变化.

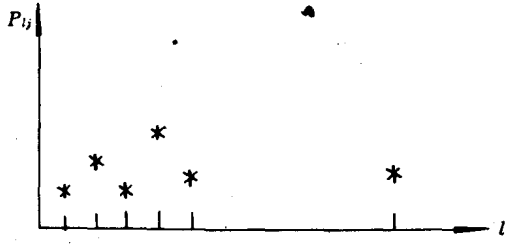

图 1.3 第 j 级频数分布图($j=1,2,\cdots,J$)

(5) 改变 n 和 n_0,重复以上各步进行考察和比较.

该法可为概率变点分析(见§1.3)提供初步的判据,在数据缺值时也能采用,只需在计算频数时分子分母减少相应个数即可.

探索分析的结果,为是否深入进行变点分析提供了初步判断,并为选择分析方法提供了依据.

§1.2 均值变点分析

均值变点是最常见最直观的一种类型. 近年来人们关注的地球气温变化问题,亦可通过对历史气温资料的均值变点分析来寻找规律.

均值变点分析的总体步骤如下：

(1) 检验变点之有无，即检验原假设 H_0：不存在变点. 若 H_0 被接受，则该数据序列没有变点.

(2) 若 H_0 被否定，则假设该序列至多存在 q 个变点，对变点 $m_1, m_2, \cdots, m_q$ 进行估计.

(3) 估计变点个数.

(4) 估计变点处均值的跳跃度.

实际数据分析时，如果先验知识足以肯定变点的存在，亦可免去第(1)步而进入后几步.

§1.2.1 变点有无的检验

一、模型和原假设 H_0

均值变点问题的离散数据模型为：设

$$x_t = a_t + e_t, \qquad t = 1, 2, \cdots, N$$

$$a_1 = \cdots = a_{m_1-1} = b_1,\ a_{m_1} = \cdots = a_{m_2-1} = b_2, \cdots,$$

$$a_{m_q} = \cdots = a_N = b_{q+1} \tag{1.2.1}$$

此处 $1 < m_1 < m_2 < \cdots < m_q \leqslant N$，如果 $b_{j+1} \neq b_j$，则 m_j 就是一个变点. 随机误差 $e_1, e_2, \cdots, e_N$ 假定为独立等方差 σ^2 且有期望值 0. 变点有无的检验，即

$$H_0: b_1 = b_2 = \cdots = b_{q+1} \tag{1.2.2}$$

需要特别强调指出的是，此检验中 $m_1, m_2, \cdots, m_q$ 为未知，且这正是本检验与通常多样本检验的差别.

二、检验的方法与步骤

原假设 H_0 的检验由以下各步运算来进行：

(1) 令 $i = 2, \cdots, N$，对每个 i 将样本分为两段：

$$x_1, x_2, \cdots, x_{i-1} \text{ 和 } x_i, x_{i+1}, \cdots, x_N$$

计算每段样本的算术平均值 $\overline{X}_{i1}$ 和 $\overline{X}_{i2}$ 及统计量：

$$S_i = \sum_{t=1}^{i-1}(x_t - \overline{X}_{i1})^2 + \sum_{t=i}^{N}(x_t - \overline{X}_{i2})^2 \qquad (1.2.3)$$

(2) 计算统计量

$$\overline{X} = \sum_{t=1}^{N} x_t/N \text{ 和 } S = \sum_{t=1}^{N}(x_t - \overline{X})^2 \qquad (1.2.4)$$

(3) 计算期望值

$$E(S-S_i),\ i=2,3,\cdots,N$$

$$E(S-S_i)^{①} = E(N^{-1}(i-1)(N-i+1)(\overline{X}_{i1}-\overline{X}_{i2})^2)^{②}$$

$$= \sigma^2 + N^{-1}(i-1)(N-i+1)(E\overline{X}_{i1} - E\overline{X}_{i2})^2 \qquad (1.2.5)$$

(4) 求极大值

$$E(S - S^*) = \max_{2\leqslant i\leqslant N} E(S - S_i),$$

在平均意义下认为

$$S^* = \min(S_2,\cdots,S_N) \qquad (1.2.6)$$

(5) 取检验显著性水平为 α，计算 C 值[3]：

$$C = \sigma^2(2\log\log N + \log\log\log N - \log\pi - 2\log(-0.5\log(1-\alpha))) \qquad (1.2.7)$$

式中 σ^2 若未知，可用下面的估计来代替：

$$\hat{\sigma}^2 = S^*/(N - 2\log\log N - \log\log\log N - 2.4) \qquad (1.2.8)$$

---

① 由 $S = \sum_{t=1}^{N} x_t^2 - N\overline{X}^2$，

$S_i^2 = \sum_{t=1}^{N} x_t^2 - (i-1)\overline{X}_{i1}^2 - (N-i+1)\overline{X}_{i2}^2$

可得：$S - S_i = (i-1)\overline{X}_{i1}^2 + (N-i+1)\overline{X}_{i2}^2 - N\overline{X}^2$

$= (i-1)\overline{X}_{i1}^2 + (N-i+1)\overline{X}_{i2}^2 - N^{-1}((i-1)\overline{X}_{i1} + (N-i+1)\overline{X}_{i2})^2$

$= N^{-1}(i-1)(N-i+1)(\overline{X}_{i1} - \overline{X}_{i2})^2$

② $E(\overline{X}_{i1}-\overline{X}_{i2})^2 = D(\overline{X}_{i1}-\overline{X}_{i2}) + (E(\overline{X}_{i1}-\overline{X}_{i2}))^2$

$= D\overline{X}_{i1} + D\overline{X}_{i2} + (E\overline{X}_{i1} - E\overline{X}_{i2})^2$

$= (i-1)^{-1}\sigma^2 + (N-i+1)^{-1}\sigma^2 + (E\overline{X}_{i1} - E\overline{X}_{i2})^2$

$= N(i-1)^{-1}(N-i+1)^{-1}\sigma^2 + (E\overline{X}_{i1} - E\overline{X}_{i2})^2$.

(6) 若 $S-S^*>C$，则否定 H_0，认为无变点，否则接受 H_0. 该检验有渐近水平 α.

由检验步骤显然可见，本检验方法的思想还是很直观的，因为平均说来，变点的存在会使 S 和 S_i 的差距增大. 然而本方法对恰有一个变点的检验最为有效，多个变点时有可能因为均值的多次升降而抵消了 S 和 S_i 间的差距. 此外，变点有无的检验也可结合变点估计来进行，这一点将在 §1.2.3 介绍局部比较法时讨论.

§1.2.2 单个变点的估计

一、Mann-Kendall 方法[4]

该法最初由 Mann 提出(1945)并用于序列平稳性的非参数检验，后由 Goossens 等将其发展用于单个均值变点的估计. 具体步骤如下：

(1) 对序列 $\{x_t\}$，$t=1,2,\cdots,l,l\leqslant N$ 构造统计量：

$$d_l=\sum_{j=1}^{l}\sum_{i=1}^{j-1}a_{ij},\ a_{ij}=\begin{cases}1, & \text{当 } i<j \text{ 时 } x_i<x_j\\ 0, & \text{其它}\end{cases} \tag{1.2.9}$$

统计量 d_l 表示长度为 l 的序列 $x_1,\cdots,x_l$ 中按大小顺序排列的样本个数，因而称为顺序统计量.

(2) 令 $l=1,2,\cdots,N$，计算 N 个统计量 $U(d_l)$ 并作图

$$U(d_l)=(d_l-E(d_l))/\sqrt{\mathrm{Var}\,d_l}$$

式中 $E(d_l)$ 和 $\mathrm{Var}\,d_l$ 在 $x_1,\cdots,x_N$ 相互独立且有相同连续分布时可计算得为(证明见本小节末尾)

$$\begin{aligned}E(d_l)&=l(l-1)/4\\ \mathrm{Var}\,d_l&=l(l-1)(2l+5)/72\end{aligned} \tag{1.2.10}$$

此时 $U(d_l)$(l 固定时)渐近服从 $\mathcal{N}(0,1)$ 分布($l>10$ 即够).

(3) 将序列 $\{x_t\}$ 反向构成序列 $\{x_t'\}$，$x_t'=x_{N-t+1}$. 对 $\{x_t'\}$ 重复前两步运算，得统计量 $U'(d_l)$，$l=1,2\cdots,N$. 令 $U^*(d_l)=-U'(d_{l'})$，$l'=N-l+1$. 将 $U^*(d_l)$ 与 $U(d_l)$ 画于同张图.

(4) 找出 $U(d_l)$ 与 $U^*(d_l)$ 两线的交点 l^*，如果该点处 U 值满

足$|U|<1.96$，则可接受l^*为变点的假设，检验置信水平$\alpha=0.05$.

该方法思想直观、计算方便.然而不宜用于多个变点的情形，也不适用于等级数据序列.

二、最小方差方法

当$q=1$时，由§1.2.1中的最小方差可直接估计单个变点，估计值即为(1.2.6)式中S^*相应的i值.

以下给出(1.2.10)式的证明：

设$x_1,\cdots,x_n$为有相同连续分布函数$F(x)$的独立随机变量

令$a(x)=\begin{cases}1 & x>0\\ 0 & x\leqslant 0,\end{cases}$ $\quad d_n=\sum_{k=1}^{n}\sum_{i=1}^{k-1}a(x_k-x_i)$

注意

$$E[a(x_k-x_i)]=\int_{-\infty}^{\infty}dF(y)\int_{-\infty}^{\infty}a(y-x)dF(x)$$

$$=\int_{-\infty}^{\infty}dF(y)\int_{-\infty}^{y}dF(x)\int_{-\infty}^{\infty}F(y)dF(y)=\frac{1}{2},k>i$$

同理有

$$E[a(x_l-x_j)a(x_k-x_i)]$$

$$\begin{cases}=\int_{-\infty}^{\infty}F(x)dF(x)=\frac{1}{2}, & l=k>j=i\\ =\int_{-\infty}^{\infty}(F(x))^2dF(x)=\frac{1}{3}, & l=k>j\neq i\\ =\int_{-\infty}^{\infty}(1-F(x))^2dF(x)=\frac{1}{3}, & l\neq k>j=i\\ =\int_{-\infty}^{\infty}F(x)(1-F(x))dF(x)=\frac{1}{6}, & l>k=j>i\\ =\left[\int_{-\infty}^{\infty}F(x)dF(x)\right]^2=\frac{1}{4}, & l,k,j,i\text{互不相同}\end{cases}$$

由此即可算出