目　录

第一章　绪论

本章对气候统计诊断和预测的基本问题、内容和方法分别作一概述，对它们之间的相互联系作一鸟瞰式介绍。这些内容是阅读本书的基础。

§1.1　气候统计诊断概述

1.1.1　气候统计诊断的含义

这里给出有关名词的含义。

诊断　“诊断”一词源于医学。医生通过对病人的了解和检查。比如：中医用询问、切脉、看舌苔等办法；西医则用测压仪、X光透视、超声波等仪器检查，从而判断病人所患何种疾病及所患疾病的原因、部位、性质及其病情程度，这一过程称为诊断。

统计诊断　“统计诊断”是指对统计建模及统计推断过程进行诊断。它是20世纪70年代中期才发展起来的一门统计学新分支[1]。统计诊断是对收集起来的数据、以数据为基础建立的模型及相应的推断方法的合理性进行分析。通过一些统计量来检查数据、模型及推断方法中可能存在的“病患”，提出“治疗”办法。为了克服模型与客观实际之间可能存在的差异，需要寻找一种诊断方法，判断实际数据与模型之间是否存在较大偏离，并采取相应对策。这就是数理统计意义下诊断的基本内容。通过统计诊断，找出严重偏离模型的异常点，区分出对于统计推断影响特别大的强影响点。其中对多元线性回归

的诊断是统计诊断的主要内容之一。

气候统计诊断 将“诊断”一词引入到气候学研究中，用某些手段根据气候观测资料对气候变化与气候异常的程度与成因作出判断，即称之为气候诊断。由于是用统计手段进行气候诊断，故将这种诊断称为气候统计诊断。

可见，气候统计诊断的含义与统计学分支——统计诊断的含义是不同的。前者是用统计学方法对气候过程进行诊断，而后者是对统计建模与统计推断过程中可能出现的问题进行诊断。

另一方面，气候统计诊断与通常的气候统计分析也有一些区别。气候统计分析是根据大量气候资料用概率论与数理统计中的方法，研究气候演变的时空变化特征和规律。气候统计诊断除了进行以上一系列分析外，还要进行一系列的科学综合和推断，期望通过统计方法这一气候诊断的重要手段对气候变化与气候异常及其成因作出正确判断[2~3]。例如，气候研究中所谓气候变化归因问题，就是气候诊断的重要内容。在实际应用时，气候统计分析与诊断又往往很难区分开。

1.1.2 气候统计诊断研究的内容

概括地讲，采用统计方法进行气候诊断研究，主要包括以下几方面内容：

(1)应用统计方法了解区域性或全球性气候变化的时空分布特征、变化规律及气候异常的程度。主要针对月以上至几十年时间尺度的变化，即主要研究月、季、年及年代4个时间尺度的气候变化。

(2)通过统计方法探索气候变量之间及与其它物理因素之间的联系，以此研究造成气候异常的原因，进而探索气候异常形成的物理机制。

(3)对气候数值模拟结果与实际变化状况之间的差异进行统计诊断。

1.1.3 现代气候统计诊断技术窥视

气候诊断使用的统计技术涉及到统计学多个分支,如统计检验、时间序列分析、谱分析、多元分析、变量场展开等等。近年来,在引入气候模式对气候异常和变化进行诊断的同时,气候统计诊断技术也有了长足的发展,引起气候工作者的关注。作者之所以在书名"气候统计诊断"前面加上"现代"一词,用意在于希望本书能够尽量反映国际气候统计诊断技术的现代水平。如同医学诊断一样,随着科学技术的发展,现代诊断技术与经典诊断技术有了显著的不同。新概念、新方法不断涌现,逐步取代了原有的经典诊断方法,为气候诊断研究提供了更科学、更有效的手段,同时也拓宽了人们认识气候系统的视野[4~6]。与经典方法相比,现代统计诊断技术的发展主要体现在如下几方面:

1.1.3.1 气候变化趋势和突变检测

从气候序列中分离气候变化趋势,不仅采用滑动平均、累积距平、线性倾向估计等传统方法,还引入了样条函数等数据拟合的新方法。采用这些方法对气候序列的分段曲线拟合,以便更好地反映其真实的变化趋势。此外,还注重对变化趋势进行显著性检验。

尽管目前突变统计诊断技术还很不成熟,但是针对突变问题,借助统计检验、概率论等发展了一些行之有效的检测方法。例如,气候均值、变率以及事件发生与否的检验。气候诊断研究中使用最多的是均值的统计检验。其中不但使用参数统计检验,而且还使用非参数统计检验。

1.1.3.2　气候振荡分析

近年来，诊断气候振荡的技术发展很快。从不连续的周期图、方差分析、谐波分解发展到连续谱、一维、多维最大熵谱。近年来，又发展了动力重构与经验正交函数相联系的新技术——奇异谱，以及能将不同波长的波幅一目了然地展现在一张二维图像上的小波变换。这些新技术与传统技术相比，分辨率更高、适用性更强，对于揭示气候序列不同时间尺度的振荡特性起到了很大的作用。

1.1.3.3　气候变化时空结构诊断

以经验正交函数为基础的对气候场进行时空分布特征的诊断技术也有了令人注目的发展。针对气候变量场特征分析的不同需要，发展了揭示变量场移动性分布结构的扩展经验正交函数、着重表现空间的相关性分布结构的旋转经验正交函数、可以展示空间行波结构的复经验正交函数和描述动力系统非线性变化特征的主振荡型、循环平稳(Cyclostationary)主振荡型、复主振荡型等等。

1.1.3.4　气候变量场间耦合特征诊断

气候研究中常常遇到两个变量场的相关问题。例如，相隔遥远的不同区域同一时间或不同时间变量场间存在的遥相关、海洋与大气的相互作用、大气环流或下垫面对气温和降水的影响等等问题。过去讨论这些问题多用相关分析。现在将典型相关这一有着坚实的数学基础、推理严谨的两组变量分析工具移植到两变量场耦合特征的诊断中。还提出了从讨论两个场主分量出发的 BP 典型相关分析。同时，奇异值分解也在两场耦合特征研究中广泛使用。

1.1.4　气候统计诊断的一般步骤

利用统计方法进行气候诊断，一般可分为下列几个步骤：

1.1.4.1　收集资料

从研究的实际问题出发，确定统计诊断的对象，收集有关的资料。选取的资料应该准确、精确，并具均一性、代表性和比较性。资料的样本长度和区域范围与所研究问题的对象有关。例如，研究中国气候年代际变化规律，应该收集半个世纪至百年以上的样本；研究准两年振荡则有数十年样本就够了。在研究气候场空间结构时，选取某一区域范围内的站点布局应该满足均一性和比较性，否则不能很好地反映变化的真实状况或诊断结果缺乏代表性，且难以比较各区域各时期气候特征的差异。

1.1.4.2　资料预处理

对于收集起来的资料，根据研究问题的具体需要进行预处理。各个气候变量的单位不相同，平均值和标准差亦不相同。为了使它们变为同一水平无量纲的变量，通常要对资料进行标准化处理。标准化的变量均值为 0，方差为 1。有时根据实际情况对资料作距平化处理，给研究带来便利。

1.1.4.3　选择诊断方法

根据研究目的和研究对象，选择合适的诊断方法进行研究。选择不恰当的方法会给研究和物理解释带来困难。例如：研究大气准两年振荡的时空演变特征，将变量场进行带通滤波后，使用扩展的经验正交函数可以展现出不同位相准两年振荡的变化。对于这种研究目的，其它方法则无法做到。再例如：划分气候区域的研究，使用旋转经验正交函数，按照分离出旋转典型空间模态的高荷载区可以进行客观的区域划分，而使用普通的经验正交函数很难收到理想的效果。

1.1.4.4　科学综合和诊断

气候统计诊断是统计学与气候学间的交叉学科。利用统

计方法进行气候诊断，不能陷入盲目套用计算公式。在一些情况下，对计算结果应该进行显著性检验。没有统计意义的结果是失真的，没有分析的价值。这一点往往被人们所忽视。要得出科学的结论，重要的是运用深厚的气候学知识，对计算结果进行科学的综合和细致的分析。如同诊断疾病，统计计算结果好比X光透视片子或超声波图像，要确定所患何种疾病及其部位、性质、病情程度，需要医生凭借医学知识和临床经验，对这些检查结果进行综合分析，才能得出正确的诊断，这是医术是否高明的重要标志。同样在气候统计诊断中，对统计计算结果需用气候专业知识进行判断、识别真伪，概括出气候系统确实存在的事实以及彼此间的联系。

§1.2 气候统计预测概述

1.2.1 气候统计预测的一般概念

按照统计学的观点，利用统计模型估计随机动态系统未来可能出现的行为或状态，称为统计预测。具体地讲，统计预测是利用历史与现时的观测值 $x_1, x_2, \cdots, x_{t-1}, x_t$（$t$ 为现在时刻），估计这个随机系统未来 m 个时刻的状态值 $x_{t+1}, x_{t+2}, x_{t+m}$。可见，统计预测是以系统的过去和现在的信息为基础，对未来时刻作出估计。利用统计模型对气候系统的未来变化状态作出估计，即为气候统计预测。当然，统计模型是在利用大量过去气候资料对气候系统内部或与其它变量之间关系的变化规律及特征的分析基础之上建立的。

1.2.2 气候统计预测的基本假设

在使用统计模型对气候系统未来状态进行统计预测时，隐含着一个基本假设——气候系统的未来状态类似于过去和

现在。这一假设体现在利用统计模型对未来状态进行预测时，是假设模型结构在预测期间内保持不变、气候系统变化及与各变量之间的相关关系在预测期间不变。

从统计学理论上讲，气候统计预测的基本假设应该满足以下两个条件：

(1)气候变化的成因和物理机制至少在预测期间与观测时期一致。

(2)气候系统在预测期间保持稳定。由于气候系统具有一定的概率特性、因果特性和相关特性，因此气候预测很大程度上依赖于统计预测。但是，由于统计预测是在假设系统未来仍按过去和现在的特性变化的前提之下，一旦气候系统出现异常甚至突变或影响气候系统的因素有所改变时，往往导致预测失败。因此，较高技巧的预测源于对气候系统变化特性的深刻了解和认识。

1.2.3 气候统计预测的基本要素

气候统计预测过程主要由以下几个要素构成：

1.2.3.1 预测对象

预测对象是指欲预测的气候要素。比如对某区域旱涝趋势、冷暖趋势或夏季降水量或某月气温等等进行预测。可以是某一测站的局地预测，也可以是大范围区域乃至全球性预测。

1.2.3.2 预测依据

在气候系统内部或影响其变化要素相互关系的诊断基础上提供的预测依据。通常为从某些统计上显著相关的预报因子群中提取的有效信息。

1.2.3.3 预测技术

根据数据性质及预测对象、预测因子的特点，选择合适的统计预测模型。

1.2.3.4　预测结果

对未来气候变化状态时间、空间、数量、性质等方面的预测。

在以上四个基本要素中，第三个要素包含的内容最丰富。

1.2.4　预测技术窥视

统计预测技术形式繁多、分类方式也是多种多样的。若按照预测性质划分，预测技术可以分为两大类，即定性预测与定量预测。

1.2.4.1　定性预测

定性预测方法主要依赖气候专家的主观认识能力，综合地分析过去、现在和将来可能出现的各种因素之间的相互影响，寻找气候要素的发展规律，对未来的发展趋势和性质作出推断。例如：气候学家根据气候学知识，对海温与副热带高压、青藏高原的热力作用、西风带环流及东亚季风等因素的过去、现在和未来可能出现的状态进行综合分析，寻找它们影响中国夏季气候异常特别是降水异常的演变规律，对未来夏季气候是否异常、异常的程度如何作出定性的推断。

定性预测纯数学的处理手段较少，所需资料数量也不必很多，但并不意味着定性预测不需要数量分析。除气候学专业知识和预测经验外，仍然需要一定的数量分析和统计处理，使预测更科学、更可靠。例如：气候预测专家们在对1998年全国夏季降水趋势作预测时，抓住了几个影响我国气候的主要因素：1997～1998年间热带太平洋海水出现异常增温、青藏高原出现空前的大雪、东亚冬季风异常偏弱、西太平洋副热带高压异常偏强等等。根据对这些异常现象的分析，得出1998年夏季长江中下游及江南北部可能出现严重洪涝的预测。其实，这个定性预测中包含了一定的统计处理。海温出现异常，确定

出现厄尔尼诺现象、确定季风偏弱、副高偏强等等均需要用一定的统计标准来确定。

1.2.4.2 定量预测

定性预测技术是对预测对象的变化趋势、发生异常的可能性及其程度作出判断。定量预测技术则是根据足够的历史数据资料，运用科学的方法建立数学模型，对预测对象未来的变化数量特征作出预测。本书所描述的就是这类通过建立概率统计数学模型进行预测的方法。气候统计预测使用的这类预测技术大致可以概括为以下几大类：

(1)时间序列模型。时间序列数学模型是描述序列自身演变规律的模型。时间序列可分为趋势项、周期项和随机项三部分。随机项通常用线性模型来描述，这类模型包括自回归(AR)、滑动平均(MA)、自回归滑动平均(ARMA)、自回归求和滑动平均(ARIMA)模型等等。其中发展较为完善的是BOX-Jenkins 途径的 ARMA(p,q)模型。其表达形式为：

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i} \qquad (1.1)$$

其中 p,q 分别为 AR 模型和 MA 模型的阶数。若 $\theta_i \equiv 0$，(1.1)式变为 AR 模型；若 $\varphi_i \equiv 0$，则(1.1)式变为 MA 模型。

描述非线性现象时，可以使用门限自回归(TAR)等非线性模型。

马尔柯夫链、方差分析周期叠加，也是这类时间序列模型。

以上时间序列模型在许多气象统计预测专著中均有较详细的介绍[7-8]，本书不再赘述。

魏凤英等提出了用多元分析手段解决时间序列预测问题的均生函数模型，为多步的短期气候预测开辟了一条新途

径[9~10]。在第十章中将对这一方法的思路及其在气候预测中的应用作较全面的叙述。

(2)动态系统模型。气候作为一个随机系统，它的状态大多并不是严格平稳的，甚至为非平稳的。Kalman 滤波可以用于描述非平稳的系统，它实质上是用一个最优递推数据处理算法建立自适应模型[11]。Kalman 滤波目前多用于短期天气的 MOS 预报中，也有人尝试用在短期气候预测中。

灰色动态模型在气候突变预测中也有一定效用[12~13]。

另外，动态系统的多层递阶预测模型亦在气候预测中使用。其基本思想是把具有时变参数的动态系统的状态预测分离成对时变参数和对系统状态两部分预测[14]。克服了回归方法中用固定参数模型来预测动态系统状态的局限性。

(3)多元回归模型。在气候预测中应用十分广泛的多元回归模型是在系统的动态方程不清楚时，描述变量之间线性关系最有效的数学模型[15]。其一般表达式为：

$$y = b_0 + \sum_{k=1}^{m} b_k x_k \qquad (1.2)$$

其中 y 为因变量(预报量、预报对象)；x_k 为影响 y 的自变量(预报因子)，$k=1,2,\cdots,m$；b_0 为回归常数；b_k 为回归系数。通常采用最小二乘法来估计回归系数。

选择最优回归方程较常用的算法是逐步回归。不过，在计算机资源十分丰富的今天，完全可以从所有可能子集回归中选择最优回归。针对不同预测问题的要求和数据存在的缺陷，发展了与最小二乘法估计不同思路的主成分回归、特征根回归及岭回归等模型。在本书中将对这部分内容作较详尽的介绍，给出在气候预测中使用的实例。

(4)变量场预测的方法。气候预测中经常遇到变量场水平

分布预测问题。预测对象是一个空间变量场，因子也为空间场。以单点资料为基础的回归分析，着眼于单点气候变量变化的统计规律，没有考虑点与点之间的相互联系，导致水平分布预测结果有时出现无法解释的跳跃。因此，变量场水平分布预测可以采用变量场展开的统计方法。其思路是，把变量场展开成各种典型特征向量与其时间系数的乘积和。假定在一定时间内，空间典型向量是稳定不变的。这时典型特征向量的时间系数变化反映了变量场随时间的变化。只要预测出未来时刻的时间系数，乘以典型特征向量就可以得到变量场的预测。常用的变量场预测展开方法有：经验正交函数展开、车贝雪夫多项式展开及典型相关等等。

(5)神经网络。注意到，近两年国内外文献中出现了将神经网络用于气候预测的研究成果[16~17]。神经网络方法是目前国际上的热点学科之一，其包含的内容十分广泛，算法也十分繁多。它以其独特的结构和处理信息方式，在许多应用领域取得了显著的成效，特别是在处理非线性问题上显示出较强的能力。神经网络预测模型的参数是网络对输入的原始数据进行不断学习训练得到的。神经网络技术是人工智能的一个分支，它并不属于概率论与数理统计，虽然其预测模型也是用观测历史数据来构建的。故本书不涉及这部分内容。

1.2.4.3 历史演变法

讲到气候预测技术不能不提到杨鉴初早在1951年提出的历史演变法[18]。从预测性质角度来分类，这一方法既可以归于定性预测，又可以算为定量预测之列。在当初气象资料十分匮乏的情况下，这一方法为当时我国的长期天气预报起到了非常积极的作用。由于它的实用性和概括性，至今仍沿用这一方法的思路制作短期气候预测。

历史演变法揭示了气候变量序列的五个特性，即持续性、相似性、周期性、最大最小可能性和转折点。持续性即气候变量的历史变化中升降趋势的持久程度；相似性为气候变化在某一时期与另一时期变化形势相似；周期性是气候变化趋势经一定时间间隔后重复出现。以上三个特性反映了气候变量变化过程中历史特征的某种重现。最大最小可能性则指气候变量历史变化的数量在一定时间内有其适当范围，给出了历史变化的概率特性；转折点则是气候变量变化中某一时期明显的特征，在另一时期有所改变，并可能出现新的特征，发生质的突变。以历史演变的五个特征及其它们的相互配合作用为依据，对气候变量未来的变化状态作出推断。

1.2.5 气候统计预测的基本步骤

气候统计预测的基本步骤如下：

1.2.5.1 收集资料

收集预测对象及预报因子资料。资料的样本量 n 应该大于 30。$n \geqslant 30$ 是根据数理统计中的大数定理推断得到的。由于对统计模型的统计检验常常是在变量遵从正态分布假设下的，因此变量资料一般应该满足正态分布，否则预先应进行必要的处理。在建立多元回归预测模型时，要选用符合一定物理规律的因子变量资料。

1.2.5.2 选择统计模型

根据预测对象、预测步长及资料状况，选择适当的统计模型。若不能保证因子变量之间相互独立时，建立回归模型可以考虑用主成分回归、特征根回归、岭回归等方法。反之，则使用一般最小二乘法估计的多元回归已足矣。一步时间序列预测问题可以选用自回归模型，而多步预测用均生函数模型效果会更佳。总之，要根据具体情况来选择统计模型。有时可以对

同一预测问题建立多种不同的统计模型进行预测试验，从中挑选出最符合实际问题的模型。

1.2.5.3 统计检验

对建立的统计模型进行统计检验。对于理论基础和实际应用比较完善的、以多元线性回归为基本形式的模型，已有成熟的检验方法。而有些模型尚没有系统的检验办法。

1.2.5.4 预测

将最临近预测时刻的数据代入到所建立的统计模型中，即可得到未来状态的预测值。

参 考 文 献

[1]韦博成，鲁国斌，史建清．统计诊断引论．南京：东南大学出版社，1991.1～16

[2]王绍武．气候诊断研究．北京：气象出版社，1993

[3]马开玉，陈星，张耀存．气候诊断．北京：气象出版社，1996

[4]H. von storch and A. Nararra (Eds.). Analysis of Climate Variability. Application of Statistical Technology. Berlin, Springer-Verlag, 1995

[5]Wilks D. S. Statistical Method in the Atmospheric Sciences. Academic Press, New York, 1992

[6]魏凤英．现代气候统计诊断和预测程序集——内容、功能和应用实例．见：曹鸿兴等主编．我国短期气候变化及成因研究．北京：气象出版社，1996.98～99

[7]黄嘉佑．气象统计分析与预报方法．北京：气象出版社，1990.271～296

[8]丁裕国，江志红．气象数据时间序列信号处理．北京：气象出版社，1998.55～115

[9]魏凤英，曹鸿兴．长期预测的数学模型及其应用．北京：气象出版

社，1990

[10]魏凤英，曹鸿兴．建立长期预测模型的新方案及其应用．科学通报，1990，35(10)

[11]朱明德，佘光辉．统计预测与控制．北京：中国林业出版社，1993．142～191

[12]曹鸿兴，郑耀文，顾今．灰色系统浅述．北京：气象出版社，1988

[13]魏凤英．月灰色动态模型试作全国温度气候预测．科学通报，1988，33(7)

[14]韩志刚．多层递阶预报方法．北京：科学出版社，1988

[15]Montgomary D. C. and E. A. Peck. Introduction to linear Regression Analysis. John Wiley & sons，1982

[16]Fredolin T. Tangang，Benyang Tang. Adam H. Monahan，and William W. Hsieh. Forecasting ENSO：Events：A Neural Network-Extended EOF Approach，J. Climate，1998，(11)：29—41

[17]Fredolin T. Tangang，W. Hsieh William，and Benyang Tang. Forecasting regional sea surface temperature of the tropical pacific by neural network models，with wind stress and sea level pressure as predictors. J. Geophys. Res. 1998. 103

[18]杨鉴初．运用气象要素历史演变的规律性作一年以上的长期预告．气象学报，1953，(24)：100～117

第二章　基本气候状态的统计量

在气候诊断与预测中，需要用统计量来表征基本气候状态的特征。归纳起来，主要包括表示气候变量中心趋势、变化幅度、分布形态和相关程度4类基本统计量。尽管它们是统计学中的基本内容，计算也很简单，但是为了保持叙述的连贯及应用的便利，这里给出几个最常用的统计量。

§2.1　中心趋势统计量

2.1.1　均值

均值是描述某一气候变量样本平均水平的量。它是代表样本取值中心趋势的统计量。均值计算简便，且由中心极限定理可以证明[11]，即使在原始数据不属于正态分布时，均值总是趋于正态分布的。因此，它是气候统计中最常用的一个基本概念。均值亦可以作为变量总体数学期望 μ 的一个估计。如果变量遵从正态分布，其均值则是 μ 的最好估计值。

我们把包含 n 个样本的一个变量 x，

$$x_1, x_2, \cdots, x_i, \cdots, x_n \tag{2.1.1}$$

视为离散随机过程的一个特定的现实。这个过程的均值定义为：

$$\mu_x(n) = \mathrm{E}(x_n) \tag{2.1.2}$$

或写为算术平均值的形式：

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1.3}$$

在计算机上编程计算算术均值，可以按照(2.1.3)式，对

(2.1.1)式的数据直接求和，再作平均得到 $\bar{x}$，亦可以用递推算法。令 $\bar{x}_0=0$，对 $i=1,2,\cdots,n$，计算中间均值

$$\bar{x}_i=\frac{i-1}{i}\bar{x}_{i-1}+\frac{1}{i}x_i=\bar{x}_{i-1}+\frac{1}{i}(x_i-\bar{x}_{i-1}) \quad (2.1.4)$$

最终得到算术均值 $\bar{x}=\bar{x}_n$。直接求和作平均运算量最省，是最常用的算法。递推算法的优点是可以进行实时资料处理，得到一系列中间均值 $\bar{x}_i$，既满足了特殊需求，也避免了增加一个样本又要从头作平均的重复计算。

应用实例[2.1]：用递推算法计算北京 1951～1996 年夏季(6～8 月)降水量序列的均值。这里样本量 $n=46$。数据见表 2.1。按(2.1.4)式递推公式编程计算出中间均值及序列的均值，结果列于表 2.1。

表 2.1　北京 1951～1996 年夏季降水量及中间均值(单位：mm)

年份	原始数据									
1951～1960	249	404	490	848	621	859	382	452	1170	410
1961～1970	411	285	660	520	185	448	484	204	675	456
1971～1980	383	228	528	372	357	578	529	511	554	243
1981～1990	293	466	319	382	620	309	469	545	268	384
1991～1996	559	364	404	697	385	612				
年份	中间均值									
1951～1960	249	327	381	498	522	579	550	538	608	589
1961～1970	572	548	557	554	530	525	522	505	514	511
1971～1980	505	492	494	488	483	487	488	489	491	483
1981～1990	477	477	472	469	474	475	474	476	471	469
1991～1996	471	468	467	472	470	473				

表 2.1 中的最后一个均值 $\bar{x}_{46}=473$，即为北京夏季降水量 46 年的平均值。

2.1.2　中位数

中位数是表征气候变量中心趋势的另一个统计量。在按大小顺序排列的气候变量 $x_1,x_2,\cdots,x_n$ 中，位置居中的那个

数就是中位数。当样本量 n 为偶数时，不存在居中的数，中位数取最中间两个数的平均值。

中位数的优点在于它不易受异常值的干扰。在样本量较小的情况下，这一点显得尤为显著。对于一个基本遵从正态分布的变量，异常值会对均值产生十分明显的影响。但是，使用中位数就不会受这种影响。

§2.2 变化幅度统计量

统计量均值和中位数描述的仅仅是气候变量分布中心在数值上的大小。换言之，它们只告诉我们气候变量变化的平均水平，却没有告诉这种变化与正常情况的偏差和变化的波动。因此，必须藉助于离散特征量，即表征距离分布中心远近程度的统计量。

2.2.1 距平

最常用的表示气候变量偏离正常情况的量是距平。一组数据的某一个数 x_i 与均值 $\bar{x}$ 之间的差就是距平，即

$$x_i - \bar{x} \tag{2.2.1}$$

气候变量的一组数据 $x_1, x_2, \cdots, x_n$ 与其均值的差异就构成了距平序列

$$x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x} \tag{2.2.2}$$

在气候诊断分析中，常用(2.2.2)式距平序列来代替气候变量本身的观测数据。任何气候变量序列，经过距平化处理，都可以化为平均值为 0 的序列。这样处理给分析带来便利，计算结果更直观。

2.2.2 方差与标准差

方差和标准差是描述样本中数据与以均值 $\bar{x}$ 为中心的平均

振动幅度的特征量，这里分别记为 s^2 和 s。它们亦可作为变量总体方差 σ^2 和标准差 σ 的估计。在气象中也常称标准差为均方差。

方差的计算公式为：

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (2.2.3)$$

标准差为：

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2.2.4)$$

在计算机上计算方差，可以直接用(2.2.3)式计算。但在处理实时资料时，采纳递推算法可以减少不小的计算量，即令 $\bar{x}_0 = 0, s_0^2 = 0$，对 $i = 1, 2, \cdots, n$，用(2.1.4)式递推算出中间均值 $\bar{x}_i$，计算中间方差

$$s_i^2 = \frac{i-1}{i}\left[s_{i-1}^2 + \frac{1}{i}(x_i - \bar{x}_{i-1})^2\right] \qquad (2.2.5)$$

最终得到 $s^2 = s_n^2$。当样本量 n 很大时，递推算法的计算量比直接计算要小得多。

应用实例[2.2]：用递推算法计算北京 1951～1996 年夏季降水量序列的方差。数据见表 2.1。这里样本量仍为 $n=46$。计算结果列于表 2.2。表中最后一个方差 $s_{46}^2 = 33\,748$ 为整个序列的方差。

表 2.2　北京 1951～1996 年夏季降水量的中间方差(单位：mm)

年份	中间均值									
1951～1960	0	6006	9945	48350	41111	49995	47581	42693	77383	73185
1961～1970	69136	69682	65206	60639	65085	61409	57888	59986	58277	55521
1971～1980	53616	54497	52182	50598	49238	47676	45974	44349	42960	43517
1981～1990	43243	41895	41356	40371	39848	38775	37728	36862	37001	36260
1991～1996	35570	34989	34269	34666	34061	33748				

§2.3 分布特征统计量

偏度系数和峰度系数

偏度系数和峰度系数是描述气候变量分布特征的两个重要统计量。偏度系数表征分布形态与平均值偏离的程度，作为分布不对称的测度。峰度系数则表征分布形态图形顶峰的凸平度。为了进行统计检验的便利，这里给出标准偏度系数和峰度系数的计算公式。

偏度系数为：

$$g_1 = \sqrt{\frac{1}{6n}} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \qquad (2.3.1)$$

峰度系数为：

$$g_2 = \sqrt{\frac{n}{24}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - 3 \right] \qquad (2.3.2)$$

(2.3.1)和(2.3.2)式中的 $\bar{x}$ 和 s 分别由(2.1.3)和(2.2.4)式算出。

标准偏度系数的意义是由 g_1 取值符号而定的。当 g_1 为正时，表明分布图形的顶峰偏左，称为正偏度；当 g_1 为负时，分布图形的顶峰偏右，称为负偏度；当 g_1 为 0 时，表明分布图形对称。

标准峰度系数的意义为：当 g_2 为正时，表明分布图形坡度偏陡；当 g_2 为负时，图形坡度平缓；当 g_2 为 0 时，坡度正好。

若 $g_1=0, g_2=0$ 时，表明研究的变量为理想正态分布变量。由此可见，利用 g_1 和 g_2 值测定出偏离 0 的程度，以此确定变量是否遵从正态分布。实际应用时，对 g_1 和 g_2 进行统计检验，以判断变量是否近似正态分布。

应用实例[2.3]：天津1951～1996年夏季(6～8月)降水量资料见表2.3，计算其偏度系数和峰度系数。

表2.3　天津1951～1996年夏季降水量(单位：mm)

年份	降水量									
1951～1960	216	251	613	680	421	412	397	299	435	420
1961～1970	493	365	239	561	341	633	408	148	567	399
1971～1980	358	171	545	363	528	384	777	569	417	241
1981～1990	469	250	237	465	421	437	362	513	236	337
1991～1996	350	253	417	564	484	328				

利用(2.3.1)和(2.3.2)式算出 $g_1=0.96$，$g_2=-0.17$。计算结果表明，天津夏季降水量的分布图形顶峰向左偏，坡度稍平。要判定天津夏季降水量是否遵从正态分布或近似正态分布，还需进一步作分布的统计检验。

§2.4　相关统计量

2.4.1　Pearson相关系数

Pearson相关系数是描述两个随机变量线性相关的统计量，一般简称为相关系数或点相关系数，用 r 来表示。它也作为两总体相关系数 ρ 的估计。

设有两个变量

$$\begin{matrix} x_1, x_2, \cdots, x_n \\ y_1, y_2, \cdots, y_n \end{matrix} \tag{2.4.1}$$

相关系数计算公式为：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2.4.2}$$

也可以用标准差形式计算：

$$r=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\mathrm{cov}(x,y)}{s_x s_y} \tag{2.4.3}$$

式中分母为变量 x 和 y 的标准差，分子为两变量 x,y 的协方差。在已经算出标准差的情况下，(2.4.3)式的计算变得十分简便。

容易证明，相关系数 r 的取值在 $-1.0\sim+1.0$ 之间。当 $r>0$ 时，表明两变量呈正相关，越接近 1.0，正相关越显著；当 $r<0$ 时，表明两变量呈负相关，越接近 -1.0，负相关越显著；当 $r=0$ 时，则表示两变量相互独立。当然，计算出的相关系数是否显著，需要经过显著性检验。

如果观测的数据不是确定的数值，而只是序号或两变量呈非线性关系时，我们不能随便去套用 Pearson 相关系数的计算公式。可以先作适当的数据变换，然后再进行相关系数计算。对于不是确定数值的数据，可以计算非参数相关——Spearman 秩相关系数或 Kendall 秩相关系数[12]，来考察两变量间的相依关系。顾名思义，它们的计算是依赖于对数据排序求秩而进行的。实际使用很少，这里不作介绍。

据统计学中大样本定理[13]，样本量大于 30 才有统计意义。当样本量较小时，计算所得相关系数可能会离总体相关系数甚远。这时，可以用计算无偏相关系数加以校正。将无偏相关系数记为 r^*，

$$r^*=r\left[1+\frac{1-r^2}{2(n-4)}\right] \tag{2.4.4}$$

应用实例[2.4]：中国 1970～1989 年年(1～12 月)平均气温和冬季(12 月～翌年 2 月)气温等级资料见表 2.4。经统计检验，两个变量均遵从正态分布。计算两变量间 Pearson 相关系数。这里 $n=20$。

表 2.4　中国 1970～1989 年年平均和冬季平均气温等级

年　份	年平均气温等级									
1970～1979	3.40	3.30	3.20	2.90	3.40	2.80	3.60	3.00	2.80	3.00
1980～1989	3.10	3.00	2.90	2.70	3.50	3.20	3.10	2.80	2.90	2.90
年　份	冬季平均气温等级									
1970～1979	3.24	3.14	3.26	2.38	3.32	2.71	2.84	3.94	2.75	1.83
1980～1989	2.80	2.81	2.63	3.20	3.60	3.40	3.07	1.87	2.63	2.47

利用(2.4.2)或(2.4.3)式编程计算，两变量间相关系数 $r=0.47$。由于变量只有 20 个样本，因此，需要用(2.4.4)式作校正。校正后 $r^*=0.48$。经显著性检验，相关系数超过 $\alpha=0.05$ 显著性水平，表明年平均气温与冬季平均气温之间存在显著的正相关关系。

2.4.2　自相关系数

自相关系数是描述某一变量不同时刻之间相关的统计量。将滞后长度为 j 的自相关系数记为 $r(j)$。$r(j)$亦是总体相关系数 $\rho(j)$的渐近无偏估计。不同滞后长度的自相关系数可以帮助我们了解前 j 时刻的信息与其后时刻变化间的联系。由此判断由 x_i 预测 x_{i+j}的可能性。对变量 x，滞后长度为 j 的自相关系数为：

$$r(j)=\frac{1}{n-j}\sum_{i=1}^{n-j}\left(\frac{x_i-\overline{x}}{s}\right)\left(\frac{x_{i+j}-\overline{x}}{s}\right) \qquad (2.4.5)$$

这里 s 为 n 长度时间序列的标准差，s 由(2.2.4)式求出。

设计自相关系数计算程序时，可以采用以下方式：

(1)连续设置滞后长度，即 $j=1,2,\cdots,k$，这样可以得到 k 个不同时刻的自相关系数 $r(1),r(2),\cdots,r(k)$。

(2)视 $i(i=1,2,\cdots,n-j)$时刻的数据为一序列，$i+j(i+j=1+j,2+j,\cdots,n)$时刻的数据为另一序列，分别计算其均值、方差及协方差，从而得到 i 时刻和 $i+j$ 时刻序列间的相关系数。

应用实例[2.5]：分别计算表 2.4 中所列中国 1970～1989 年年平均和冬季平均气温等级的自相关系数 $r_1(j)$和$r_2(j)$。这里 $n=20$，滞后长度 $j=1,2,\cdots,5$。计算结果见表 2.5。

表 2.5　年平均和冬季平均气温等级的自相关系数

j	1	2	3	4	5
$r_1(j)$	−0.1372	0.0655	−0.2384	0.1603	0.1491
$r_2(j)$	0.1101	−0.1940	−0.0899	−0.3575	−0.3050

由表可见，年平均气温等级序列在滞后长度 $j=3$ 时达最大，而冬季序列则在 $j=4$ 时达最大。

2.4.3　关联度

表征气候变量关系密切程度的相关系数是以数理统计为基础，要求足够大的样本量及数据遵从一定的概率分布。灰色关联度是一种相对性排序的量，来源于几何相似，其实质是进行曲线间几何形态的比较。几何形状相近的序列，变化趋势就接近，其关联程度就越高，反之亦然。关联度适合表征小样本变量间的关联程度。灰色系统理论中有绝对值关联度和速率关联度[4]，这里给出一种更适合于气候变量的关联度计算方案[5]，称之为优序度。

设有一因变量 $x_0=\{x_{01},x_{02},\cdots,x_{0n}\}$及 m 个自变量 $x_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$，$i=1,2,\cdots,m$。$n$ 为样本量。对原始数据进行极差标准化处理，即

$$x'_{ij} = \frac{x_{ij} - \min x_{ij}}{\max x_{ij} - \min x_{ij}} \quad \begin{pmatrix} i = 0,1,2,\cdots,m \\ j = 1,2,\cdots,n \end{pmatrix} \tag{2.4.6}$$

为简便起见，下面对标准化后数据仍记为 x_{ij}。

因变量与自变量之间的关联系数为：

$$\xi_{ij} = \frac{1}{1 + a_i(x_{0j} - x_{ij})^2} \quad \begin{pmatrix} i = 1,2,\cdots,m \\ j = 1,2,\cdots,n \end{pmatrix} \tag{2.4.7}$$

式中 a_i 为权重系数，有如下两种取法：

(1)令

$$a_i = (l_i + 1)/(s_i + 1) \tag{2.4.8}$$

其中 l_i 和 s_i 为第 i 个自变量所有样本在 m 个自变量序列中与因变量序列取最大距离 $\Delta\max x_{ij}$ 的个数和最小距离 $\Delta\min x_{ij}$ 的个数。当取定样本 $j=j_l, j=j_s$ 时，

$$\begin{aligned} \Delta\max x_{ij_l} &= \max|x_{0j_l} - x_{ij_l}| \\ \Delta\min x_{ij_s} &= \min|x_{0j_s} - x_{ij_s}| \end{aligned} \quad (i = 1,2,\cdots,m) \tag{2.4.9}$$

在某时间截口上各取出 $\Delta\max$ 和 $\Delta\min$，再计算二者的个数。权重 a_i 体现了第 i 个自变量在 m 个自变量中远近程度的相对关系。

(2)从数量角度定义权重系数

$$a_i = u_i/v_i \tag{2.4.10}$$

其中

$$\begin{aligned} u_i &= \frac{1}{l_i + 1}\sum_{l=1}^{l_i}\Delta\max x_{ij_l} \\ v_i &= \frac{1}{s_i + 1}\sum_{s=1}^{s_i}\Delta\min x_{ij_s} \end{aligned} \tag{2.4.11}$$

这里 l_i，s_i 与(2.4.8)式意义相同。但 $\Delta\max x_{ij_l}$ 和 $\Delta\min x_{ij_s}$ 的定义与(2.4.9)式不同，

$$\begin{aligned}\Delta\max x_{ij_l} &= \max(x_{0j_l} - x_{ij_l})^2 \\ \Delta\min x_{ij_s} &= \min(x_{0j_s} - x_{ij_s})^2\end{aligned} \tag{2.4.12}$$

关联系数表征的是各个序列在不同时刻的关联程度，关联度则是表征序列间关联程度大小的综合指标。一般简单地取关联系数的平均值作为关联度：

$$r_i = \frac{1}{n}\sum_{j=1}^{n}\xi_{ij} \quad (i = 1,2,\cdots,m) \tag{2.4.13}$$

应用实例[2.6]：计算 1958～1987 年二氧化碳(CO_2)浓度序列与全球、北半球、南半球、中国加权 160 站、中国未加权 160 站和西安 6 个单站年气温序列的关联度。这里 $n=20$，$m=6$。计算时，权重 a_i 采用第 2 种方法。计算结果见表 2.6。

表 2.6 CO_2 浓度与气温序列间的关联度

气温序列	全球	北半球	南半球	中国加权	中国未加权	西安
关联度	0.9239	0.8808	0.9517	0.8679	0.8601	0.8328

由表 2.6 可见，CO_2 浓度与南半球气温关系最密切，与西安单站气温的关系最差。

参考文献

[1]C. R. 劳著，张燮等译．线性统计推断及其应用．北京：科学出版社，1987.144～145

[2]陶澍．应用数理统计方法．北京：中国环境科学出版社，1994.308～313

[3]王梓坤．概率论基础及其应用．北京：科学出版社，1976.143～150

[4]邓聚龙．灰色系统——社会，经济．北京：国防工业出版社，1985.26～29

[5]曹鸿兴，江野．二氧化碳浓度增加与温度变化的关联分析．见：么枕生主编．气候学研究——气候与中国气候问题．北京：气象出版社，1993．148～154

第三章　基本气候状态的统计检验

第二章介绍了表征基本气候状态的几种常用统计量。我们通过某一气候变量序列的均值和方差了解其变化平均状况和变化幅度，但不清楚这种状况是否稳定、变化是否显著。因此，需要进行统计检验。自相关系数和 Pearson 相关系数仅仅显示气候变量前后时刻和两变量间的相关程度，但还不能由此贸然断言存在显著的相关，必须经过统计检验。本章首先简单介绍一下统计检验的概念和统计检验的流程。然后按照气候状态的稳定性、相关性和分布形态分别介绍进行统计检验的方法。由于我们分析的气候变量大多遵从正态分布，因此这里着重介绍与正态分布有关的统计检验的方法及其计算。

§3.1　统计检验概述

3.1.1　统计检验与统计假设

统计检验的基本思想是针对要检验的实际问题，提出统计假设。所谓统计假设实际上是用统计语言表达出期望得出结论的问题。例如：想了解北京和天津两地夏季降水量是否相同，可以将统计假设表达为：北京与天津两地夏季降水量均值没有差异。然后用特定的检验方法计算，并按给定的显著性水平对接受还是拒绝假设作出推断。需要强调的是，由于所有统计检验无一例外都针对总体而言，因此统计假设也必须与总体有关。例如：北京与天津两地降水量均值的比较，统计假设不能表述为两地降水量均值相同或不同，必须表述为两总体

均值相同或两样本来自均值相同的总体。

由上述可知，统计检验是对二者择一作出判断的方法。其统计假设包括相互对立的两方面，即原假设和对立假设。原假设是检验的直接对象，常用 H_0 表示。对立假设是检验结果拒绝原假设时必然接受的结论，用 H_1 表示。统计假设多数情况下可以用数学符号表达。例如原假设 $H_0:\mu_1=\mu_2$，就是检验两总体均值相等的统计假设。

由于选择显著性水平 α 的取值与是否拒绝原假设密切相关，为保证检验的客观性，应该在检验前就确定出适当的显著性水平。一般通取 0.05，有时也取 0.01。也就是说，在原假设正确的情况下，接受这一原假设的可能性有 95%或 99%，而拒绝这一假设的可能性较小。

3.1.2 统计检验的一般流程

进行统计检验的一般流程如下：

(1)明确要检验的问题，提出统计假设。

(2)确定显著性水平 α。

(3)针对研究的问题，选取一个适当的统计量。例如：检验两组样本均值差异可选用 t 检验，检验方差的显著性选用 F 检验等等。通常这些统计量的分布有表可查。

(4)根据观测样本计算有关统计量。

(5)对给定的 α，从表上查出与 α 水平相应的数值，即确定出临界值。

(6)比较统计量计算值与临界值，看其是否落入否定域中。若落入否定域则拒绝原假设。

关于统计检验的具体计算，在介绍检验方法时，根据实例再作详细说明。

§3.2 气候稳定性检验

某一地区的气候是否稳定，可以通过比较不同时段气候变量的均值或方差是否发生显著变化来判断。另外，比较两个地区的气候变化是否存在显著差异也可以通过检验均值和方差来判断。为叙述方便，对于适用上述两种情形的检验方法，在具体介绍时将一并加以说明。

3.2.1 u 检验

u 检验用于两方面的检验：

(1)总体均值的检验，可用于检验一地气候是否稳定。

(2)两个总体均值的检验，用于检验两地气候变化是否存在显著差异。

当方差 σ^2 已知，且比较稳定时，只需对均值进行检验。所谓均值检验就是检验样本均值 $\bar{x}$ 和总体均值无偏估计 μ_0 之间的差异是否显著。用 u 检验就可以进行这方面的检验。

构造统计量

$$u = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} \tag{3.2.1}$$

其中 $\bar{x}$ 为样本均值；μ_0 和 σ 为原总体均值和标准差；n 为样本量。如果假设总体均值无改变，即 $\mu=\mu_0$，则 $\bar{x}$ 遵从正态分布 $N(\mu_0, \frac{\sigma^2}{n})$，$u$ 遵从标准正态分布 $N(0,1)$。由正态分布表查得 u_{α_1}，u_{α_2}，使得

$$P(u \leqslant u_{\alpha_1}) + P(u \geqslant u_{\alpha_2}) = \alpha_1 + \alpha_2 = \alpha \tag{3.2.2}$$

由于正态分布的对称性，令 $\alpha_1=\alpha_2=\frac{\alpha}{2}$，于是有：

$$P(|u| \geqslant u_\alpha) = \alpha$$

在给定显著性水平 $\alpha=0.05$，查得 u_α 的值，若 $|u|\geqslant u_\alpha=1.96$ 时否定假设，若 $|u|<u_\alpha$ 则接受原假设。

下面以实例来说明这种检验方法。

应用实例[3.1]：经正态检验，中国 1910～1989 年年平均气温等级遵从正态分布，其均值为 2.94，标准差为 0.30。又观测得到 1990～1994 年中国年平均气温等级分别为 2.60，3.30，3.70，3.10 和 2.40，样本均值 $\bar{x}=3.02$。检验在显著性水平 $\alpha=0.05$ 下，中国年平均气温等级的总体均值与样本均值有无显著差异，即总体均值有无改变。这里样本量 $n=5$。检验步骤为：

(1)提出原假设 $H_0:\mu=\mu_0$。

用统计语言表述为：总体均值与样本均值之间没有显著差异。

(2)计算统计量，将特征值代入(3.2.1)有：

$$u=\frac{3.02-2.94}{0.30}\sqrt{5}\approx 0.604$$

(3)当 $\alpha=0.05$ 时，查正态分布函数表(附表 1b)，$u_\alpha=1.96$，那么，$u<u_\alpha$，接受原假设。至此，可以得出结论：在 $\alpha=0.05$ 显著性水平上，可以认为 1990～1994 年样本均值与年平均气温总体均值无显著差异，即年平均气温变化是稳定的。这里应强调，这一结论是在 $\alpha=0.05$ 的显著性水平上得出的，因为以更低的显著性水平进行检验，有可能得出不同的结论。

u 检验还可以用来检验两个总体的均值是否相等。比如：诊断两地气候状况是否有显著差异就可以用 u 检验。假设观测数据 x,y 分别遵从正态分布 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$，若要检验两个均值是否相等，即检验原假设 $H_0:\mu_1=\mu_2$

x 和 y 的样本量为 n_1 和 n_2，样本均值为 $\bar{x},\bar{y}$。它们均为正

态分布，且相互独立，因此 $\bar{x}-\bar{y}$ 也是正态分布，构造统计量

$$u=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \tag{3.2.3}$$

u 遵从标准正态分布 $N(0,1)$。

应用实例[3.2]：赤道东太平洋地区(0°～10°S，180°～90°W)春季(3～5 月)平均海表温度 39 年平均值为 27.5℃，方差为 2.07℃。西风漂流区(40°～20°N，180°～145°W)春季平均海表温度 39 年平均值为 17.3℃，方差为 2.08℃。检验两地区海温平均值有无显著差异。这里样本量 $n_1=n_2=39$。检验步骤为：

(1)提出原假设 H_0：$\mu_1=\mu_2$。用统计语言表述为：两总体均值之间没有显著差异。

(2)计算统计量，将特征量代入(3.2.3)，算得 $u=21.4$。

(3)当显著性水平 $\alpha=0.05$，$u_\alpha=1.96$。那么，$u>u_\alpha$，拒绝原假设。在 $\alpha=0.05$ 的显著性水平上，认为赤道东太平洋春季海温均值与西风漂流区春季海温的均值之间存在显著性差异。

归纳起来，u 检验适用于下列三种情况：

(1)方差是已知的。

(2)对遵从正态分布的观测对象样本量大或小均适用。

(3)若样本量足够大，即使观测对象不遵从正态分布也适用。因为样本量足够大时，可以认为其样本均值近似遵从正态分布。

3.2.2 t 检验

t 检验也是一种均值统计检验方法。它适用于下列两种情况：

(1)方差未知时。

(2)遵从正态分布的均值检验,小样本也适用。

和 u 检验一样,t 检验也构造了检验总体均值和两个总体均值两种统计量。在总体方差 σ^2 未知的情况下,是用样本方差 s^2 来估计的。

构造检验总体均值的 t 统计量

$$t=\frac{\bar{x}-\mu_0}{s}\sqrt{n} \tag{3.2.4}$$

其中 $\bar{x}$ 和 s 分别代表样本均值和标准差;μ_0 为总体均值;n 为样本量。在确定显著性水平 α 之后,根据自由度 $\nu=n-1$ 查 t 分布表(附表 2),若 $|t|\geqslant t_\alpha$,则拒绝原假设。

应用实例[3.3]:经正态检验赤道东太平洋地区(0°~10°S,180°~90°W)春季(3~5 月)平均海温遵从正态分布,1952~1981 年 30 年平均值 $\mu_0=27.4$℃。又观测得到,1982~1990 年 9 年春季海温分别为:27.5,28.7,27.3,27.1,27.3,28.3,27.6,26.8 和 27.4℃。9 年样本均值 $\bar{x}=27.6$℃,样本方差 $s^2=3.1$℃,标准差为 $s=1.76$℃,$n=9$。

(1)提出原假设 $H_0:\mu=\mu_0$。

(2)计算统计量:$t=\dfrac{27.4-27.6}{1.76}\sqrt{9}\approx-0.34$

(3)确定显著性水平 $\alpha=0.05$,这里自由度 $\nu=9-1=8$,查 t 分布表(附表 2)得 $t_\alpha=2.31$,因此 $|t|<t_\alpha$,接受原假设,认为赤道东太平洋海温的总体均值没有发生显著性变化,即这一时段赤道东太平洋海温是稳定的。

构造检验两个总体的均值有无显著差异的统计量:

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \tag{3.2.5}$$

显见(3.2.5)式遵从自由度 $\nu=n_1+n_2-2$ 的 t 分布。式中 $\bar{x}$，$\bar{y}$，n_1，n_2 的意义与(3.2.3)式相同。s_1^2 和 s_2^2 分别表示两个样本的方差。

如果样本量 n_1，n_2 均较大，可用下式近似计算

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \tag{3.2.6}$$

应用实例[3.4]：赤道东太平洋地区 1982～1990 年春季海温已在应用实例[3.3]中给出。西风漂流区(40°～20°N，180°～145°W)1982～1992 年 11 年春季海温(℃)分别为：17.0，16.1，17.4，17.7，16.8，16.2，16.9，17.5，17.1，17.1 和 16.7。在总体方差 σ^2 未知的情况下，检验来自两个总体的样本均值有无显著差异。赤道东太平洋地区春季海温的 9 年样本均值 $\bar{x}=27.6$℃，样本方差 $s_1^2=3.1$℃；西风漂流区春季海温的 11 年样本均值 $\bar{y}=17.0$℃，样本方差 $s_2^2=2.3$℃。

(1)提出原假设 H_0：$\mu_1=\mu_2$。

(2)计算统计量，将特征量代入(3.2.5)，$t\approx14.9$。

(3)确定显著性水平 $\alpha=0.05$，自由度 $\nu=9+11-2=18$，查 t 分布表 $t_\alpha=2.10$，由于 $t>t_\alpha$，拒绝原假设，认为在 $\alpha=0.05$ 显著性水平上，赤道东太平洋地区的海温均值与西风漂流区海温均值有显著性差异。这一结论与 u 检验用 39 年样本的结果一致。

3.2.3 χ^2 检验

上面讲到的 u 检验和 t 检验是对均值的统计检验。方差反映了某一变量观测数据的偏离程度，它是变量稳定与否的重要测度。因此，对方差的检验与均值检验一样重要。用 χ^2 检验就可以对正态总体方差有无显著改变进行检验。

若 s^2 是来自正态总体 $N(\mu,\sigma^2)$中的样本方差，则可以构造统计量：

$$\chi^2=\frac{(n-1)s^2}{\sigma^2} \tag{3.2.7}$$

可见，(3.2.7)式统计量适用于总体方差 σ^2 已知的情况，且仅限于对总体方差显著性的检验。确定显著性水平后，查 χ^2 分布表(附表 4)，查出自由度 $\nu=n-1$ 的上界 $\chi^2_{\frac{\alpha}{2}}$和下界 $\chi^2_{1-\frac{\alpha}{2}}$。若 $\chi^2>\chi^2_{\frac{\alpha}{2}}$或 $\chi^2<\chi^2_{1-\frac{\alpha}{2}}$，则认为总体方差有显著变化。

应用实例[3.5]：已知上海 10 月逐日地面相对湿度(单位：%)近似遵从正态分布，且 $\sigma^2=102.9$，又测得 5 天相对湿度，算出 $s^2=46.4$。

(1)提出原假设 $H_0:\sigma=\sigma_0$。可表述为总体方差与样本方差无显著差异。

(2)计算统计量。将特征量代入(3.2.7)，算出 $\chi^2\approx1.80$。

(3)确定显著性水平 $\alpha=0.10$，自由度 $\nu=5-1=4$，查 χ^2 分布表(附表 4)，$\chi^2_{\frac{\alpha}{2}}=9.49$，$\chi^2_{1-\frac{\alpha}{2}}=0.711$。$\chi^2=1.80<\chi^2_{\frac{\alpha}{2}}=9.49$，且 $\chi^2>\chi^2_{1-\frac{\alpha}{2}}=0.711$，所以接受原假设，认为总体方差与样本方差之间无显著差异。

统计量(3.2.7)是在正态总体均值 μ 未知时检验总体方差的。若总体均值 μ 已知时，可用下面统计量：

$$\chi^2=\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \tag{3.2.8}$$

进行检验。其中 n 为样本量，$x_i(i=1,2,\cdots,n)$为观测样本，(3.2.8)式遵从自由度 $\nu=n$ 的 χ^2 分布。

3.2.4 F 检验

检验两个总体的方差是否存在显著差异，可以用 F 检验。在总体方差未知的情况下，假定 s_1^2 和 s_2^2 是分别来自两个

相互独立的正态总体的样本方差，统计量

$$F=\left(\frac{n_1}{n_1-1}s_1^2\right)\Big/\left(\frac{n_2}{n_2-1}s_2^2\right) \tag{3.2.9}$$

遵从自由度 $\nu_1=n_1-1,\nu_2=n_2-1$ 的 F 分布。给定显著性水平 α 之后，查 F 分布表(附表 3a)，若 $F\geqslant F_{\frac{\alpha}{2}}$，则拒绝原假设。

应用实例[3.6]：对应用实例[3.4]的问题用 F 检验两个总体的样本方差有无显著差异。赤道东太平洋春季海温样本方差 $s_1^2=3.1℃$，西风漂流区春季海温样本方差 $s_2^2=2.3℃$。这里 $n_1=9,n_2=11$。

(1)原假设 $H_0:\sigma_1=\sigma_2$。

(2)计算统计量，算得 $F\approx1.38$。

(3)给定显著性水平 $\alpha=0.10$，自由度 $\nu_1=9-1=8,\nu_2=11-1=10$，查 F 分布表(附表 3b)，$F_{\frac{\alpha}{2}}=3.07$，$F<F_{\frac{\alpha}{2}}$，接受原假设，认为赤道东太平洋春季海温与西风漂流区春季海温的样本方差无显著差异。

顺便指出，F 检验还可用于方差分析中。将数据按不同时间间隔进行分组，然后利用 F 检验来检验不同组的组内方差与组间方差的显著性。此外，F 检验常被作为确定线性回归模型自变量入选和剔除的标准。利用 F 检验还可以判断 ARMA(自回归滑动平均)模型降阶后与原模型之间是否有显著性差异，以此确定模型的阶数。

§3.3 相关性检验

对于气候变量不同时刻间的线性相关(2.4.5)式和两气候变量间的线性相关(2.4.2)式是否显著，即相关数值达到多少算是存在显著相关关系，必须进行统计检验。

正态总体的相关检验实质上是两个变量间或不同时刻间观测数据的独立性检验。(2.4.2)和(2.4.5)式给出的相关系数 r 和 $r(j)$是总体相关系数 ρ 和 $\rho(j)$的渐近无偏估计。所谓相关检验，就是检验 ρ 和 $\rho(j)$为 0 的假设是否显著，即提出原假设 $H_0:\rho=0$ 或 $H_0:\rho(j)=0$。相关系数检验大致有以下三种方法。

(1)在假设总体相关系数 $\rho=0$ 成立的条件下，相关系数 r 的概率密度函数正好是 t 分布的密度函数，因此可以用 t 检验来对 r 进行显著性检验。统计量

$$t=\sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \tag{3.3.1}$$

遵从自由度 $\nu=n-2$ 的 t 分布。给定显著性水平 α，查 t 分布表，若 $t>t_\alpha$，则拒绝原假设，认为相关系数是显著的。

(2)当样本量足够大时，对于滞后长度为 j 的自相关系数的显著性检验，可以通过统计量

$$u(j)=\sqrt{n-j}\,r(j) \tag{3.3.2}$$

进行检验。(3.3.2)式遵从渐近 $N(0,1)$分布。通过对自相关系数的检验，可以判断气候变量是否具有持续性。

(3)为检验方便，已构造出不同自由度，不同显著性水平的相关系数检验表。在实际检验过程中，自由度已知，给定显著性水平，就可直接查表对相关系数进行显著性检验，这是实际研究工作中最通用的办法。

应用实例[3.7]：在应用实例[2.4]中，算得中国年平均气温与冬季平均气温之间的相关系数为 $r=0.48$。用第 1 种方法检验 r 是否显著，$n=20$。

(1)提出原假设　$H_0:\rho=0$。统计表述为总体相关系数为 0。

(2)计算统计量：

$$t = \sqrt{20-2}\,\frac{0.48}{\sqrt{1-0.48^2}} \approx 2.33$$

(3)给定显著性水平 $\alpha=0.05$，查自由度 $\nu=20-2=18$ 时 t 分布表，$t_\alpha=2.10$，由于 $t=2.33>t_\alpha$ 拒绝原假设，在 $\alpha=5\%$ 显著性水平上，认为年平均气温与冬季平均气温之间的相关系数是显著的。

应用实例[3.8]：应用实例[2.4]算出中国年平均气温与冬季平均气温之间相关系数 $r=0.48$。给定显著性水平 $\alpha=0.05$，用查相关系数表的办法对 r 进行检验。这里自由度 $\nu=20-2=18$，查相关系数表(附表 5)，自由度 18 对应 $\alpha=0.05$ 时，$r_\alpha=0.44$。由于 $r=0.48>r_\alpha$，因此认为，在 $\alpha=0.05$ 的显著性水平上，相关系数是显著的。若取显著性水平 $\alpha=0.01$ 查表 $r_\alpha=0.56$，$r<r_\alpha$，因此认为，在 $\alpha=0.01$ 的显著性水平上相关系数是不显著的。

由应用实例[3.8]可以看到，取不同的 α，得出的结论可能是不同的。这两个结论并不矛盾，因为它们是在不同显著性水平下作出的结论。因此，在诊断分析中，对希望作出否定原假设，即判断两变量间是否存在相关关系时，应该注意显著性水平的选取，α 取得小一些得出的结论可靠性就大一些。

在实际工作中，人们依赖相关系数提供的信息对未来作出预测。那么，相关系数是否具有稳定性是预报效果好坏的关键问题。因此，许多学者都在探索检验相关稳定性的方法[2]。若两个变量的统计特征不随时间改变，相关系数必然有良好的稳定性。根据这一特性，有学者提出用计算滑动相关系数和序贯检验的方法来检验相关是否稳定[3~4]。

§3.4 分布的统计检验

从上述内容中，可以看到正态分布在统计学中处在何等重要的位置。大多数气候诊断方法和预测模型是在气候变量呈正态分布假定的前提下进行的。因此，对于气候变量是否呈正态分布形态的检验是十分必要的。正态分布检验不仅可以判断原始变量是否遵从正态分布，还可以检验那些原本不遵从正态分布而经某种数学变换后的变量是否已成为正态分布形式。

3.4.1 正态分布偏度——峰度检验

对变量进行正态分布统计检验最简便的方法是对描述观测数据总体分布密度图形特征量偏度系数和峰度系数进行检验。

当样本量 n 足够大时，标准偏度系数(2.3.1)式和标准峰度系数(2.3.2)式都以标准正态分布 $N(0,1)$ 为渐近分布。因此，对某一变量作正态性检验，就是提出变量遵从正态分布的原假设，对计算出的样本标准偏度系数和峰度系数作检验。由于已有标准正态分布表，使得检验十分简便。确定出显著性水平 α，查表即可得出结论。需要注意的是，由于正态分布的对称性，查表时显著性水平应为 $\frac{\alpha}{2}$。例如：给定 $\alpha=0.05$，查表时，要找对应 $\frac{\alpha}{2}=0.025$ 的分布函数值。

这里顺便讲一下正态分布表的查法。附表 1 给出两种格式的正态分布表。附表 1a 是已知正态分布函数 u_α，求 α 值，附表 1b 是已知(给定)α，查 u_α 值。例如：已知分布函数 $u_\alpha=$

1.96，求对应的$\frac{\alpha}{2}$。其方法是，在附表 1a 中，从左栏找到 1.9，平行向右移，移到对应上栏为 0.06 处，相交点的值为 0.025 即为$\frac{\alpha}{2}$。再如：求$\frac{\alpha}{2}=0.05$时分布函数 u_α 值。其查表顺序是，在附表 1b 中，先从左栏找到 0.0，然后平行向右移，移至对应上栏为 5 处，相交点的值为 1.64485，即为 u_α 值。

应用实例[3.9]：应用实例[2.3]中计算出天津夏季(6～8月)降水的标准偏度系数 $g_1=0.96$，标准峰度系数 $g_2=-0.17$。检验天津夏季降水量是否遵从或近似遵从正态分布。

(1)提出原假设 H_0：天津夏季降水遵从正态分布。

(2)给定显著性水平 $\alpha=0.05$，查正态分布表(附表 1b)，查得 $u_\alpha=1.96$，由于 $g_1=0.96<1.96$，且 $|g_2|=0.17<1.96$，因此，接受原假设，认为在 $\alpha=0.05$ 显著性水平下，天津夏季降水近似遵从正态分布。

值得注意的是，对一个变量进行检验，只有偏度和峰度均接受原假设，才可以认为样本来自正态分布总体。

3.4.2 正态分布的 Liffifors 检验

正态分布的 Liffifors 检验通过对累积频率分布的比较，判断样本是否遵从正态分布。它也是提出变量遵从正态分布的原假设。具体检验步骤如下：

(1)对具有 n 个样本的观测数据 x_i，按从小到大顺序排列，并进行标准化处理

$$x_i=\frac{x_i-\bar{x}}{s}\quad(i=1,2,\cdots,n)\qquad(3.4.1)$$

(2)计算观测的累积频率：

$$f_i=\frac{i}{n}\quad(i=1,2,\cdots,n)\qquad(3.4.2)$$

(3)从附表6中查出理论的累积频率 $\hat{f}_i$。由于附表6所列的数值对应于标准正态分布函数曲线下方从0到自变量绝对值范围内的面积。当 x_i' 为正值时，累积频率 $\hat{f}_i$ 等于0.5加上查出的数值；当 x_i' 为负值时，$\hat{f}_i$ 等于0.5减去查出的数值。

(4)计算理论累积频率与观测累积频率之差的绝对值：

$$\begin{aligned} E_i &= |\hat{f}_i - f| \\ E_i' &= |\hat{f}_i - f_{i-1} - 1| \end{aligned} \quad (i = 1,2,\cdots,n) \quad (3.4.3)$$

在 E_i' 时，取 $f_0=0$。

(5)挑选所有 E_i 和 E_i' 中的最大值作为检验统计量：

$$E = \max(E_i, E_i') \quad (3.4.4)$$

(6)给定显著性水平 α，查 Lillifors 检验临界值表(附表7)，若 $E<E_\alpha$，则接受检验原假设，认为在 α 显著性水平下变量遵从正态分布。

3.4.3 数据正态化变换

对于不遵从正态分布的变量可以作适当的变换，使其正态化。这里给出几种常用的变换公式。

(1)对数变换：对数变换是一种很常用的正态化变换方法。它的优点是计算简便。对原始数据 x_i 取对数

$$x_i' = \ln x_i \quad (i = 1,2,\cdots,n) \quad (3.4.5)$$

(2)平方根变换：对离散型变量用平方根变换十分奏效，即

$$x_i' = \sqrt{x_i + 0.5} \quad (i = 1,2,\cdots,n) \quad (3.4.6)$$

(3)角变换：对于遵从二项分布的变量，可采用角变换

$$x_i' = \arcsin\sqrt{x_i} \quad (i = 1,2,\cdots,n) \quad (3.4.7)$$

(4)幂变换：对于不清楚分布形式的变量，使用幂变换是

最合适的。其中有 Box-Cox 幂变换、Hinkley 幂变换和 Box-Tidwell 幂变换等。由于选取最佳幂次涉及到优化问题，计算较繁杂，这里不详细介绍。

参考文献

[1]M. 费史著，王福保译．概率论及数理统计．上海：上海科学出版社，1962

[2]朱盛明．相关系数稳定性分析及其应用．气象学报，1982，49(4)

[3]林学椿．统计天气预报中相关系数的不稳定性．大气科学，1978，3(2)：55～63

[4]丁裕国．气象变量间相关系数的序贯检验及其应用．南京气象学院学报，1987，(23)：340～347

第四章　气候变化趋势分析

随时间变化的一列气候数据构成了一个气候时间序列。我们研究的变量常常是离散观测得到的随机序列，例如：年降水总量序列、月海表温度序列、季平均气温序列等等均属于这类时间序列。气候时间序列一般具有以下特征：

(1)数据的取值随时间变化。

(2)每一时刻取值的随机性。

(3)前后时刻数据之间存在相关性和持续性。

(4)序列整体上有上升或下降趋势，并呈现周期振荡。

(5)在某一时刻的数据取值出现转折或突变。

前 2 种特征是时间序列的一般规律，其分析方法已有许多书籍介绍[1~2]。本章重点介绍气候序列变化趋势的诊断方法。在下面的第五和第六章中将陆续介绍气候序列突变现象、周期性及时频结构等特征的诊断方法。

对任何一个气候时间序列 x_t 都可以看成由下列几个分量构成：

$$x_t = H_t + P_t + C_t + S_t + a_t$$

其中 H_t 为气候趋势分量，是指几十年的时间尺度显示出的气候变量上升下降趋势，它是一种相对序列长度的气候波动；P_t 为气候序列存在的一种固有的周期性变化，例如：年、月变化；C_t 为循环变化分量，代表气候序列周期长度不严格的隐含周期性波动，例如：几年、十几年十或几十年长度的波动；S_t 是平稳时间序列分量；a_t 是随机扰动项，又称白噪声。

分离气候变化趋势的常用做法是用年总量、年平均值或

月、季总量来构造气候时间序列，这样就消除了固有周期分量 P_i。然后再作统计处理，消除或削弱循环变化分量 C_i 和随机扰动项 a_i。这就可以将趋势分量 H_i 显现出来。S_i 则可以由平稳随机序列分析方法处理。下面介绍几种常用的分离气候趋势的统计方法。

§4.1 线性倾向估计

4.1.1 方法概述

用 x_i 表示样本量为 n 的某一气候变量，用 t_i 表示 x_i 所对应的时间，建立 x_i 与 t_i 之间的一元线性回归

$$\hat{x}_i = a + bt_i \quad (i = 1,2,\cdots,n) \tag{4.1.1}$$

(4.1.1)式可以看作一种特殊的、最简单的线性回归形式。它的含意是用一条合理的直线表示 x 与其时间 t 之间的关系。由于(4.1.1)式右边的变量是 x 对应的时间 t，而不是其它变量，因此这一方法属于时间序列分析范畴。(4.1.1)式中 a 为回归常数，b 为回归系数。a 和 b 可以用最小二乘进行估计。

对观测数据 x_i 及相应的时间 t_i，回归系数 b 和常数 a 的最小二乘估计为

$$\begin{cases} b = \dfrac{\sum\limits_{i=1}^{n} x_i t_i - \dfrac{1}{n}\left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} t_i\right)}{\sum\limits_{i=1}^{n} t_i^2 - \dfrac{1}{n}\left(\sum\limits_{i=1}^{n} t_i\right)^2} \\ a = \bar{x} - b\bar{t} \end{cases} \tag{4.1.2}$$

其中

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

利用回归系数 b 与相关系数之间的关系，求出时间 t_i 与变量 x_i 之间的相关系数

$$r=\sqrt{\frac{\sum_{i=1}^{n}t_i^2-\frac{1}{n}(\sum_{i=1}^{n}t_i)^2}{\sum_{i=1}^{n}x_i^2-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2}} \qquad (4.1.3)$$

4.1.2 计算步骤

(1)对变量 x_i 构造其对应的时间序列 t_i。t_i 可以是年份，例如：1951，1952，…，1990 年，也可以是序号，例如：1，2，…，30 或其它时间单位值。

(2)按照(4.1.2)和(4.1.3)式求出回归系数 b，回归常数 a 及相关系数 r。

(3)将 a，b 代入(4.1.1)式，求出回归计算值 $\hat{x}_i$。

4.1.3 计算结果分析

对于线性回归计算结果，主要分析回归系数 b 和相关系数 r。

4.1.3.1 回归系数 b——倾向值

回归系数 b 的符号表示气候变量 x 的趋势倾向。b 的符号为正，即当 $b>0$ 时，说明随时间 t 的增加 x 呈上升趋势；当 b 的符号为负，即 $b<0$ 时，说明随时间 t 的增加，x 呈下降趋势。b 值的大小反映了上升或下降的速率，即表示上升或下降的倾向程度。因此，通常将 b 称为倾向值，将这种方法叫做线性倾向估计。

4.1.3.2 相关系数 r

相关系数 r 表示变量 x 与时间 t 之间线性相关的密切程度。当 $r=0$ 时，回归系数 b 为 0，即用最小二乘法估计确定的回归直线平行于 x 轴，说明 x 的变化与时间 t 无关；当 $r>0$

时，$b>0$，说明 x 随时间 t 的增加呈上升趋势；当 $r<0$ 时，$b<0$，说明 x 随时间 t 增加呈下降趋势。$|r|$ 越接近 0，x 与 t 之间的线性相关就越小。反之，$|r|$ 越大，x 与 t 之间的线性相关就越密切。当然，要判断变化趋势的程度是否显著，就要对相关系数进行显著性检验。确定显著性水平 α，若 $|r|>r_\alpha$，表明 x 随时间 t 的变化趋势是显著的，否则表明变化趋势是不显著的。

应用实例[4.1]：用线性倾向估计分析华北地区 1951～1995 年夏季(6～8 月)干旱指数的变化趋势[3]。这一实例包括两方面内容：

(1)分析代表华北整个区域干旱状况干旱指数的变化趋势。这里 $n=45$，x_i 为干旱指数，t_i 为 x_i 一一对应的年份。利用(4.1.2)式计算出 $a=40.7602$，$b=-0.0182$，用(4.1.3)式求出相关系数 $r=-0.3395$，将 a，b 代入(4.1.1)式，求出回归计算值 $\hat{x}_i$。绘制出线性趋势图(图 4.1)，图中曲线为干旱指

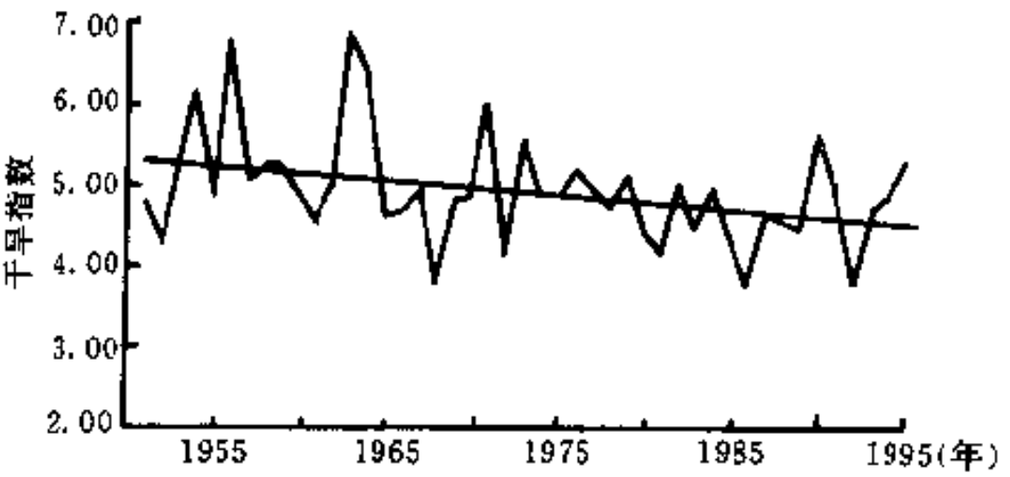


图 4.1　华北地区夏季干旱指数线性变化趋势

数，直线为方程(4.1.1)式配制的回归直线。计算结果表明，从总体上考察，该地区夏季干旱指数呈下降趋势，相关系数 $|r|>r_{0.05}=0.2875$，表明这种下降趋势在 $\alpha=0.05$ 显著性水平

上是显著的。由图 4.1 可以看出，回归直线向下倾斜比较明显。

(2)分别计算华北地区 24 个站(承德、张家口、北京、天津、石家庄、德州、邢台、安阳、烟台、青岛、潍坊、济南、临沂、荷泽、连云港、淮阴、徐州、阜阳、郑州、南阳、信阳、长治、太原、临汾)夏季干旱指数的线性变化趋势，绘制出线性趋势分布图(图 4.2)。图中虚线表示下降趋势，实线表示上升趋势。可以看出，除张家口一站外，华北其余各站均呈下降趋势，其中烟台、长治等站下降趋势明显，相关系数超过 0.05 显著性水平。

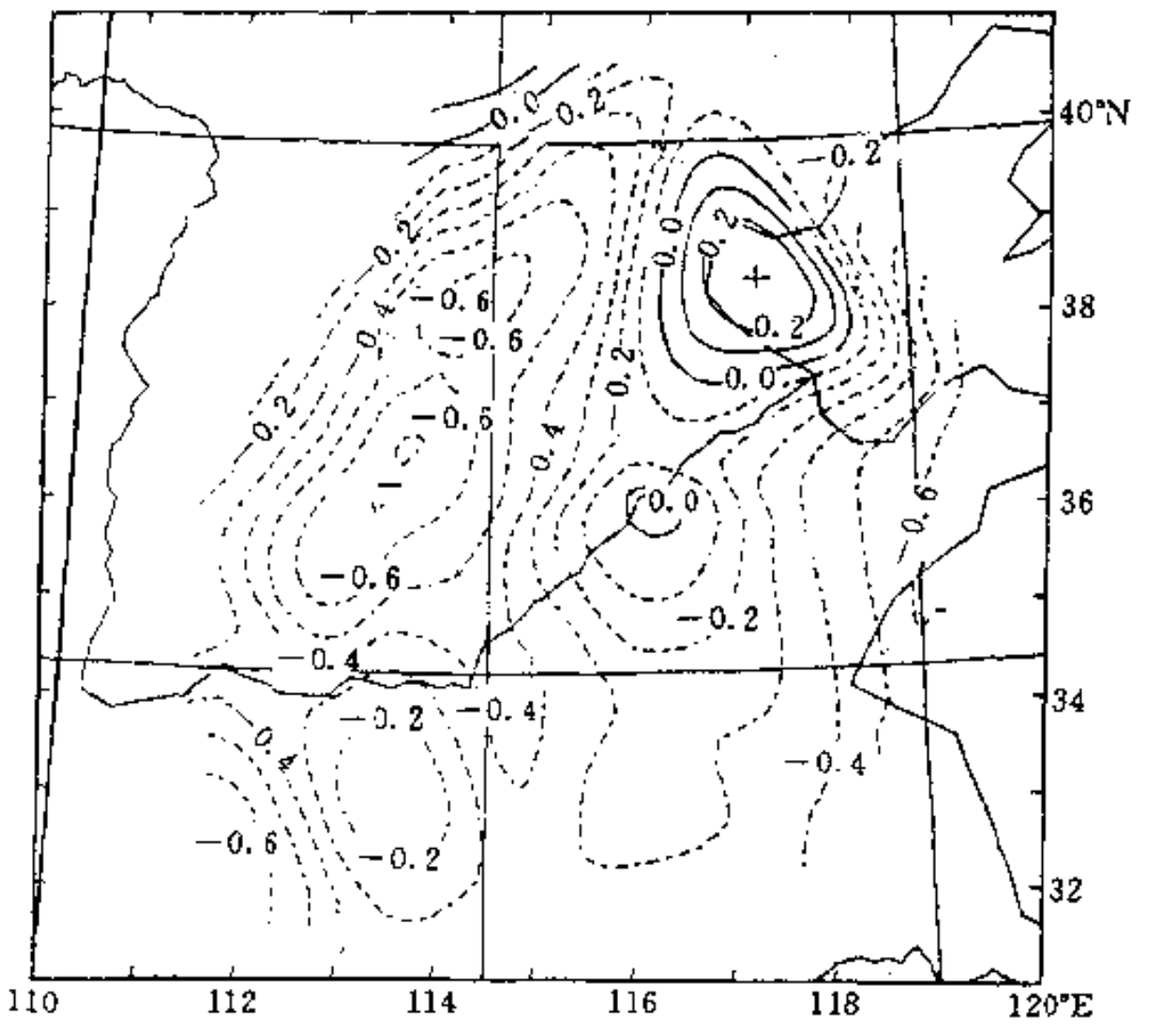


图 4.2　华北地区干旱指数线性趋势分布

由此可以得出这样的结论：近 45 年来，无论从区域整体

还是从各个站的角度分析，华北地区夏季干旱指数均呈下降趋势，即趋于干旱且趋势比较明显。

应用实例[4.2]：利用线性倾向估计研究全球冰雪变化趋势[4]。资料取自1973～1992年美国NOAA卫星资料中的北极冰(NI)、南极冰(SI)覆盖范围和北半球雪盖(NS)。计算结果列于表4.1。可以看到，除夏季南极冰呈正倾向外，其余均呈负倾向。表明近20年来，北极冰和北半球雪盖面积都在收缩，其中北半球夏季雪盖的收缩最显著，相关系数超过0.01($r_{0.01}=0.561$)显著性水平，且收缩速度较快，为$-14.34\times10^5 km^2\cdot(100年)^{-1}$。其次是全年的北半球雪盖的收缩亦较显著，相关系数亦超过0.01显著性水平，其余均未达到一定的显著性水平，虽然也趋于收缩，但趋势不明显。

表4.1 极冰和雪盖的变化趋势

项　目	时　段	倾向值b	相关系数α
北极冰(NI)	12～2月冬	−0.015	−0.026
	6～8月夏	−0.032	−0.053
	年	−0.051	−0.123
南极冰(SI)	12～2月冬	0.244	0.234
	6～8月夏	−0.386	−0.291
	年	−0.131	−0.131
北半球雪盖(NS)	12～2月冬	−0.006	−0.020
	6～8月夏	−0.143	−0.633
	年	−0.103	−0.585

§4.2 滑动平均

4.2.1 方法概述

滑动平均是趋势拟合技术最基础的方法，它相当于低通

滤波器。用确定时间序列的平滑值来显示变化趋势。对样本量为 n 的序列 x，其滑动平均序列表示为：

$$\hat{x}_j = \frac{1}{k}\sum_{i=1}^{k} x_{i+j-1} \quad (j = 1,2,\cdots,n-k+1) \quad (4.2.1)$$

式中 k 为滑动长度。作为一种规则，k 最好取奇数，以使平均值可以加到时间序列中项的时间坐标上。若 k 取偶数，可以对滑动平均后的新序列取每两项的平均值，以使滑动平均对准中间排列。

可以证明，经过滑动平均后，序列中短于滑动长度的周期大大削弱，显现出变化趋势。

4.2.2 计算步骤

根据具体问题的要求及样本量大小确定滑动长度 k，用(4.2.1)式直接对观测数据进行滑动平均值计算。n 个数据可以得到 $n-k+1$ 个平滑值。编程计算时可采用这样的形式：首先将序列的前 k 个数据求和得到一值，然后依次用这个值减去平均时段的第一个数据，并加上第 $k+1$ 个数据，再用求出的值除以 k，循环这样的过程计算出 $1,2,\cdots,n-k+1$ 个平滑值。

4.2.3 计算结果分析

分析时主要从滑动平均序列曲线图来诊断其变化趋势。例如：看其演变趋势有几次明显的波动，是呈上升还是呈下降趋势。

应用实例[4.3]：北京 1951～1996 年夏季(6～8 月)降水量见表 2.1。计算 11 年滑动平均。样本量 $n=46$，滑动平均后得到 $46-11+1=36$ 个平滑值。图 4.3 中较光滑曲线即为滑动平均曲线。可以看出，50 年代中期至 60 年代末北京夏季降水量呈逐渐下降趋势。70 年代初降至低点后变化平缓，处于

少雨阶段，并持续至今，虽有小的波动，但没有出现明显的上升或下降趋势。

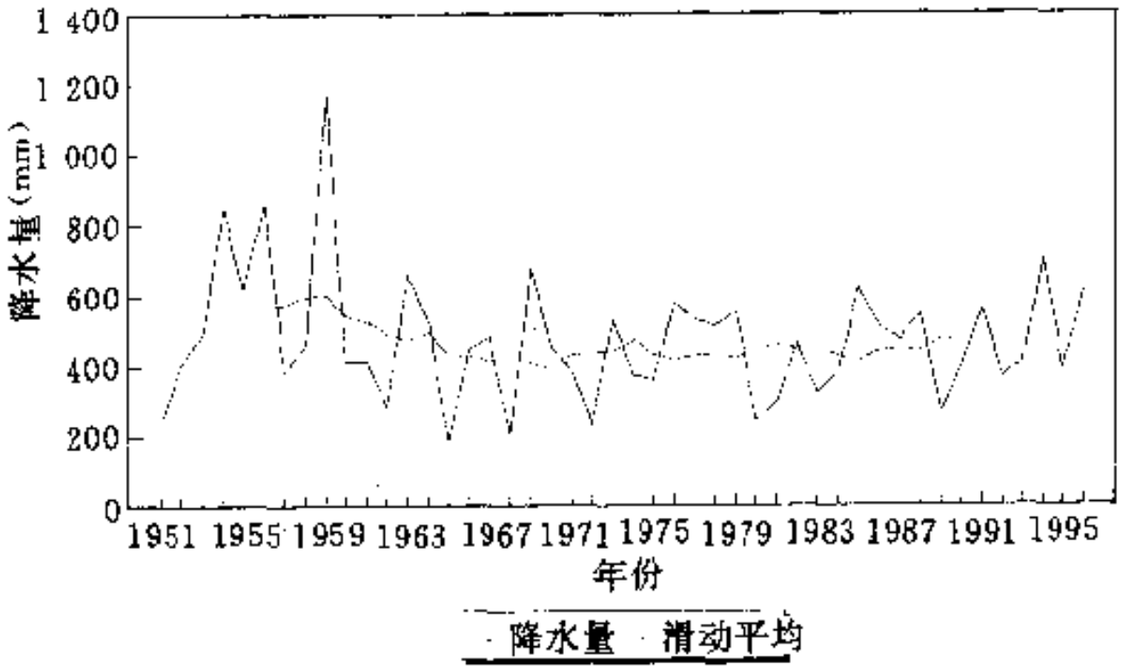


图 4.3　北京 1951～1996 年夏季降水变化趋势
（其中较光滑曲线为 11 年滑动平均）

§4.3　累积距平

4.3.1　方法概述

累积距平也是一种常用的、由曲线直观判断变化趋势的方法。对于序列 x，其某一时刻 t 的累积距平表示为：

$$\hat{x}_t = \sum_{i=1}^{t}(x_i - \bar{x}) \quad (t = 1,2,\cdots,n) \qquad (4.3.1)$$

其中

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

将 n 个时刻的累积距平值全部算出，即可绘出累积距平曲线进行趋势分析。

4.3.2　计算步骤

(1)计算出序列 x 的均值 $\bar{x}$。

(2)按(4.3.1)式逐一算出各个时刻的累积距平值。

4.3.3　计算结果分析

累积距平曲线呈上升趋势，表示距平值增加，呈下降趋势则表示距平值减小。从曲线明显的上下起伏，可以判断其长期显著的演变趋势及持续性变化，甚至还可以诊断出发生突变的大致时间。从曲线小的波动变化可以考察其短期的距平值变化。

应用实例[4.4]：分别计算 1958～1994 年全球二氧化碳含量和全球气温序列的累积距平值。这里 $n=37$。图 4.4a 和图 4.4b 为它们的累积距平曲线图。尽管二氧化碳的累积距平均为负值(图 4.4a)，但曲线的变化形态却十分直观、清晰地展示出近 37 年来全球二氧化碳经历了一次显著的波动。从 37 年的平均值来看，50 年代末至 70 年代中期二氧化碳呈下降趋势，70 年代末 80 年代初开始增长，上升趋势至今未减。全球气温的累积距平曲线(图 4.4b)的变化趋势与二氧化碳有着十分一致的配合。从 70 年代末 80 年代初开始增温，直至 1994 年。

§4.4　五、七和九点二次平滑

4.4.1　方法概述

对时间序列 x 作五点二次、七点二次和九点二次平滑，与滑动平均作用一样，亦是起到低通滤波器作用，以展示出变化趋势，它可以克服滑动平均削弱过多波幅的缺点。对于时间序列 x，用二次多项式拟合：

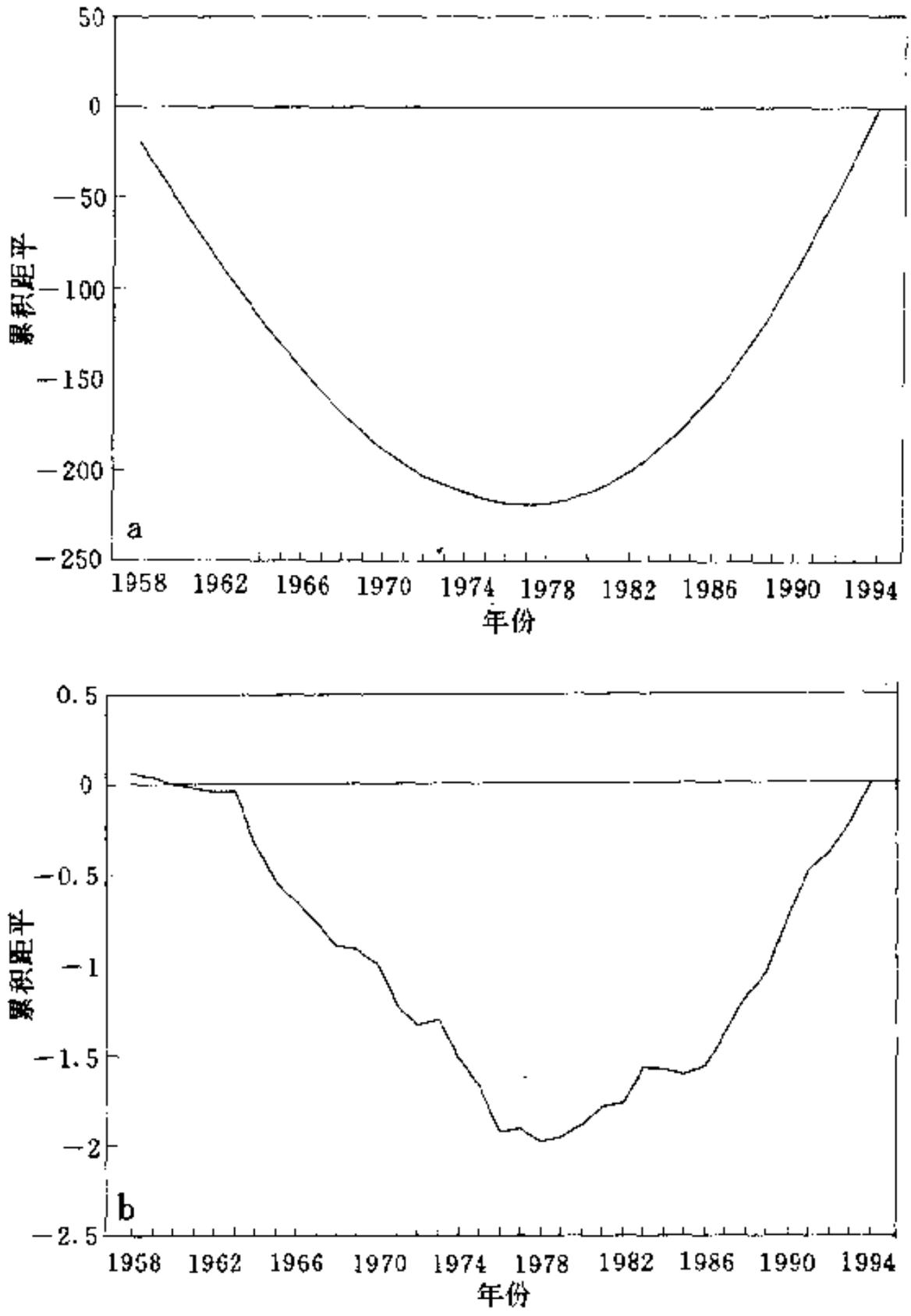


图 4.4　1958～1994 年全球二氧化碳(a)和全球气温(b)累积距平曲线

$$\hat{x} = a_0 + a_1 x + a_2 x^2 \qquad (4.4.1)$$

根据最小二乘法原理确定系数 a_0, a_1, a_2，可以分别得到五点二次、七点二次和九点二次平滑公式：

$$\hat{x}_{i-2} = \frac{1}{35}(-3x_{i-2} + 13x_{i-1} + 17x_i + 12x_{i+1} - 3x_{i-2}) \qquad (4.4.2\,a)$$

$$\hat{x}_{i-3} = \frac{1}{21}(-2x_{i-3} + 3x_{i-2} + 6x_{i-1} + 7x_i + 6x_{i+1} - 3x_{i+2} - 2x_{i+3}) \qquad (4.4.2\,b)$$

$$\hat{x}_{i-4} = \frac{1}{231}(-21x_{i-4} + 14x_{i-3} + 39x_{i-2} + 54x_{i-1} + 59x_i + 54x_{i+1} + 39x_{i+2} + 14x_{i+3} - 21x_{i+4}) \qquad (4.4.2\,c)$$

4.4.2 计算步骤

(1)根据实际问题的需要及样本量的大小确定平滑的点数 k，然后按照(4.4.2)式直接对观测数据进行平滑计算，得到 $n-k+1$ 个平滑值。

(2)对五、七及九点端点的平滑值，分别由相邻的二、三和四点平滑值求平均得到。这样就可以得 n 个平滑值。

在编程计算时，分别设计计算五点、七点和九点二次平滑的子程序，每个子程序含有计算 $n-k+1$ 个平滑值及端点平滑值过程。在主程序中用条件语句控制执行指定平滑点数的子程序。

应用实例[4.5]：对北京 1951～1996 年夏季降水量(表 2.1)进行九点二次平滑。样本量 $n=46$，平滑后仍得到 46 个平滑值。九点二次平滑曲线(图 4.5)显然不像 11 年滑动平均曲线(图 4.3)那么光滑了，除显现出 60 年代末至 70 年代初降水量的下降趋势外，还保留了几次明显的波动。如在相对少

雨阶段的70～90年代，曾经历了两次几年周期的振动。可见，滑动长度选取的不同，得到的变化趋势会有差别。因此，根据分析的目的和对象选取恰当的平滑时段是较重要的。

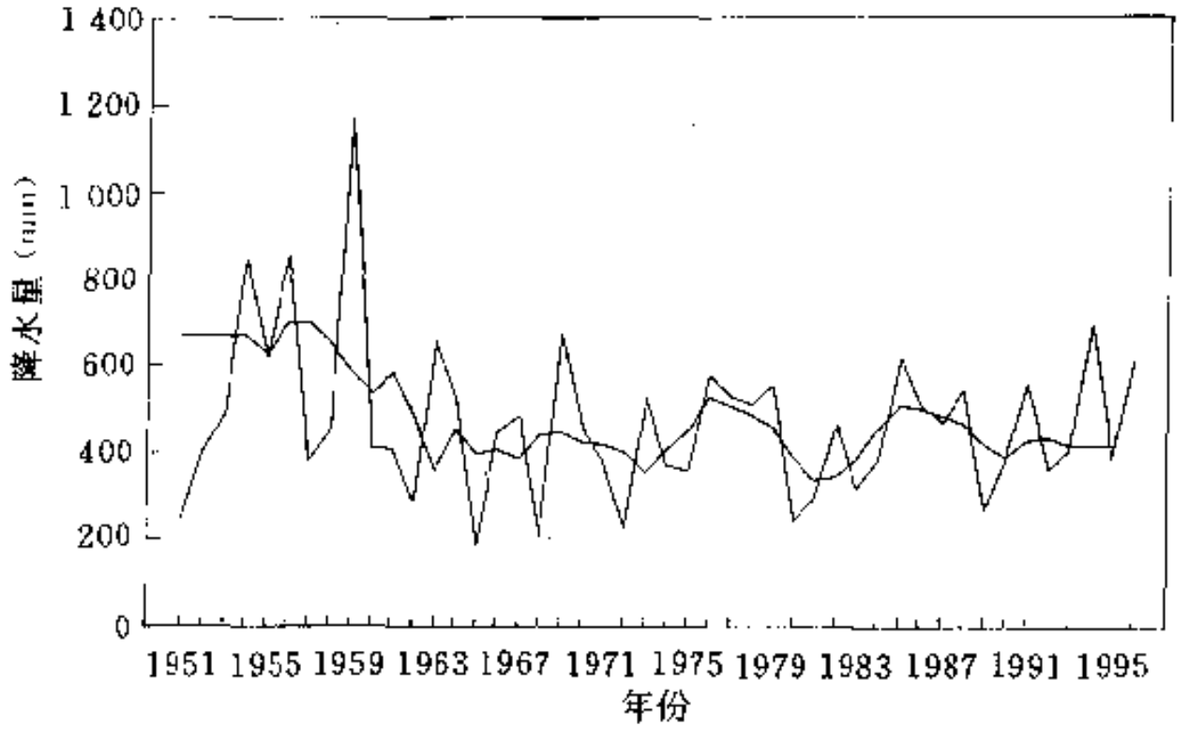


图 4.5　北京 1951～1996 年夏季降水变化趋势
（其中较光滑曲线为九点二次平滑值）

§4.5　五点三次平滑

4.5.1　方法概述

五点三次平滑与上述的二次平滑一样，是一种常用的多项式平滑方法。它可以很好地反映序列变化的实际趋势，特别适合于作相对短时期变化趋势的分析。

对序列 x，在其每个数据点前后各取两相邻数据，用三次多项式拟合

$$\hat{x} = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \qquad (4.5.1)$$

根据最小二乘法原理确定系数 a_0, a_1, a_2 和 a_3，可得到五点三

次平滑公式：

$$\hat{x}_{i-2}=\frac{1}{70}(69x_{i-2}+4x_{i-1}-6x_i+4x_{i+1}-x_{i-2}) \tag{4.5.2 a}$$

$$\hat{x}_{i-1}=\frac{1}{35}(2x_{i-2}+27x_{i-1}+12x_i-8x_{i+1}-x_{i-2}) \tag{4.5.2 b}$$

$$\hat{x}_i=\frac{1}{35}(-3x_{i-2}+12x_{i-1}+17x_i+12x_{i+1}-3x_{i+2}) \tag{4.5.2 c}$$

$$\hat{x}_{i+1}=\frac{1}{35}(2x_{i-2}-8x_{i-1}+12x_i+27x_{i+1}+2x_{i+2}) \tag{4.5.2 d}$$

$$\hat{x}_{i+2}=\frac{1}{70}(-x_{i-2}+4x_{i-1}-6x_i+4x_{i+1}+69x_{i+2}) \tag{4.5.2 e}$$

由上述公式可见，这一方法要求样本量 $n\geqslant 5$。

4.5.2　计算步骤

对序列的开始两点用(4.5.2 a)、(4.5.2 b)式平滑，最后两点用(4.5.2 d)和(4.5.2 e)式进行平滑，其余各点均按(4.5.2 c)式进行平滑。

§4.6　三次样条函数

4.6.1　方法概述

三次样条函数是近二十几年来统计界十分瞩目的数据拟合方法。它以对给定的时间序列进行分段曲线拟合的方式，来反映其本身真实的变化趋势。

对样本量为 n 的序列 x_i，其对应的时刻为 t_i。欲将 $t_1,t_2,\cdots,t_n$ 分成 m 段，需在 t_i 中插入 $m-1$ 个分点，即有：

$$t_1 < \eta_1 < \eta_2 < \cdots < \eta_{m-1} < t_n \tag{4.6.1}$$

为方便起见，两端各引入一个新分点 η_0, η_m，并令 $\eta_0 < t_1, t_n \leqslant \eta_m$。这样，就可以在每个新分点上构造拟合函数：

$$F(t) = \begin{cases} \hat{x}_1(t) & \eta_0 < t \leqslant \eta_1 \\ \hat{x}_2(t) & \eta_1 < t \leqslant \eta_2 \\ \vdots & \vdots \\ \hat{x}_m(t) & \eta_{m-1} < t \leqslant \eta_m \end{cases} \tag{4.6.2}$$

其中

$$\hat{x}_k(t) = \sum_{j=0}^{3} V_{kj} a_{kj}(s) \quad (k = 1, 2, \cdots, m) \tag{4.6.3}$$

$$s = (2t - \eta_{k-1} - \eta_1)/(\eta_k - \eta_{k-1}) = 2(t - \eta_{k-1})/(\eta_k - \eta_{k-1}) - 1 \tag{4.6.4}$$

(4.6.3)式中 $a_{kj}(s)$ 是切比雪夫第一类多项式：

$$\begin{cases} a_{k0}(s) = 1 \\ a_{k1}(s) = s \\ a_{k2}(s) = 2s^2 - 1 \\ a_{k3}(s) = 4s^3 - 3s \end{cases} \tag{4.6.5}$$

$\hat{x}_k(t)$ 在 $m-1$ 个分点上相邻的两个多项式满足函数 $x_k(t)$ 及其二阶导数在 η_k 处均连续，分段多项式 $F(t)$ 即为三次样本函数。用最小二乘法原理确定出 V_{kj}，就可以得到分段拟合曲线。

假设第 k 个区间的 t_k 共有 q 个，即

$$\eta_{k-1} < t_{k1} \leqslant t_{k2} \leqslant \cdots \leqslant t_{kq} \leqslant \eta_k \quad (k = 1, 2, \cdots, m) \tag{4.6.6}$$

要确定 V_{kj}，使得

$$Q_0 = \sum_{k=1}^{m} \sum_{l=1}^{q} [x_{kl} - \hat{x}_k(t_{kl})]^2 \tag{4.6.7}$$

达到最小，并满足函数 $\hat{x}_k(t)$ 及其二阶导数在各分点处都连续

的约束条件。

应用拉格朗日乘子法，上述条件极值问题化为无条件极值问题，即确定 V_{kj}，使得

$$Q=\sum_{k=1}^{m}\sum_{l=1}^{q}(x_{kl}-\bar{x}_k(t_{kl}))^2+\sum_{k=1}^{m-1}\sum_{s=0}^{2}\lambda_{ks}\left[x_k^{(s)}(t_k')-x_{k+1}^{(s)}(t_k)\right] \tag{4.6.8}$$

达到最小，其中 $\lambda_{ks}(k=1,2,\cdots,m-1,s=0,1,2)$ 为拉格朗日乘子，也是需要定出的。

这时，V_{kj} 和 λ_{ks} 满足

$$\begin{cases}\dfrac{\partial Q}{\partial V_{kj}}=0 \quad (k=1,2,\cdots,m;j=0,1,2,3)\\ \dfrac{\partial Q}{\partial \lambda_{ks}}=0 \quad (k=1,2,\cdots,m-1;s=0,1,2)\end{cases} \tag{4.6.9}$$

方程组(4.6.9)是关于 V_{kj} 和 λ_{ks} 的线性方程组。经过推导，(4.6.9)式可以表示为下列矩阵形式：

$$\begin{cases}H_kV_k+\dfrac{1}{2}(C_k^{\mathrm{T}}\lambda_k+D_{k-1}^{\mathrm{T}}\lambda_{k-1})=bk \quad (k=1,2,\cdots,m)\\ C_kV_k+D_kV_{k+1}=0 \quad (k=1,2,\cdots,m-1)\end{cases} \tag{4.6.10}$$

其中

$$H_k=A_kA_k^{\mathrm{T}}$$

$$A_k=\begin{bmatrix}a_{k0}(s_{k1}) & a_{k1}(s_{k1}) & a_{k2}(s_{k1}) & a_{k3}(s_{k1})\\ a_{k0}(s_{k2}) & a_{k1}(s_{k2}) & a_{k2}(s_{k2}) & a_{k3}(s_{k2})\\ \vdots & \vdots & \vdots & \vdots\\ a_{k0}(s_{kq}) & a_{k1}(s_{kq}) & a_{k2}(s_{kq}) & a_{k3}(s_{kq})\end{bmatrix}$$

符号“T”代表矩阵转置。

$$V_k^{\mathrm{T}} = (V_{k0}, V_{k1}, \cdots, V_{km})$$

$$C_k = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & h_k & 4h_k & 9h_k \\ 0 & 0 & 4h_k^2 & 24h_k^2 \end{bmatrix}$$

$$C_k^{\mathrm{T}} = \begin{bmatrix} 0.5 & 0 & 0 \\ 0.5 & 0.5h_k & 0 \\ 0.5 & 2h_k & 2h_k^2 \\ 0.5 & 4.5h_k & 12h_k^2 \end{bmatrix}$$

$$D_{k-1} = \begin{bmatrix} -1 & 1 & -1 & 1 \\ 0 & -h_k & 4h_k & -9h_k \\ 0 & 0 & -4h_k^2 & 24h_k \end{bmatrix}$$

$$D_{k-1}^{\mathrm{T}} = \begin{bmatrix} -0.5 & 0 & 0 \\ 0.5 & -0.5h_k & 0 \\ -0.5 & 2h_k & -2h_k^2 \\ 0.5 & -4.5h_k & 12h_k^2 \end{bmatrix}$$

$$h_k^s = [2/(\eta_k - \eta_{k-1})]^s \quad (s=0,1,2)$$

$$\lambda_k^{\mathrm{T}} = [\lambda_{k0}, \lambda_{k1}, \lambda_{k2}]$$

$$b_k = A_s x_k = \begin{bmatrix} \sum_{l=1}^{q} a_{k0}(s_{kl}) x_{kl} \\ \sum_{l=1}^{q} a_{k1}(s_{kl}) x_{kl} \\ \sum_{l=1}^{q} a_{k2}(s_{kl}) x_{kl} \\ \sum_{l=1}^{q} a_{k3}(s_{kl}) x_{kl} \end{bmatrix}$$

这时，(4.6.10)式变成一般带型线性方程组：

$$FV = B \tag{4.6.11}$$

其中

$$V = (V_{12}, V_{12}, V_{13}, V_{14}, \lambda_{11}, \lambda_{12}, \lambda_{13}, \cdots, \lambda_{m1}, \lambda_{m2}, \lambda_{m3}, V_{m1}, V_{m2}, V_{m3}, V_{m4})^T$$

$$B = (b_{11}, b_{12}, b_{13}, b_{14}, 0, 0, 0, \cdots, 0, 0, 0, b_{m1}, b_{m2}, b_{m3}, b_{m4})^T$$

这时我们就可以采用标准求解方法来解(4.6.11)方程组。

4.6.2 计算步骤

三次样条函数拟合的计算步骤如下：

(1)根据原序列的长度及实际问题的需要，将序列划分为 m 段，即给定新分点 $\eta_0, \eta_1, \cdots, \eta_m$ 和每段对应的点数。实际计算时，每段点数只包括右端点不含左端点。

(2)用形如(4.6.2)式的样条函数，对各分段作最小二乘拟合。其中(4.6.3)式的 a_{kj} 由(4.6.4)和(4.6.5)式给出。V_{kj} 由(4.6.10)式算出。

(3)将分段拟合曲线连接起来，即可得序列 x 光滑的拟合曲线。虽然每段上的多项式可能各不相同，但却在相邻段的连接处是光滑的。

应用实例[4.6]：对近百年(1884～1988年)西太平洋热带气旋年频数用三次样条函数进行拟合，分析其变化趋势[5]。在105年年频数序列中插入5个分点分成6段，两端各引进一个新分点 η_0 和 η_6，这样分点 $\eta_0, \eta_1, \cdots, \eta_6$ 定为 0.5，10.5，20.5，30，40，60，105，每段点数为 10，10，10，10，20，45。经过样条函数拟合，得到一条光滑的变化趋势曲线(见图 4.6)。

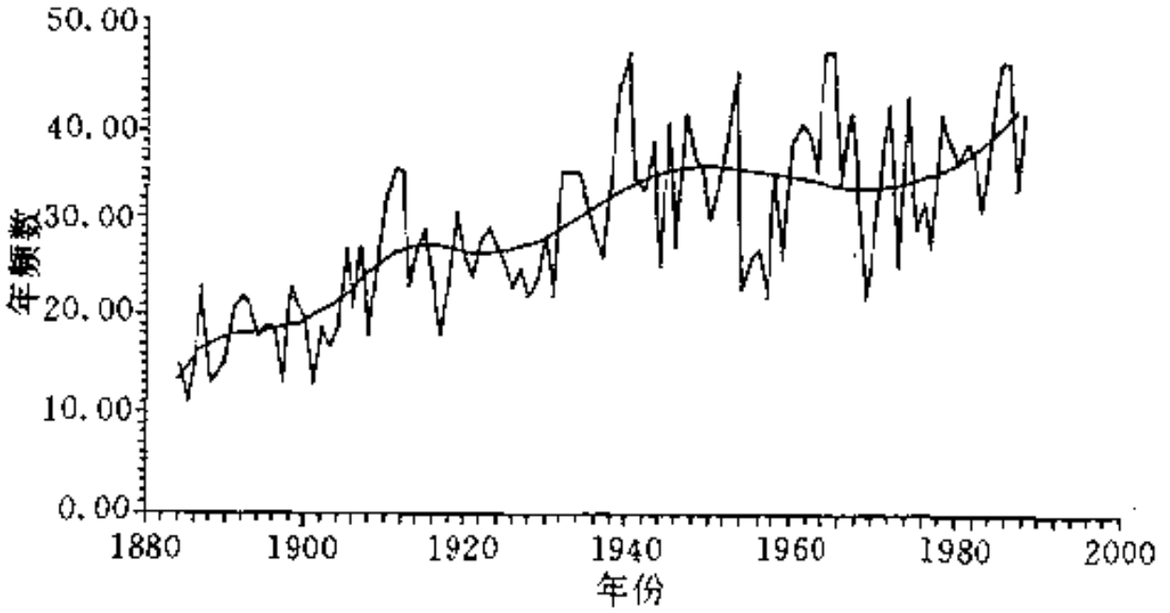


图 4.6　1884～1988 年西太平洋热带气旋年频数变化
（图中光滑曲线为三次样本函数拟合）

§4.7　变化趋势的显著性检验

前面讲到的用线性倾向估计方法考察气候序列的变化趋势，其变化趋势是否显著可以通过对相关系数的显著性检验进行判断。但是，滑动平均、累积距平、多项式等方法是根据变化趋势曲线图直观判断的，对趋势十分明显的容易得出结论，而有时则很难直观得到结论。这时可以借助统计检验的办法。这里给出一种非参数统计检验方法[6]。

4.7.1　方法概述

对于气候序列 x_i，在 i 时刻，$i=1,2,\cdots,n-1$，有

$$r_i=\begin{cases}1 & \text{当 } x_j>x_i\\ 0 & \text{否则}\end{cases}\qquad (j=i+1,\cdots,n)\tag{4.7.1}$$

可见，r_i 是 i 时刻以后的数值 x_j，$j=i+1,\cdots,n$ 大于该时刻值 x_i 的样本个数。

计算统计量：

$$Z=\frac{4\sum_{i=1}^{n-1}r_i}{n(n-1)}-1 \qquad (4.7.2)$$

显见，对于递增直线，r_i 序列为 $n-1, n-2, \cdots, 1$，这时 $Z=1$，对于递减直线 $Z=-1$，Z 值在 1～−1 之间变化。

给定显著性水平 α，假定 $\alpha=0.05$，则判据

$$Z_{0.05}=1.96\left[\frac{4n+10}{9n(n-1)}\right]^{1/2}. \qquad (4.7.3)$$

若 $|Z|>Z_{0.05}$，则认为变化趋势在 $\alpha=0.05$ 显著性水平下是显著的。

4.7.2 计算步骤

(1)对原气候序列或用某种方法得到的趋势序列，计算其秩统计量 r_i。

(2)计算统计量 Z。

(3)计算判据 $Z_{0.05}$，若 $|Z|>Z_{0.05}$ 则判断变化趋势是显著的。

应用实例[4.7]：北京 1951～1996 年夏季降水量见表 2.1。用累积距平进行变化趋势分析。用上述非参数统计量对变化趋势作显著性检验。由(4.7.1)式得到 r_i 序列(表 4.2)。$Z=-0.4531$，$Z_{0.05}=0.20$，$|Z|>Z_{0.05}$，因此认为在 $\alpha=0.05$ 显著性水平下，夏季降水量的变化趋势是显著的。

表 4.2 北京夏季降水量累积距平序列的 r_i

42	44	43	29	25	12	13	14	0	0
1	2	1	0	0	1	0	3	0	0
0	3	1	3	5	3	2	1	0	0
0	1	5	8	3	1	1	0	2	4
2	3	3	1	1					

参 考 文 献

[1]杨位钦，顾岚．时间序列分析与动态数据建模．北京：北京工业学院出版社，1986

[2]项静恬等．动态和静态数据处理——时间序列和数理统计分析．北京：气象出版社，1991

[3]魏凤英．华北干旱不同时间尺度的变化特征．见：王馥棠等主编．华北农业干旱研究进展．北京：气象出版社，1997．1～10

[4]曹鸿兴等．全球冰雪变化趋势及其对全球增暖的蕴示．见：曹鸿兴等主编．我国短期气候变化及成因研究．北京：气象出版社，1996．84～88

[5]张光智等．近百年西北太平洋热带气旋年频数的变化特征．热带气象学报，1995．11(4)：315～323

[6]黄嘉佑．气候状态变化趋势与突变分析．气象，1995，21(7)：54～57

第五章　气候突变检测

所有变量的变化方式不外乎两种基本形式，一种是连续性变化，另一种是不连续的飞跃。不连续变化现象的特点是突发性，所以人们称不连续现象为“突变”。对突变有不同的理解和定义。其实，突变可以理解为一种质变，一种当量变达到一定的限度时发生的质变。形形色色的突变现象向传统的分析方法提出了挑战。20 世纪 60 年代中期，法国数学家 Thom 创立了突变理论，很快突变理论风靡一时，经过十几年从理论到实际应用方面的改进与完善，使其在科学界造成很大影响。随后，突变理论在数学、生物、天文、地震、气象、社会科学等领域得到广泛应用。

突变理论是以常微分方程为数学基础的[1]，其精髓是关于奇点的理论，其要点在于考察某种系统或过程从一种稳定状态到另一种稳定状态的飞跃。从统计学的角度，可以把突变现象定义为从一个统计特性到另一个统计特性的急剧变化，即从考察统计特征值的变化来定义突变。例如：考察均值、方差状态的急剧变化。目前，突变统计分析还相当不成熟，针对常见的突变问题，人们藉助统计检验、最小二乘法、概率论等发展了一些行之有效的检验方法。主要涉及检验均值和方差有无突然漂移、回归系数有无突然改变及事件的概率有无突然变化等方面。这里仅介绍几种在检测气候突变现象中最常用的方法。

顺便指出，突变理论研究中最为活跃，同时争议最大的就是有关应用问题。对一些物理机制目前还不甚明确的突变现

象，人们很难给予解释，有时使用的检测方法不当，可能会得出错误的结论。因此，建议在确定某气候系统或过程发生突变现象时，最好使用多种方法进行比较。另外，要指定严格的显著性水平进行检验。运用气候专业知识对突变现象进行判断也十分重要。

§5.1 滑动 t-检验

5.1.1 方法概述

滑动 t-检验是考察两组样本平均值的差异是否显著来检验突变。其基本思想是把一气候序列中两段子序列均值有无显著差异看为来自两个总体均值有无显著差异的问题来检验。如果两段子序列的均值差异超过了一定的显著性水平，可以认为均值发生了质变，有突变发生。

对于具有 n 个样本量的时间序列 x，人为设置某一时刻为基准点，基准点前后两段子序列 x_1 和 x_2 的样本分别为 n_1 和 n_2，两段子序列平均值为 $\bar{x}_1$ 和 $\bar{x}_2$，方差为 s_1^2 和 s_2^2。定义统计量：

$$t=\frac{\bar{x}_1-\bar{x}_2}{s\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{5.1.1}$$

其中

$$s=\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}}$$

(5.1.1)式遵从自由度 $\nu=n_1+n_2-2$ 的 t 分布。

这一方法的缺点是子序列时段的选择带有人为性。为避免任意选择子序列长度造成突变点的漂移，具体使用这一方法时，可以反复变动子序列长度进行试验比较，提高计算结果

的可靠性。

5.1.2 计算步骤

(1)确定基准点前后两子序列的长度，一般取相同长度，即 $n_1=n_2$。

(2)采取滑动的办法连续设置基准点，依次按(5.1.1)式计算统计量。由于进行滑动的连续计算，可得到统计量序列 $t_i, i=1,2,\cdots,n-(n_1+n_2)+1$。

(3)给定显著性水平 α，查 t 分布表(附表 2)得到临界值 t_α，若 $|t_i|<t_\alpha$，则认为基准点前后的两子序列均值无显著差异，否则认为在基准点时刻出现了突变。

在编程计算时，滑动计算两子序列平均值 $\bar{x}_1$ 和 $\bar{x}_2$，相当于执行两子序列的滑动平均过程。设子序列长度 $n_1=n_2=IH$，以前 IH 个数据之和为基数，依次减前一个数往后加一个数求平均，这是第一个序列的滑动平均过程。第二个滑动平均是以第 $IH+1$ 个至 $2\times IH$ 个数据之和为基数，再依次减前一个数往后加一个数求平均。再用滑动的方式依次计算两子序列各自的方差。

5.1.3 计算结果分析

根据 t 统计量曲线上的点是否超过 t_α 值来判断序列是否出现过突变，如果出现过突变，确定出大致的时间。

另外，根据诊断出的突变点分析突变前后序列的变化趋势。

应用实例[5.1]：用滑动 t-检验检测 1911～1995 年中国年平均气温等级序列的突变。这里 $n=85$，两子序列长度 $n_1=n_2=10$。给定显著性水平 $\alpha=0.01$，按 t 分布自由度 $\nu=n_1+n_2-2=18$，$t_{0.01}=\pm 2.898$。这里为编程方便，给定 $t_{0.01}=\pm 3.20$，实际上给出了更严格的显著性水平。将计算出的 t 统计

量序列和 $t_{0.01}=\pm 3.20$ 绘出图 5.1。从图中看出，自 1920 年以来，t 统计量有两处超过 0.01 显著性水平，一处是正值(出现在 1920 年)，另一处是负值(出现在 1950 年)。说明中国年平均气温在近 85 年来，出现过两次明显的突变。20 年代经历了一次由冷到暖的转变，50 年代出现了一次由增暖转为冷的明显突变，尽管 70 年代末 80 年代初，中国气温与全球气温同步在回升，但没有达到显著性水平。

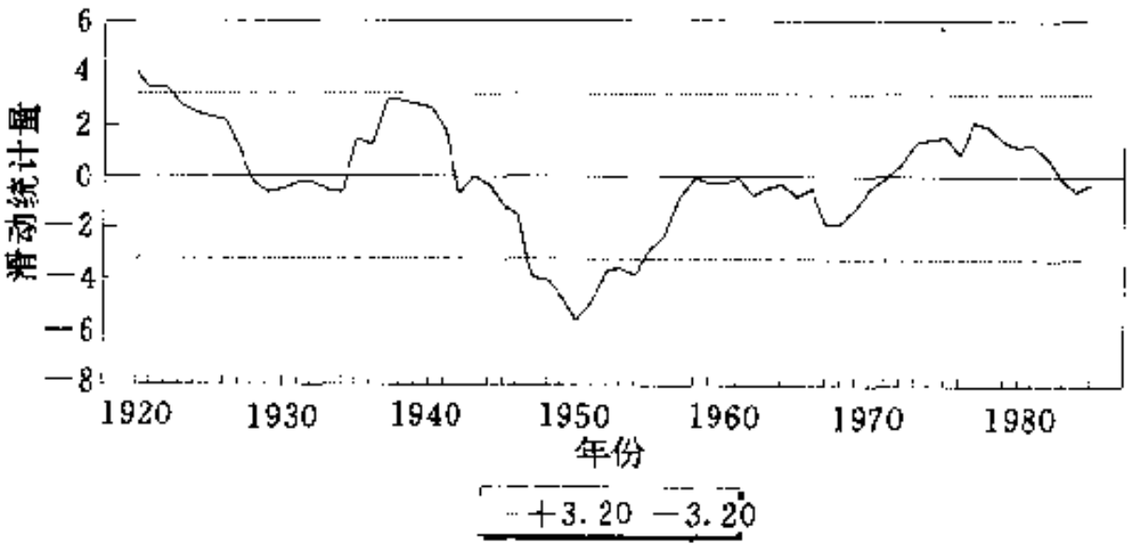


图 5.1　中国气温等级滑动 t-统计量曲线

(直线为 $\alpha=0.01$ 显著性水平临界值)

§5.2　Cramer's 法

5.2.1　方法概述

Cramer 方法的原理与 t-检验类似，区别仅在于它是用比较一个子序列与总序列的平均值的显著差异来检测突变。

设总序列 x 和子序列 x_1 的均值分别为 $\bar{x}$ 和 $\bar{x}_1$，总序列方差为 s，定义统计量：

$$t=\sqrt{\frac{n_1(n-2)}{n-n_1(1+\tau)}}\cdot\tau \qquad (5.2.1)$$

其中 n 为序列样本长度；n_1 为子序列样本长度。τ 由下式求出

$$\tau=\frac{\bar{x}_1-\bar{x}}{s}$$

(5.2.1)式遵从自由度 $n-2$ 的 t 分布。

由于这一方法也要人为地确定子序列长度，因此在具体使用时，应采取反复变动子序列长度的办法来提高计算结果的可靠性。

5.2.2 计算步骤

(1)确定子序列 x_1 的长度 n_1。

(2)按(5.2.1)式以滑动的方式计算 t 统计量，可得到 t 统计量序列 t_i，$i=1,2,\cdots,n-n_1+1$。

(3)给定显著性水平 α，查 t 分布表得到临界值 t_α，若 $|t_i|<t_\alpha$，则认为子序列均值与总体序列均值之间无显著差异，否则认为在 t_i 对应的时刻发生了突变。

§5.3 Yamamoto 法

Yamamoto 方法是从气候信息与气候噪声两部分来讨论突变问题的。由于是由 Yamamoto 最先将信噪比用于确定日本地面气温、降水、日照时数等序列的突变，故称其为 Yamamoto 法[2]。

5.3.1 方法概述

对于时间序列 x，人为设置某一时刻为基准点，基准点前后样本量分别为 n_1 和 n_2 的两段子序列 x_1 和 x_2 的均值为 $\bar{x}_1$ 和 $\bar{x}_2$，标准差为 s_1 和 s_2，定义信噪比为：

$$SNR = \frac{|\overline{x}_1 - \overline{x}_2|}{s_1 + s_2} \qquad (5.3.1)$$

(5.3.1)式的含意是，两段子序列的均值差的绝对值为气候变化的信号，而它们的变率（用标准差 s_1 和 s_2 表示）则视为噪声。信噪比还有一些不同的定义，但与其类似。

在 t-检验中，我们曾假定两段子序列样本相同，即 $n_1 = n_2 = III$。那么比较(5.1.1)和(5.3.1)两式，可以得到：

$$t > SNR\ \sqrt{IH}$$

若取 $IH = 10$，$SNR = 1.0$，相当于 $|t| > 3.162$，$t_\alpha = t_{0.01} = 2.878$，即 $|t| > t_\alpha$，超过 $\alpha = 0.01$ 显著性水平，说明两段子序列的均值存在显著性差异，认为在基准点发生了突变。显然，$SNR > 2.0$，相当于 $t > 6.324$，超过 $\alpha = 0.0001$ 显著性水平，表明在基准点出现了强的突变。

由(5.3.1)式可见，Yamamoto 方法也是用检验两子序列均值的差异是否显著来判别突变的。从形式上它比 t-检验更简单明了。但它也存在与 t-检验相同的缺点，由于人为设置基准点，子序列长度的不同可能引起突变点的漂移。因此，应该通过反复变动子序列的长度进行试验比较，以便得到可靠的判别。

5.3.2　计算步骤

(1)确定基准点前后两段子序列长度，一般取 $n_1 = n_2 = IH$。

(2)连续设置基准点，以滑动方式依次按(5.3.1)式计算信噪比，得到信噪比序列 SNR_i，$i = 1, 2, \cdots, n - 2 \times IH - 1$。

(3)若 $SNR_i > 1.0$，则认为在 i 时刻有突变发生。若 $SNR_i > 2.0$，则认为在 i 时刻有强突变发生。

5.3.3　计算结果分析

根据信噪比曲线上的点是否超过 1.0 或 2.0 直线判断序列是否发生过突变或强突变，并确定出发生突变的时间。同时，根据信噪比曲线的变化，分析序列的演变趋势，特别是长期演变趋势。

应用实例[5.2]：用 Yamamoto 法研究中国、北半球和全球气温序列的突变[3]。具体计算时，子序列的长度分别取为 30，25，20，15，10，5 年。

子序列长度为 30，25，20，15，10 年的结果指出，中国气温序列在 1948～1953 年间信噪比值均超过 $\alpha=0.01$ 显著性水平，且最大者大都出现在 1949～1950 年间。北半球和全球气温序列的信噪比变化大体一致。在子序列长度取 30，25，20 和 15 年时，1925～1926 年间信噪比出现了最大值，且超过 0.01 显著性水平。这两个序列的子序列长度取 10 年时，除 20 年代中期有一突变点外，北半球序列在 1893～1894 年间、全球序列在 1895～1896 年间的信噪比也超过了 0.01 显著性水平。在子序列长度取 5 年时，中国气温序列在 60 年代中期又有一次超过显著性水平的突变点出现。北半球和全球也分别在 60 和 70 年代间各有一次信噪比超过显著性检验。

由计算结果可见，取长短不同的平均时段作出的突变事实是有差异的。但是，揭示的显著突变基本上是一致的，且信噪比最大值出现的年份也基本相同。可以确定，在 1949～1950 年间中国气温曾出现过一次显著的突变，由 40 年代明显暖转为冷，这种变冷是近百年来最为显著的一次突变。上世纪末、本世纪 20 年代，北半球和全球经历过两次突变。可见，突变事实的揭露有助于我们了解气候系统的行为，同时也为建立气候预测模型提供了必要的根据。

§5.4 Mann-Kendall 法

Mann-Kendall 法是一种非参数统计检验方法。在第三章统计检验中介绍的检验方法都是参数方法，即假定了随机变量的分布。非参数检验方法亦称无分布检验，其优点是不需要样本遵从一定的分布，也不受少数异常值的干扰，更适用于类型变量和顺序变量，计算也比较简便。由于最初由 Mann 和 Kendall 提出了原理并发展了这一方法，故称其为 Mann-Kendall 法。但是，当时这一方法仅用于检测序列的变化趋势。后来经其他人进一步完善和改进，才形成目前的计算格式。

5.4.1 方法概述

对于具有 n 个样本量的时间序列 x，构造一秩序列：

$$s_k = \sum_{i=1}^{k} r_i \qquad (k = 2,3,\cdots,n) \tag{5.4.1}$$

其中

$$r_i = \begin{cases} +1 & \text{当 } x_i > x_j \\ 0 & \text{否则} \end{cases} \qquad (j = 1,2,\cdots,i)$$

可见，秩序列 s_k 是第 i 时刻数值大于 j 时刻数值个数的累计数。

在时间序列随机独立的假定下，定义统计量：

$$UF_k = \frac{[s_k - \mathrm{E}(s_k)]}{\sqrt{\mathrm{Var}(s_k)}} \qquad (k = 1,2,\cdots,n) \tag{5.4.2}$$

其中 $UF_1 = 0$，$\mathrm{E}(s_k)$，$\mathrm{Var}(s_k)$ 是累计数 s_k 的均值和方差，在 $x_1, x_2, \cdots, x_n$ 相互独立，且有相同连续分布时，它们可由下式算出：

$$E(s_k)=\frac{n(n+1)}{4}$$
$$Var(s_k)=\frac{n(n-1)(2n+5)}{72} \tag{5.4.3}$$

UF_i 为标准正态分布，它是按时间序列 x 顺序 $x_1,x_2,\cdots,x_n$ 计算出的统计量序列，给定显著性水平 α，查正态分布表(附表 1b)，若 $UF_i|>U_\alpha$，则表明序列存在明显的趋势变化。

按时间序列 x 逆序 $x_n,x_{n-1},\cdots,x_1$，再重复上述过程，同时使 $UB_k=-UF_k,k=n,n-1,\cdots,1,UB_1=0$。

这一方法的优点在于不仅计算简便，而且可以明确突变开始的时间，并指出突变区域。因此，是一种常用的突变检测方法。

5.4.2 计算步骤

(1)计算顺序时间序列的秩序列 s_k，并按(5.4.2)式计算 UF_k。

(2)计算逆序时间序列的秩序列 s_k，也按(5.4.2)式计算出 UB_k。

(3)给定显著性水平，例如：$\alpha=0.05$，那么临界值 $u_{0.05}=\pm1.96$。将 UF_k 和 UB_k 两个统计量序列曲线和 ±1.96 两条直线均绘在一张图上。

5.4.3 计算结果分析

分析绘出的 UF_k 和 UB_k 曲线图。若 UF_k 或 UB_k 的值大于 0，则表明序列呈上升趋势，小于 0 则表明呈下降趋势。当它们超过临界直线时，表明上升或下降趋势显著。超过临界线的范围确定为出现突变的时间区域。如果 UF_k 和 UB_k 两条曲线出现交点，且交点在临界线之间，那么交点对应的时刻便是突变开始的时间。

应用实例[5.3]：用 Mann-Kendall 法检测 1900～1990 年上海年平均气温序列(表 5.1)的突变。给定显著性水平 $\alpha=0.05$，即 $u_{0.05}=\pm 1.96$。计算结果绘成图 5.2。

表 5.1　1900～1990 年上海年平均气温序列

1900～1909	15.4	14.6	15.8	14.8	15.0	15.1	15.1	15.0	15.2	15.4
1910～1919	14.8	15.0	15.1	14.7	16.0	15.7	15.4	14.5	15.1	15.3
1920～1929	15.5	15.1	15.6	15.1	15.1	14.9	15.5	15.3	15.3	15.4
1930～1939	15.7	15.2	15.5	15.5	15.6	16.1	15.1	16.0	15.0	15.8
1940～1949	16.2	16.2	16.0	15.6	15.9	15.2	16.7	15.8	16.2	15.9
1950～1959	15.8	15.5	15.9	16.8	15.5	15.8	15.0	14.9	15.3	16.0
1960～1969	16.1	16.5	15.5	15.6	16.1	15.6	16.0	15.4	15.5	15.2
1970～1979	15.4	15.6	15.1	15.8	15.5	16.0	15.2	15.8	16.2	16.2
1980～1989	15.2	15.7	16.0	16.0	15.7	15.9	15.7	16.7	15.3	16.1
1990	16.2									

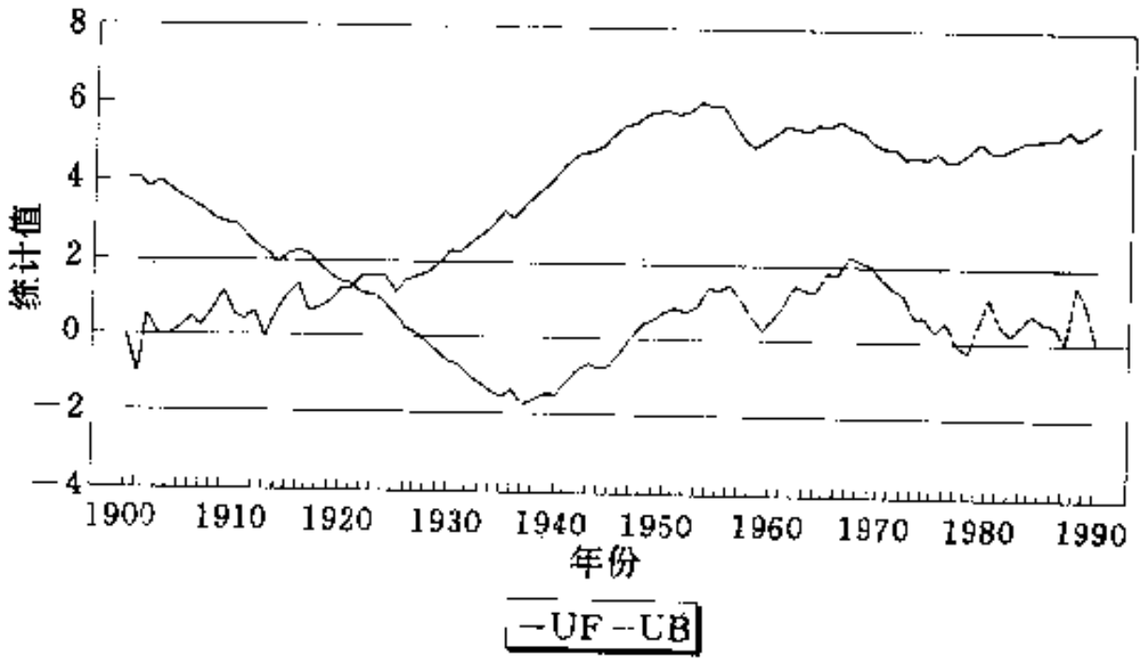


图 5.2　上海年平均气温 Mann-Kendall 统计量曲线

(直线为 $\alpha=0.05$ 显著性水平临界值)

由 UF 曲线可见，自本世纪 20 年代以来，上海年平均气温有一明显的增暖趋势。30～90 年代这种增暖趋势均大大超

过 0.05 临界线甚至超过 0.001 显著性水平($u_{0.001}=2.56$),表明上海气温的上升趋势是十分显著的。根据 UF 和 UB 曲线交点的位置,确定上海年平均气温 20 年代的增暖是一突变现象,具体是从 1921 年开始的。

上海年平均气温的增暖趋势及发生突变的时间均与北半球年平均气温完全一致。作为突变的典型实例,许多文献都给出了北半球平均气温的 Mann-Kendall 统计量曲线图[4]。与图 5.2 比较,我们发现两张图的变化趋势与突变点完全吻合。可见,上海的气温变化与北半球是同步的。

§5.5 Pettitt 方法

Pettitt 方法是一种与 Mann-Kendall 法相似的非参数检验方法。由于是由 Pettitt, A. N 最先用于检测突变点的,故将其称为 Pettitt 法[5]。

与 Mann-Kendall 法一样,构造形如(5.4.1)式的一秩序列。不同的是 r_i 是分三种情况定义的,即

$$r_i=\begin{cases}+1 & \text{当 } x_i>x_j\\ 0 & \text{当 } x_i=x_j\\ -1 & \text{当 } x_i<x_j\end{cases}\qquad (j=1,2,\cdots,i)\quad(5.5.1)$$

可见,这里的秩序列 s_k 是第 i 时刻数值大于或小于 j 时刻数值个数的累计数。

Pettitt 法是直接利用秩序列来检测突变点的。若 t_0 时刻满足

$$k_{t_0}=\mathrm{Max}|s_k|\quad(k=2,3,\cdots,n)\qquad(5.5.2)$$

则 t_0 点处为突变点。

计算统计量

$$P = 2\exp[-6k_{t0}^2(n^3 + n^2)] \qquad (5.5.3)$$

若 $P \leqslant 0.5$，则认为检测出的突变点在统计意义上是显著的。

§5.6 Lepage 法

Lepage 法是一种无分布双样本的非参数检验方法。它的统计量是由标准的 Wilcoxon 检验和 Ansariy-Bradley 检验之和构成的。由于将两个检验联合在一起的原理最早是由 Lepage 在 1971 年提出的，因此将其称为 Lepage 检验[6]。已有研究证明，与其它检验相比，它是一种十分有效的检验方法。但是，迄今为止，它还没有像上述介绍的方法那样广泛地应用到气候研究领域。

5.6.1 方法概述

Lepage 检验原本是用于检验两个独立总体有无显著差异的非参数统计检验方法。用它来检测序列的突变，其基本思想简而言之即是，视序列中的两个子序列为两个独立总体，经过统计检验，如果两子序列有显著差异，则认为在划分子序列的基准点时刻出现了突变。

假设基准点之前的子序列样本量为 n_1，之后的子序列样本量为 n_2，n_{12}为 n_1 和 n_2 之和。在 n_{12}范围内计算秩序列 s_i

$$s_i = \begin{cases} 1 & \text{如果最小值出现在基准点之前} \\ 0 & \text{如果最小值出现在基准点之后} \end{cases} \qquad (5.6.1)$$

构造一秩统计量：

$$W = \sum_{i=1}^{n_{12}} i s_i \qquad (5.6.2)$$

(5.6.2)式是两子序列的累计数。其均值和方差为

$$\begin{cases} \mathrm{E}(W) = \dfrac{1}{2}n_1(n_1 + n_2 + 1) \\ \mathrm{Var}(W) = \dfrac{1}{12}n_1 n_2(n_1 + n_2 + 1) \end{cases} \tag{5.6.3}$$

再构造一秩统计量：

$$A = \sum_{i=1}^{n_1} is_i + \sum_{i=n_1+1}^{n_{12}} (n_{12} - i + 1)s_i \tag{5.6.4}$$

可见，(5.6.4)式是两个子序列各自累计数之和。前半部分是基准点之前子序列的累计数，后半部分是基准点之后子序列的累计数。A 的均值和方差为

$$\begin{cases} \mathrm{E}(A) = \dfrac{1}{4}n_1(n_1 + n_2 + 2) \\ \mathrm{Var}(A) = \dfrac{n_1 n_2(n_1 + n_2 - 2)(n_1 + n_2 + 2)}{48(n_1 + n_2 - 1)} \end{cases} \tag{5.6.5}$$

至此，可以构造 Wilcoxon 和 Ansariy-Braley 联合统计量：

$$WA = \frac{[W - \mathrm{E}(W)]^2}{\mathrm{Var}(W)} + \frac{[A - \mathrm{E}(A)]^2}{\mathrm{Var}(A)} \tag{5.6.6}$$

(5.6.6)式是 Lepage 统计量。注意到当样本量足够大时，WA 渐近具有自由度为 2 的 χ^2 分布。

由于需要人为确定子序列长度，因此使用时也应该反复变动子序列的长度，以避免由于突变点的漂移而给解释带来的困难。

5.6.2 计算步骤

(1)确定基准点前后两子序列的样本长度，一般取 $n_1 = n_2 = IH$。

(2)采用连续设置基准点的办法以滑动的方式计算 $n_1 + n_2$ 范围内的 s_i，并按照(5.6.2)～(5.6.6)式计算。由于是以滑动方式计算，因此可以最终得到统计量序列 $WA_i, i = 1, 2, \cdots,$

$n-(n_1+n_2)+1$。n 为一时间序列 x 的样本量。

(3)给定显著性水平，查 χ^2 分布表(附表4)得到自由度为2的临界值。当 WA_i 超过临界值时，表明第 i 时刻前时段的样本与第 i 时刻后的样本之间存在显著性差异，认为 i 时刻发生了突变。

5.6.3 计算结果分析

从绘出的 WA 曲线上的点是否超过临界值来判断序列是否出现突变，并确定出现突变的时间。

应用实例[5.4]：利用 Lepage 检测 1951～1995 年华北地区夏季旱涝指数序列的突变[7]。$n=45$，子序列长度 $n_1=n_2=10$。

计算结果表明，在 1966 和 1979 年处，Lepage 统计量值出现了极大值，且均超过 $\alpha=0.01$ 的显著性水平。说明近 45 年来，华北地区夏季旱涝指数曾经历过两次明显的趋势突变。1966 年的剧烈变化标志着华北地区从相对湿润转变为干旱少雨，这一突变现象与北半球及全国大范围气候变化的大背景是一致的。80 年代又经历了一次明显振动，进入更为干旱时期，这一振动与 80 年代全球气候激烈振荡相协调。从这一实例可见，Lepage 检验有较强的检验突变功效。

子序列长度取为 5 年的计算结果显示，统计量有三处超过 $\alpha=0.05$ 显著性水平，除了在 1966 年和 1979 年处外，还有一处在 1988 年。这一结果验证了上述 1966 和 1979 年的突变。同时表明，从较小的时间尺度来看，80 年代末又有一次突变发生。事实上，华北旱涝实况资料中已显露出 80 年代末 90 年代初干旱趋于缓解的端倪。

参考文献

[1]谷松林．突变理论及其应用．兰州：甘肃教育出版社，1993. 2～5

[2]yamamoto R. T. Iwashima and N. K. Sanga. An analysis of climatic jump. Meteor. Soc. Japan，1986，64(2)：273～281

[3]魏凤英，曹鸿兴. 中国、北半球和全球的气温突变分析及其趋势预测研究．大气科学，1995，19(2)：140～148

[4]符淙斌，王强．气候突变的定义和检验方法．大气科学，1992，16(4)：483～493

[5]Pettitt A. N. A non-Parametric approach to the change point problem，Appl. Statis. 1979，28：125～135

[6]Tsuneharu yonetani. Discontinuous change of precipitation in Japan after 1900 detected by the Lepage test，Meteor. Soc. Japan，1992，70(1)：95～103

[7]魏凤英．华北干旱不同时间尺度的变化特征. 见：王馥棠等主编. 华北农业干旱研究进展．北京：气象出版社，1997. 1～10

第六章　气候序列周期提取方法

近年来，提取时间序列振荡周期的统计方法发展十分迅速。从离散的周期图、方差分析过渡到连续谱分析。然而，周期图不能处理周期的位相突变和周期振幅的变化。方差分析在具体实施时，对原序列寻找一个隐含的显著周期的统计推断是十分巧妙的，但用剩余序列推断第二和第三个周期时，从假设检验意义上讲，就很牵强。就其结果而言，上述两种方法及经典的谐波分析均是从时间域上研究气候序列中周期振荡的方法，它们将气候序列中的周期性视为正弦波，有其固有的局限性，这里不作介绍。

1807 年法国数学家傅立叶提出了在有限时间间隔内定义的任何函数均可以用正弦分量的无限谐波的叠加来表示，这样就出现了与时域相对应的频域。特别是 1965 年出现快速傅里叶变换以来，使频域分析走向实用并迅速拓展。这里将介绍的重点放在以傅立叶变换概念为基础的功率谱、交叉谱及以自回归模型为基础的最大熵谱。

近年来又出现了研究周期现象的新技术——奇异谱分析和时频结构分析的新方法——小波分析，使得提取气候序列周期技术有了新的飞跃，这些将在本章进行介绍。

§6.1　功率谱

功率谱分析是以傅里叶变换为基础的频域分析方法，其意义为将时间序列的总能量分解到不同频率上的分量，根据

不同频率的波的方差贡献诊断出序列的主要周期，从而确定出周期的主要频率，即序列隐含的显著周期。功率谱是应用极为广泛的一种分析周期的方法。有关功率谱的概念、谱分解及傅里叶变换的算法，许多书籍中都有详尽的阐述[1]。这里仅给出有关提取显著周期的具体方法、计算流程及结果分析要点。

6.1.1 方法概述

对于一个样本量为 n 的离散时间序列 $x_1, x_2, \cdots, x_n$，可以使用下面两种完全等价的方法进行功率谱估计。

(1)直接使用傅里叶变换。序列 x_t 可以展成傅里叶级数

$$x_t = a_0 + \sum_{k=1}^{\infty} (a_k \cos wkt + b_k \sin wkt) \qquad (6.1.1)$$

其中 a_0, a_k, b_k 为傅里叶系数。它们可以由下式求得：

$$\begin{cases} a_0 = \dfrac{1}{n} \sum_{t=1}^{n} x_t \\ a_k = \dfrac{2}{n} \sum_{t=1}^{n} x_t \cos \dfrac{2\pi k}{n}(t-1) \\ b_k = \dfrac{2}{n} \sum_{t=1}^{n} x_t \sin \dfrac{2\pi k}{n}(t-1) \end{cases} \qquad (6.1.2)$$

其中 k 为波数，$k=1,2,\cdots,[\dfrac{n}{2}]$，[]表示取整数。不同波数 k 的功率谱值为：

$$\hat{s}_k^2 = \frac{1}{2}(a_k^2 + b_k^2) \qquad (6.1.3)$$

(2)根据谱密度与自相关函数互为傅里叶变换的重要性质，通过自相关函数间接作出连续功率谱估计。对一时间序列 x_t，最大滞后时间长度为 m 的自相关系数 $r(j), j=0,1,2,\cdots, m$ 为：

$$r(j) = \frac{1}{n-j} \sum_{t=1}^{n-j} \left(\frac{x_t - \bar{x}}{s}\right)\left(\frac{x_{t+j} - \bar{x}}{s}\right) \qquad (6.1.4)$$

式中 $\bar{x}$ 为序列的均值，s 为序列的标准差。

由下列得到不同波数 k 的粗谱估计值：

$$\hat{s}_k = \frac{1}{m}\left[r(0) + 2\sum_{j=1}^{m-1} r(j)\cos\frac{k\pi j}{m} + r(m)\cos k\pi\right]$$
$$(k = 0,1,\cdots,m) \qquad (6.1.5)$$

其中 $r(j)$ 表示第 j 个时间间隔上的相关函数。在实际计算中考虑端点特性，常用下列形式：

$$\begin{cases} \hat{s}_0 = \dfrac{1}{2m}[r(0) + r(m)] + \dfrac{1}{m}\sum_{j=1}^{m-1} r(j) \\ \hat{s}_k = \dfrac{1}{m}\left[r(0) + 2\sum_{j=1}^{m-1} r(j)\cos\dfrac{k\pi j}{m} + r(m)\cos k\pi\right] \\ \hat{s}_m = \dfrac{1}{2m}[r(0) + (-1)^m r(m)] - \dfrac{1}{m}\sum_{j=1}^{m-1}(-1)^j r(j) \end{cases} \qquad (6.1.6)$$

最大滞后时间长度 m 是给定的。在已知序列样本量为 n 的情况下，功率谱估计随 m 的不同而变化。当 m 取较大值时，谱的峰值就多，但这些峰值并不表明有对应的周期现象，而可能是对真实谱的估计偏差造成的虚假现象。当 m 取太小值时，谱估计过于光滑，不容易出现峰值，难以确定出主要周期。因此最大滞后长度的选取十分重要，一般 m 取为 $\frac{n}{3}\sim\frac{n}{10}$ 为宜。

上述两种方法得到的谱估计都与真实谱存在一定误差。因而对粗谱估计需要作平滑处理，以便得到连续性的谱值。常用 Hanning 平滑系数

$$\begin{cases} s_0 = 0.5\hat{s}_0 + 0.5\hat{s}_1 \\ s_k = 0.25\hat{s}_{k-1} + 0.5\hat{s}_k + 0.25\hat{s}_{k+1} \\ s_m = 0.5\hat{s}_{m-1} + 0.5\hat{s}_m \end{cases} \qquad (6.1.7)$$

来进行平滑。

6.1.2 计算步骤

上面给出了计算谱估计值的两种方法。那么，如何利用功

率谱提取隐含在气候序列中的显著周期呢？下面给出通过自相关系数间接求谱估计值，从而确定显著周期的计算步骤：

(1)据(6.1.4)式计算自相关系数。

(2)据(6.1.6)式计算粗谱估计值。

(3)据(6.1.7)式计算平滑谱估计值。

(4)确定周期。周期值与波数 k 的关系是：

$$T_k = \frac{2m}{k} \tag{6.1.8}$$

(5)对谱估计作显著性检验。为了确定谱值在哪一波段最突出并了解该谱值的统计意义，需要求出一个标准过程谱以便比较。标准谱有两种情况：

①红噪音标准谱：

$$s_{0k} = \bar{s}\left[\frac{1 - r(1)^2}{1 + r(1)^2 + 2r(1)\cos\dfrac{\pi k}{m}}\right] \tag{6.1.9}$$

其中 $\bar{s}$ 为 $m+1$ 个谱估计值的均值，即

$$\bar{s} = \frac{1}{2m}(s_0 + s_m) + \frac{1}{m}\sum_{k=1}^{m-1} s_k \tag{6.1.10}$$

②白噪音标准谱：

$$s_{0k} \equiv \bar{s} \tag{6.1.11}$$

如果序列的滞后自相关系数 $r(1)$ 为较大正值时，表明序列具有持续性，用红噪音标准谱检验。若 $r(1)$ 接近 0 或为负值时，表明序列无持续性，用白噪音标准谱检验。

假设总体谱是某一随机过程的谱，记为 $E(s)$ 则

$$\frac{s}{E(s)/\nu} = \chi^2_\nu \tag{6.1.12}$$

遵从自由度为 ν 的 χ^2 分布。自由度 ν 与样本量 n 及最大滞后长度 m 有关，即

$$\nu = (2n - \frac{m}{2})/m \qquad (6.1.13)$$

给定显著性水平 α，查附表 4 得到 χ^2_α 值。计算

$$s'_{0k} = s_{0k}(\frac{\chi^2_\alpha}{\nu}) \qquad (6.1.14)$$

若谱估计值 $s_k > s'_{0k}$，则表明 k 波数对应的周期波动是显著的。

编程计算时，可以给定一显著性水平，如 $\alpha = 0.05$，将 χ^2 分布表中对应的不同自由度的 χ^2 值赋给某一数组，然后依(6.1.14)式计算出 s'_{0k}。

6.1.3 计算结果分析

将功率谱估计和标准谱绘成曲线图。根据绘出的曲线确定序列的显著周期。首先看功率谱估计曲线的峰点是否超过标准谱，若超过则说明峰点所对应的周期是显著的。这一周期是序列存在的第一显著周期。再从图上找次峰点，再次峰点…看其是否超过标准谱，从中找出第二、第三…显著周期。

应用实例[6.1]：取 1882～1995 年南方涛动指数序列计算功率谱。$n = 114$，最大滞后长度 m 取为 $\frac{n}{3}$。计算时首先对指数序列作 10 年滑动平均处理。计算标准谱的显著性水平 α 取为 0.05。计算结果如图 6.1 所示。横坐标值为周期，纵坐标为谱值。

由图 6.1 可以清楚地看出，在周期长度为 6.8 年处功率谱估计值为一峰值且大大超过标准谱，因此 6.8 年是第一显著周期。其次在 6.1 和 7.5 年处功率谱估计值也超过标准谱。因此，可以确定南方涛动指数存在 6～7 年的周期振荡。另外在 4.25 年处还有一峰点，谱估计值超过标准谱。可见，南方涛动指数还存在 4 年左右的另一显著周期。

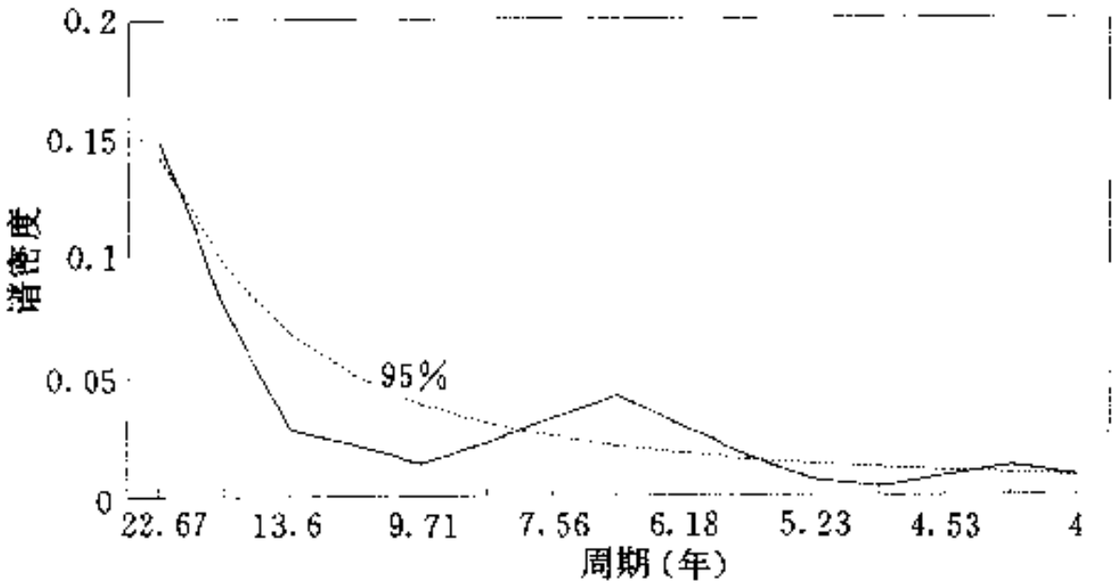


图 6.1　南方涛动指数功率谱

(光滑曲线为 $\alpha=0.05$ 的红噪音标准谱)

§6.2　最大熵谱

从上述介绍中我们知道连续功率谱估计需要借助于谱窗函数对粗谱加以平滑而求得。因此，其统计稳定性和分辨率都与选择的窗函数有关。例如：(6.1.7)式就是一种对应于 Hamming 窗函数的平滑公式。由于使用了与分析的系统毫无关系的窗函数，有时可能会得出虚假的结论。另外，在连续功率谱估计中，自相关函数估计与样本量大小有关，这也会造成谱估计的误差，影响分辨率。可见，功率谱存在分辨率不高和有可能产生虚假频率分量等缺点。由于功率谱不需要由时间序列本身提供某种参数模式，因而是一种非参量谱估计。1967 年 Burg 提出了一种称之为"最大熵"谱估计的方法，从而将谱估计推进了一个新的阶段。最大熵谱的基本思想是，以信息论中熵的概念为基础，选择这样一种谱估计——在外推已知

时间序列的自相关函数时，其外推原则是使相应的序列在未知点上取值的可能性具有最大的不确定性，也就是不对结果作人为主观地干预，因而所得信息最多。最大熵谱估计是以确定时间序列的参数模式——自回归模型有关的方法，是一种参数谱估计。最大熵谱具有分辨率高等优点，尤其适用于短序列，因此它受到人们的广泛重视。

6.2.1 方法概述

Burg 将“熵”的概念引入到谱估计中，提出了最大熵谱估计。在统计学中用“熵”作为各种随机事件不确定性程度的度量。假定研究的随机事件只有 n 个相互独立的结果，它们相应的概率为 $P_i, i=1,2,\cdots,n$，且满足 $\sum_{i=1}^{n} P_i=1$。已经证明，可以用熵 H 来度量随机事件不确定性的程度：

$$H=-\sum_{i=1}^{n} P_i \lg P_i \tag{6.2.1}$$

对均值为 0，方差为 σ^2 的正态分布序列 x 有

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/2\sigma^2} \tag{6.2.2}$$

则有

$$H=\ln\sigma\sqrt{2\pi e} \tag{6.2.3}$$

由信息论可知，随机事件以等概率可能性出现时，熵值达到极大。由(6.2.3)式可知，熵谱越大，对应的方差 σ^2 越大。将(6.2.3)式推广，且考虑方差与功率谱的关系，则有

$$H=\int_{-\omega}^{\omega} \ln s(w)\mathrm{d}w \tag{6.2.4}$$

功率谱与自相关函数间有下列关系：

$$r(j)=\int_{-\omega}^{\omega} s(w)e^{jw}\mathrm{d}w \tag{6.2.5}$$

(6.2.5)式表明，自相关函数 $r(j)$ 与谱密度 $s(w)$ 按傅立叶变换一一对应。然而对有限的样本序列，只有有限个 $r(j)$ 估计值来代替 $r(j)$。因此，关键的问题在于如何利用 $r(j)$ 提供的信息去估计谱密度 $s(w)$。利用泛函分析中拉格朗日乘子法可以证明，欲使谱估计满足(6.2.5)式，且使熵谱为最大，则其谱密度

$$S_H(w)=\frac{\sigma_{k_0}^2}{\left|1-\sum_{k=1}^{k_0}a_k^{(k_0)}e^{-iwk}\right|^2} \qquad (6.2.6)$$

式中 k_0 为自回归的阶数；$a_i^{(k_0)}$ 为自回归系数；$\sigma_{k_0}^2$ 为预报误差方差估计。由(6.2.6)式可见，最大熵谱估计实质上是自回归模型的谱。

最大熵谱最流行的算法是由 Burg 设计的算法。Burg 算法的思路是，建立适当阶数的自回归模型，并利用(6.2.6)式计算出最大熵谱。在建立自回归模型中，必须根据某种准则截取阶数 k_0，并递推算出各阶自回归系数。

变量 x 的自回归模型为：

$$x_t=a_1x_{t-1}+a_2x_{t-2}+\cdots+a_kx_{t-k}+\varepsilon_t \qquad (6.2.7)$$

式中 $a_1,a_2,\cdots,a_k$ 为自回归系数；ε_t 为白噪声。在线性系统中，将自回归模型看作预报误差滤波器，输入为 x_t，输出为 ε_t，(6.2.7)式可以写为：

$$\varepsilon_t=x_t-a_1x_{t-1}-a_2x_{t-2},\cdots,a_kx_{t-k} \qquad (6.2.8)$$

假设均值为 0，k 阶预报误差滤波器输出方差为 σ_k^2，则相应的系数为 $a_{k_1},a_{k_2},\cdots,a_{k_k}$。那么，零阶($k=0$)预报误差滤波器输出方差的估计值为：

$$\sigma_0^2=\frac{1}{n}\sum_{t=1}^{n}x_t^2=r(0)>0 \qquad (6.2.9)$$

根据 Yule-Walk 方程可以推出 $k=1$ 时预报误差滤波器输出方差的估计值为：

$$\begin{cases} r(1) = a_{11}\sigma_0^2 \\ \sigma_1^2 = (1 - a_{11}^2)\sigma_0^2 \\ a_{11} = 2\sum_{t=2}^{n} x_t x_{t-1} \Big/ \sum_{t=2}^{n} (x_t^2 + x_{t-1}^2) \end{cases} \tag{6.2.10}$$

$k=2$ 时，

$$\begin{cases} r(2) = a_{11}r(1) + a_{11}\sigma_1^2 \\ \sigma_2^2 = \sigma_1^2 - a_{22}[r(2) - a_{11}r(1)] = [1 - a_{22}^2]\sigma_1^2 \\ a_{21} = a_{11} - a_{22}a_{11} \\ a_{22} = \dfrac{\sum_{t=3}^{n} (x_t - a_{11}x_{t-1})(x_{t-2} - a_{11}x_{t-1})}{\sum_{t=3}^{n} [(x_t - a_{11}x_{t-1})^2 - (x_{t-2} - a_{11}x_{t-1})^2]} \end{cases} \tag{6.2.11}$$

由归纳法可以导出递推公式，在 $a_{kj}(j=1,2,\cdots,k)$ 已知情况下，求 $a_{k+1,k+1}$

$$\begin{cases} r(k+1) = \sum_{j=1}^{k} a_{kj} \cdot r(k+1-j) + a_{k+1,k+1}\sigma_k^2 \\ \sigma_{k+1}^2 = (1 - a_{k+1,k+1}^2)\sigma_k^2 \\ a_{k+1,j} = a_{kj} - a_{k+1,k+1}a_{k,k+1-j} \\ a_{k+1,k+1} = \dfrac{\left[2\sum_{t=k+2}^{n} (x_t - \sum_{j=1}^{k} a_{kj}x_{t-j})(x_{t-k-1} - \sum_{j=1}^{k} a_{kj}x_{t-k-1+j})\right]}{\sum_{t=k+2}^{n} \left[(x_t - \sum_{j=1}^{k} a_{kj}x_{t-j})^2 + (x_{t+k-1} - \sum_{j=1}^{k} a_{kj}x_{t-k-1+j})^2\right]} \end{cases} \tag{6.2.12}$$

由上面递推过程可以看到，Burg 算法巧妙之处在于直接从序列来计算谱密度中的参数，不必提前算出自相关函数。

确定自回归模型的阶数 k_0 可以采用下面几种准则：

(1)最终预测误差(FPE)准则。这一准则是由 Akaike 提出的，其含义为：如果用由过程的一组采样所算出的自回归模型来估计同一过程的另一组采样，则会有预测均方误差，该误差在某一个 k 值时最小。当过程的均值为 0 时，k 阶自回归模型的 FPE 定义为：

$$FPE(k) = \frac{n+k}{n-k}\sigma_k^2 \quad (k = 1,2,\cdots,n-1) \qquad (6.2.13)$$

由于 σ_k^2 随 k 的增加而减少，而 $\frac{n+k}{n-k}$ 项随 k 的增加而增大，所以在某一个 k 值时，$FPE(k)$ 将出现最小值。根据最终预测误差准则，这个 k 值就定义为自回归模型的最佳阶数。

(2)信息论准则(AIC)。AIC 准则是由 Akaike 将统计学中根据极大似然原理估计参数的方法加以改进而提出来的。AIC 准则定义为：

$$AIC(k) = \ln\sigma_k^2 + 2k/n \qquad (6.2.14)$$

由(6.2.14)式可以看出，AIC 准则是通过预测均方误差与模型阶数的权衡来确定模型的。显而易见，以 AIC 值达到最小为准则确定自回归模型的阶数。从数学上可以证明，在一定条件下 FPE 与 AIC 是等价的。

(3)自回归传输函数准则(CAT)。CAT 准则是由 Parzen 提出的。按照这个准则，当自回归模型与估计自回归模型二者均方误差之差的估计值为最小时，自回归的阶数就是最佳阶数。CAT 准则定义为：

$$CAT(k) = \frac{1}{n}\sum_{j=1}^{k}\frac{n-j}{n\sigma_j^2} - \frac{n-k}{n\sigma_k^2} \qquad (6.2.15)$$

6.2.2 计算步骤

归纳起来，用 Burg 算法的最大熵谱提取序列显著周期

的计算步骤为：

(1)对时间序列 $x_1, x_2, \cdots, x_n$ 用 Burg 递推公式(6.2.10)～(6.2.12)，计算 $k=1,2,\cdots,n-1$ 各阶试验模型，同时以 FPE 准则或其它准则确定自回归最佳阶数 k_0。

(2)用 k_0 代入递推公式中，计算出最终的自回归系数 a_1，$a_2, \cdots, a_{k_0}$。

(3)用(6.2.6)式计算最大熵谱。在实际计算中，通常采用离散形式。因为，

$$e^{-iwk} = \cos wk - i\sin wk$$

所以，

$$\left|1 - \sum_{k=1}^{k_0} a_k^{(k_0)} e^{-iwk}\right|^2 = \left(1 - \sum_{k=1}^{k_0} a_k^{(k_0)} \cos wk\right)^2 + \left(\sum_{k=1}^{k_0} a_k^{(k_0)} \sin wk\right)^2$$

在计算离散谱值时，频率取 $w_l = \frac{2\pi l}{2m}$ $(l=0,1,2,\cdots,m)$，m 为选取的最大波数，在序列样本量不大时，m 通常取为 $\frac{n}{2}$。m 对应的周期为 $T_l = \frac{2m}{l}$，这时，可以得到最大熵谱的离散形式：

$$S_H(l) = \frac{\sigma_{k_0}^2}{\left[1 - \sum_{k=1}^{k_0} a_k^{(k_0)} \cos\left(\frac{\pi lk}{m}\right)\right]^2 + \left[\sum_{k=1}^{k_0} a_k^{(k_0)} \sin\left(\frac{\pi lk}{m}\right)\right]^2}$$

6.2.3 计算结果分析

将计算出的最大熵谱谱密度绘成图。如果谱密度有尖锐的峰点，其对应的周期就是序列存在的显著周期。

用 Burg 递推估计出的谱密度有时也会出现峰值漂移或出现将真实峰值估计成两个或多个接近的峰值现象，对这种

现象可以采用 Marple 方法进行纠正[2]。

应用实例[6.2]：用最大熵谱提取 1952～1995 年华北地区春季干旱指数序列的显著周期[3]。样本量 $n=44$，最大波数 m 取为 22，计算各阶试验自回归模型。用 FPE 准则确定出最佳阶数 $k_0=4$。计算结果绘成最大熵谱图(图 6.2)。由图可见，有两个明显的峰点，最高峰值对应在 6.29 年周期上。次峰值对应在 2.93 年周期上。

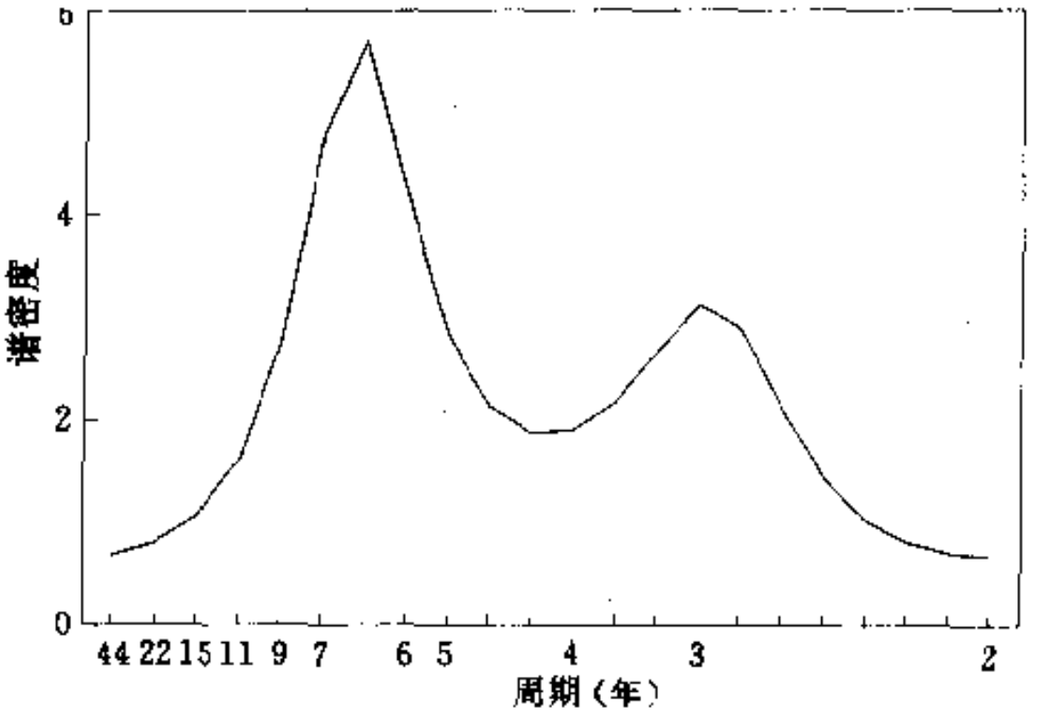


图 6.2 华北春季干旱指数最大熵谱

§6.3 交叉谱

在实际问题中，我们不仅要研究单个气候序列的频域结构和周期特性，还要分析不同序列在频域变化上的相互关系。因此，需要讨论多个序列(这里仅限两个序列)之间的交叉谱。

6.3.1 方法概述

对于时间序列 $x_1(t)$ 和 $x_2(t)$，两个序列的互相关函数反

映它们的交叉能量，可以表示为：

$$r_{12}(j) = \int_{-\infty}^{\infty} x_1(t)x_2(t)\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_2(w)f_1^*(w)\mathrm{e}^{iwj}\mathrm{d}w = \frac{1}{2\pi}\int_{-\infty}^{\infty} s_{12}(w)\mathrm{e}^{iwj}\mathrm{d}w \quad (j = 0,1,2,\cdots,m) \tag{6.3.1}$$

式中 $f_1(w)$ 和 $f_2(w)$ 分别是 $x_1(t)$ 和 $x_2(t)$ 的复谱；$s_{12}(w)$ 称为 $x_1(t)$ 和 $x_2(t)$ 的交叉谱；m 为最大滞后时间长度。交叉谱由下式求出：

$$s_{12}(w) = f_2(w)f_1^*(w) = \int_{-\infty}^{\infty} r_{12}(j)\mathrm{e}^{-iwj}\mathrm{d}j \tag{6.3.2}$$

交叉谱是复谱，可以用实部与虚部形式表示：

$$s_{12}(w) = P_{12}(w) - iQ_{12}(w) \tag{6.3.3}$$

式中 $P_{12}(w)$ 为实部谱，称为协谱；$Q_{12}(w)$ 为虚部，称为正交谱。其中

$$\begin{cases} P_{12}(w) = \int_{-\infty}^{\infty} r_{12}(j)\cos wj\mathrm{d}j \\ Q_{12}(w) = \int_{-\infty}^{\infty} r_{12}(j)\sin wj\mathrm{d}j \end{cases} \tag{6.3.4}$$

互相关函数具有交叉关系对称性，即

$$\begin{cases} r_{12}(-j) = r_{21}(j) \\ r_{12}(j) = r_{21}(-j) \end{cases} \tag{6.3.5}$$

应用交叉关系对称性(6.3.5)，协谱可以化为：

$$P_{12}(w) = \int_{-\infty}^{0} r_{12}(j)\cos wj\mathrm{d}j + \int_{0}^{\infty} r_{12}(j)\cos wj\mathrm{d}j = -\int_{0}^{\infty} r_{12}(-j)\cos(-wj)\mathrm{d}(-j) + \int_{0}^{\infty} r_{12}(j)\cos wj\mathrm{d}j = \int_{0}^{\infty} r_{21}(j)\cos wj\mathrm{d}j + \int_{0}^{\infty} r_{12}(j)\cos wj\mathrm{d}j = \int_{0}^{\infty} [r_{12}(j) + r_{21}(j)]\cos wj\mathrm{d}j \tag{6.3.6}$$

协谱的含义为两个时间序列在某一频率 w 上同位相的相关程度。

同样，正交谱可以化为：

$$Q_{12}(w) = \int_0^\infty [r_{12}(j) - r_{21}(j)]\sin wj\,dj \qquad (6.3.7)$$

正交谱的含义是某一频率上两序列位相差 90°时的交叉相关关系。

根据协谱和正交谱可以得到两个序列的振幅谱、位相谱和凝聚谱

$$C_{12}(w) = \sqrt{P_{12}^2(w) + Q_{12}^2(w)} \qquad (6.3.8)$$

振幅谱 $C_{12}(w)$ 反映的是两个序列分解出的某一频率振动的能量关系。

$$\Theta_{12}(w) = \arctan \frac{Q_{12}(w)}{P_{12}(w)} \qquad (6.3.9)$$

显见，位相谱 $\Theta_{12}(w)$ 反映的是两序列各个频率波动的位相差关系，其值在 $-\frac{\pi}{2} \sim \frac{\pi}{2}$ 之间变化。

$$R_{12}^2(w) = \frac{P_{12}^2(w) + Q_{12}^2(w)}{P_{11}(w) \cdot P_{22}(w)} \qquad (6.3.10)$$

式中 $P_{11}(w)$ 和 $P_{22}(w)$ 分别为序列 $x_1(t)$ 和 $x_2(t)$ 自身的交叉谱，即单个序列的功率谱。凝聚谱 $R_{12}^2(w)$ 代表两序列各个频率之间的相关程度，其值在任何频率下都在 0～1 之间变化。

与功率谱一样，交叉谱也有直接和间接两种计算方法。

(1)直接使用傅里叶变换。将一个函数的傅里叶变换：

$$\begin{aligned} f_1(w) &= a_1(w) - ib_1(w) \\ f_1^*(w) &= a_1(w) + ib_1(w) \\ f_2(w) &= a_2(w) - ib_2(w) \end{aligned} \qquad (6.3.11)$$

代入(6.3.2)式得到：

$$S_{12}(w) = f_2(w)f_1^*(w) = [a_2(w) - ib_2(w)][a_1(w) + ib_1(w)] = [a_1(w)a_2(w) + b_1(w)b_2(w)] - i[a_1(w)b_2(w) - a_2(w)b_1(w)] \quad (6.3.12)$$

其中 $a_1(w)\to\frac{1}{2}a_{1k}$，$b_1(w)\to\frac{1}{2}b_{1k}$，$a_2(w)\to\frac{1}{2}a_{2k}$，$b_2(w)\to\frac{1}{2}b_{2k}$，$a_{1k}$，$b_{1k}$，$a_{2k}$，$b_{2k}$分别为序列 $x_1(t)$，$x_2(t)$离散傅里叶系数，这里作为相应频率的近似估计值。

将(6.3.12)式写成离散交叉谱形式：

$$S_{12k} = \frac{1}{4}[(a_{1k}a_{2k} + b_{1k}b_{2k}) - i(a_{1k}b_{2k} - a_{2k}b_{1k})] \quad (6.3.13)$$

离散形式的协谱与正交谱为：

$$P_{12k} = \frac{1}{4}(a_{1k}a_{2k} + b_{1k}b_{2k})$$
$$Q_{12k} = \frac{1}{4}(a_{1k}b_{2k} - a_{2k}b_{1k}) \quad (6.3.14)$$

将(6.3.14)式代入(6.3.8)、(6.3.9)和(6.3.10)式就可以得到振幅谱、位相谱和凝聚谱。

(2)通过计算落后互相关系数间接求出连续交叉谱估计。

首先计算落后互相关函数：

$$r_{12}(j) = \frac{1}{n-j}\sum_{i=1}^{n-j}\left(\frac{x_{1i}-\bar{x}_1}{s_1}\right)\left(\frac{x_{2(i+j)}-\bar{x}_2}{s_2}\right)$$
$$r_{21}(j) = \frac{1}{n-j}\sum_{i=1}^{n-j}\left(\frac{x_{1(i+j)}-\bar{x}_1}{s_1}\right)\left(\frac{x_{2i}-\bar{x}_2}{s_2}\right) \quad (6.3.15)$$

式中 $\bar{x}_1$，$\bar{x}_2$ 为 $x_1(t)$和 $x_2(t)$的平均值，s_1 和 s_2 是它们的标准差。将(6.3.6)和(6.3.7)式化为有限求和形式

$$P_{12}(k)=\frac{1}{m}\{r_{12}(0)+\sum_{j=1}^{m-1}[r_{12}(j)+r_{21}(j)]\cos\frac{k\pi}{m}j+$$
$$r_{12}(m)\cos k\pi\} \qquad (6.3.16)$$
$$Q_{12}(k)=\frac{1}{m}\sum_{j=1}^{m-1}[r_{12}(j)-r_{21}(j)]\sin\frac{k\pi}{m}j$$

应用 Hanning 光滑系数(6.1.7)式计算谱的估计值 $\hat{P}_{12}(k)$和 $\hat{Q}_{12}(k)$。再将它们代入(6.3.8)、(6.3.9)和(6.3.10)式得到振幅谱、位相谱和凝聚谱。

6.4.2 计算步骤

下面给出用间接方法进行两序列交叉谱分析的计算步骤：

(1)确定出最后滞后长度 m，利用(6.3.15)式计算滞后交叉相关系数 $r_{12}(j)$，$r_{21}(j)$。

(2)利用(6.3.16)式计算协谱 $P_{12}(k)$和正交谱 $Q_{12}(k)$。

(3)利用 Hanning 平滑公式对 $P_{12}(k)$和 $Q_{12}(k)$进行平滑得到 $\hat{P}_{12}(k)$和 $\hat{Q}_{12}(k)$。

(4)分别计算 $x_1(t)$和 $x_2(t)$的光滑功率谱，得到 $\hat{P}_{12}(k)$和 $\hat{P}_{22}(k)$。

(5)将 $\hat{P}_{12}(k)$，$\hat{Q}_{12}(k)$，$\hat{P}_{11}(k)$和 $\hat{P}_{22}(k)$代入(6.3.8)、(6.3.9)和(6.3.10)式得到振幅谱 $C_{12}(k)$，位相谱 $\Theta_{12}(k)$和凝聚谱 $R^2_{12}(k)$。在实际使用中，位相谱 $\Theta_{12}(k)$通常用时间长度来表示，利用(6.1.8)式，可以从位相角与同期的关系计算落后时间长度谱：

$$L(k)=\frac{m\Theta_{12}(k)}{\pi k} \qquad (6.3.17)$$

(6)对凝聚谱 $R^2_{12}(k)$值进行显著性检验。原假设：在某一频率上两序列振动的相关程度为 0，即凝聚谱为 0。计算统计

量：

$$F = \frac{(\nu - 1)R_{12}^2}{1 - R_{12}^2} \tag{6.3.18}$$

上面统计量是遵从分子自由度为 2，分母自由度为 $2(\nu-1)$的 F 分布。其中 $\nu=\frac{2n-(m-1)/2}{m-1}$。确定显著性水平 α，查附表 3b 得 F_α，若 $F>F_\alpha$ 则拒绝原假设，认为在某一频率上两序列振动的凝聚是显著的。

6.4.3 计算结果分析

由上述计算流程可知，两序列交叉谱分析得到 5 种谱估计——协谱、正交谱、振幅谱、位相谱（落后时间长度谱）和凝聚谱，5 种谱估计相互有密切联系，其中凝聚谱和位相谱是分析的主要对象。

(1)利用协谱估计和正交谱估计可以分别分析两序列在某一频率上同位相相关关系和位相差 90°的相关关系。

(2)利用振幅谱分析在某一频率上同位相相关和位相差 90°相关关系的能量大小。

(3)利用凝聚谱分析两序列在某一频率上振动相关的程度。如果在某一频率上所对应的凝聚值通过显著性检验，则证明两序列在这一频率上存在密切的相关关系。那么，存在怎样的相关关系呢？依据其对应的位相谱（或更经常使用落后时间长度谱）来分析这两序列在这一频率上存在落后多长时间尺度的相关关系。

应用实例[6.3]：对南京、上海两站 1951～1996 年 7 月降水量序列进行交叉谱分析。样本量 $n=46$，最大滞后长度 $m=20$。主要结果列于表 6.1。

表 6.1　南京、上海两站 7 月降水量交叉谱分析

k	$T(k)$	$R_{12}^2(k)$	$\Theta_{12}(k)$	$L(k)$	k	$T(k)$	$R_{12}^2(k)$	$\Theta_{12}(k)$	$L(k)$
0	∞	—	—	—	10	3.6	0.665	0.410	0.248
1	40.0	1.223	1.277	7.723	11	3.3	0.273	1.369	0.753
2	20.0	0.369	0.343	1.037	12	3.1	0.380	−0.290	−0.146
3	10.3	0.177	1.462	2.947	13	2.9	0.923	−0.397	−0.185
4	10.0	0.188	0.574	0.868	14	2.7	0.767	−0.268	−0.116
5	8.0	0.104	0.754	0.912	15	2.5	0.781	0.366	0.147
6	6.7	0.733	0.634	0.639	16	2.4	0.592	0.168	0.064
7	5.0	0.325	1.073	0.927	17	2.2	0.405	−0.358	−0.128
8	4.4	0.158	0.506	0.382	18	2.1	0.369	−0.040	−0.014
9	4.0	0.546	0.00	0.00	19	2.0	—	—	—

从表 6.1 凝聚谱 $R_{12}^2(k)$一栏中看到，在 2.9 年周期上凝聚出现了最大值，其附近 2.7～2.4 年周期段上的凝聚也较高。另外，在 6.7 和 3.6 年周期上凝聚也比较高。要确定南京、上海两地降水量序列在上述几个周期上是否存在显著的相关关系，要进行显著性检验。由于 $n=46, m=20$，因此自由度 $\nu=(2n-\frac{m-1}{2})/(m-1)=(2\times46-\frac{20-1}{2})/(20-1)=4.34$。将上述几个较高凝聚值逐一代入(6.4.18)式。6.7 年的 F 值为 9.18；3.6 年为 6.63；2.9 年为 40.04；2.7 年为 10.99；2.5 年为 11.91；2.4 年为 4.85。确定显著性水平 $\alpha=0.05$，查分子自由度为 2，分母自由度为 $2(\nu-1)=6.68$ 的 F 检验表，$F_{0.05}=4.74$。上述几个周期的 F 计算值均大于 $F_{0.05}$，因此上述几个周期上振动的凝聚是显著的。也就是说，南京、上海两站 7 月降水量在上述几个周期上的振动存在显著的相关关系。

从表 6.1 的落后长度谱 $L(k)$一栏中可以查到上述高凝聚对应的落后长度。6.7 年对应 $L(6)=0.639$，表明在 6 年左

右周期相关关系中，南京 7 月降水量比上海 7 月降水量落后半年左右。类似，在 3.6 年的周期关系中，南京 7 月降水量比上海 7 月降水量落后 0.248 年，在 2.9 和 2.7 年的周期关系中，南京 7 月降水量则比上海 7 月降水量分别超前 0.185 和 0.116 年。

§6.4 多维最大熵谱

在上一节我们介绍了使用交叉谱研究两个气候序列之间的凝聚和位相关系。在提取单个气候序列的周期时，最大熵谱表现出比普通功率谱的优越。这一节将最大熵谱推广到多变量形式——多维最大熵谱，它描述的是多个不同气候时间序列之间的交叉能量关系，是一种估计复合谱[4]。多维最大熵谱最早主要应用于地质信号处理中，后来推广到雷达、通讯的信号处理领域，各种计算方法在应用中相继应运而生。在雷达、通讯领域通常将其称为多信道最大熵谱。因此在方法的描述上有关术语通常使用通讯信号处理的用法。

6.4.1 方法概述

假设有 l 个变量，样本量均为 n，用一个 $l\times 1$ 的列向量 x_i 表示 l 个序列在 i 时刻的值，即

$$x_i = \begin{bmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(l)} \end{bmatrix} \tag{6.4.1}$$

这里假设 x_i 是复数，均值为 0，那么，滞后长度为 j 的相关函数定义为：

$$R_x(j) = \mathrm{E}\left[x_i x_{i-j}^{\mathrm{H}}\right] \tag{6.4.2}$$

(6.4.2)式中 E[]表示数学期望算子，上标 H 表示共轭转置(关于复数、共轭转置，Hermite 矩阵的概念及运算，在第七章介绍复经验正交函数时加以叙述，必要时可参阅)。$R_x(j)$类似于单变量的滞后长度为 j 的自相关函数，对角线元素代表每个序列的自相关函数，其余元素则表示两两序列之间的相关函数。值得注意的是，由于 x_i 是复数，故矩阵 $R_x(j)$是复数矩阵——Hermite 阵。

对于 m 阶多维预测误差滤波器(同单变量一样，将自回归模型视为预测误差滤波器)，可以定义$(m+1)\times(m+1)$的分块相关矩阵

$$R_x=\begin{bmatrix} R_x(0) & R_x(-1) & \cdots & R_x(-m) \\ R_x(1) & R_x(0) & \cdots & R_x(1-m) \\ \vdots & \vdots & & \vdots \\ R_x(m) & R_x(m-1) & \cdots & R_x(0) \end{bmatrix} \tag{6.4.3}$$

当滤波器的阶数 $m\geqslant|j|$时，序列 $R_x(j)$的最大熵一般可以表示为下列 z 变换形式：

$$S_H(w)=A_m^{-1}(z)P_{f,m}A_m^{-\mathrm{H}}(z^{-1})= B_m^{-1}(z)P_{b,m}B_m^{-\mathrm{H}}(z^{-1}) \tag{6.4.4}$$

式中上标-1表示相应矩阵的逆。其中

$$A_m(z)=I+A_1^{(m)}z^{-1}+\cdots+A_{m-1}^{(m)}z^{1-m}+A_m^{(m)}z^{-m} \tag{6.4.5}$$

$$B_m(z)=B_m^{(m)}+B_{m-1}^{(m)}z^{-1}+\cdots+B_1^{(m)}z^{1-m}+Iz^{-m} \tag{6.4.6}$$

(6.4.5)和(6.4.6)式中 z^{-1}为单位延时算子，(6.4.4)中 $P_{f,m}$和 $P_{b,m}$为向前和向后预测误差功率：

$$P_{f,m}=\mathrm{E}[(e_{f,i}^{(m)})(e_{f,i}^{(m)})^{\mathrm{H}}] \tag{6.4.7}$$

$$P_{b,m}=\mathrm{E}[(e_{b,i}^{(m)})(e_{b,i}^{(m)})^{\mathrm{H}}] \tag{6.4.8}$$

(6.4.7)和(6.4.8)式中 $e_f^{(m)}$ 和 $e_b^{(m)}$ 分别为向前和向后预测误差。(6.4.5)和(6.4.6)式中的滤波器系数矩阵 $A_k^{(m)}$、$B_k^{(m)}$ $(k=0,1,\cdots,m)$，可以利用递推方法算出。但是用这种方法不能得到唯一的最大熵谱估计。

Morf 等人提出了一种归一化递推方法，系数矩阵直接由已知数据估计出来，产生唯一的最小相位预测误差滤波器，从而确保多维最大熵谱估计的唯一性。对于一个 m 阶多维预测误差滤波器，利用递推关系，可以分别定义一系列归一化向前和向后的计算公式。式中带"～"表示归一化的变量。

6.4.1.1 系数矩阵

$$\tilde{A}_k^{(m)} = P_m^{-1/2}\left[\tilde{A}_k^{(m-1)} - \rho_m \tilde{B}_{m-k}^{(m-1)}\right] \tag{6.4.9}$$

$$\tilde{B}_k^{(m)} = Q_m^{-1/2}\left[\tilde{B}_k^{(m-1)} - \rho_m^H \tilde{A}_{m-k}^{(m-1)}\right] \tag{6.4.10}$$

$$(k=0,1,\cdots,m-1)$$

当阶数 $m=0$ 时，滤波器系数的初始条件为

$$(\tilde{A}_0^{(0)})^{-1} = \left(\sum_{i=1}^{n} x_i x_i^H\right)^{\frac{1}{2}} \tag{6.4.11}$$

$$(\tilde{B}_0^{(0)})^{-1} = \left(\sum_{i=0}^{n-1} x_i x_i^H\right)^{\frac{1}{2}} \tag{6.4.12}$$

(6.4.9)和(6.4.10)式中 ρ_m 称为反射系数矩阵，定义为

$$\rho_m = P_{f,m-1}^{1/2}\,\Delta m\, P_{b,m-1}^{-H/2} \tag{6.4.13}$$

其中

$$\Delta m = \sum_{k=0}^{m-1} A_k^{(m-1)} R_x(m-k) = \sum_{k=0}^{m-1} B_k^{(m-1)} R_x(m-k) \tag{6.4.14}$$

(6.4.9)和(6.4.10)式中 $P_m^{1/2}$ 和 $Q_m^{1/2}$ 分别定义为：

$$P_m^{1/2} = P_{f,m-1}^{1/2} \cdot P_{f,m}^{1/2} \tag{6.4.15}$$

$$Q_m^{1/2} = P_{b,m-1}^{1/2} \cdot P_{b,m}^{1/2} \tag{6.4.16}$$

在阶数 $m=0$ 时，其初始条件为

$$P_0 = Q_0 = R_x(0) \tag{6.4.17}$$

利用反射系数，可以求出：

$$P_m = I - \rho_m \rho_m^{\mathrm{H}} \tag{6.4.18}$$

$$Q_m = I - \rho_m^{\mathrm{H}} \rho_m \tag{6.4.19}$$

6.4.1.2 向前和向后预测误差功率

$$\tilde{P}_{f,m} = \sum_{i=m+1}^{n} \tilde{e}_{f,i}^{(m)} (\tilde{e}_{f,i}^{(m)})^H \tag{6.4.20}$$

$$\tilde{P}_{b,m} = \sum_{i=m-1}^{n} \tilde{e}_{b,i-1}^{(m)} (\tilde{e}_{b,i-1}^{(m)})^H \tag{6.4.21}$$

向前和向后预测误差的互功率为：

$$\tilde{P}_{fb,m} = \sum_{i=m+1}^{n} \tilde{e}_{f,i}^{(m)} (\tilde{e}_{b,i-1}^{(m)})^H \tag{6.4.22}$$

6.4.1.3 传输函数

$$\tilde{A}_m(z) = P_m^{-1/2}[\tilde{A}_{m-1}(z) - z^{-1}\rho_m \tilde{B}_{m-1}(z)] \tag{6.4.23}$$

$$\tilde{B}_m(z) = Q_m^{-1/2}[z^{-1}\tilde{B}_{m-1}(z) - \rho_m^H \tilde{A}_{m-1}(z)] \tag{6.4.24}$$

在 $m=0$ 时，其初始条件为：

$$\tilde{A}_0(z) = \tilde{B}_0(z) = R_x^{-\frac{1}{2}}(0) \tag{6.4.25}$$

6.4.1.4 预测误差

$$\tilde{e}_{f,i}^{(m)} = \sum_{k=0}^{m} \tilde{A}_k^{(m)} x_{i-k} \tag{6.4.26}$$

$$\tilde{e}_{b,i}^{(m)} = \sum_{k=0}^{m} \tilde{B}_k^{(m)} x_{i-k} \tag{6.4.27}$$

仿照单变量最大熵谱，利用多维形式的最终预测误差准则来确定滤波器的最佳阶数：

$$FPE(m,l) = \left(\frac{n+lm+1}{n-lm-1}\right)^l \det\left[\frac{1}{2}(P_{m+1} + Q_{m+1})\right] \tag{6.4.28}$$

det[]表示广义方差。

6.4.2 计算步骤

归纳起来，多维最大熵谱估计计算步骤大致如下：

(1)首先利用(6.4.11)和(6.4.12)式计算0阶多维预测误差滤波器系数 $\tilde{A}_0^{(0)}$，$\tilde{B}_0^{(0)}$。用(6.4.26)和(6.4.27)式计算0阶向前和向后预测误差 $\tilde{e}_{f,l}^{(0)}$ 和 $\tilde{e}_{b,l}^{(0)}$，用(6.4.20)～(6.4.22)计算向前、向后及互预测误差功率 $\tilde{P}_{f,0}$，$\tilde{P}_{b,0}$、$\tilde{P}_{fb,0}$。

(2)将 $\tilde{P}_{f,0}$，$\tilde{P}_{b,0}$，$\tilde{P}_{fb,0}$ 代入下列归一化反射系数公式

$$\rho_{m+1} = \tilde{P}_{f,m}^{-1/2} \cdot \tilde{P}_{fb,m} \cdot \tilde{P}_{b,m}^{-H/2} \quad m = 0 \qquad (6.4.29)$$

得到 ρ_1，并将其代入(6.4.18)和(6.4.19)式，求出 P_1，Q_1，进而利用(6.4.28)式计算出 $FPE(0,l)$。

(3)利用(6.4.9)和(6.4.10)式计算 $k=1,2,\cdots,m-1$ 阶数的系数 $\tilde{A}_k^{(m)}$ 和 $\tilde{B}_k^{(m)}$，并重复过程(1)的其它计算，求出 $\tilde{e}_{f,l}^{(m)}$ 和 $\tilde{e}_{b,l}^{(m)}$ 及 $\tilde{P}_{f,m}$，$\tilde{P}_{b,m}$ 和 $\tilde{P}_{fb,m}$，进一步求出 ρ_{m+1}，P_{m+1} 和 Q_{m+1}。

(4)计算 $FPE(m,l)$，若 $FPE(m,l) \leqslant FPE(m-1,l)$ 则回到上一步计算过程，否则停止计算。

(5)将最终 m 阶的计算结果代入 z 变换形式(6.4.4)，得到归一化多维最大熵谱估计。

(6)与交叉谱类似，通过计算出的 P_m 和 Q_m，求出凝聚谱和位相谱。多维最大熵谱是复合谱，可以用实部与虚部形式表示：

$$S_{ij} = P_{ij} - iQ_{ij} \qquad (6.4.30)$$

凝聚谱为：

$$R_{ij}^2 = \frac{P_{ij} + Q_{ij}}{P_{ij}Q_{ij}} \qquad (6.4.31)$$

位相谱为：

$$\Theta_{ij} = \arctan \frac{Q_{ij}}{P_{ij}} \qquad (6.4.32)$$

为便于研究，通常是计算两序列的交叉最大熵谱，即计算 R_{12}^2 和 Q_{12}，分析两序列在某一频率上振动相关程度及两序列的相关状况。计算结果分析可以仿照交叉谱操作。

§6.5 奇异谱分析

奇异谱分析(Singular Spectrum Analysis，SSA)是从时间序列的动力重构出发，并与经验正交函数(Empirical Orthogonal Function，EOF)相联系的一种统计技术。它已广泛使用在时间范围上的信号处理中。SSA 的具体操作过程是，将一个样本量为 n 的时间序列 $x(t)$ 按给定的嵌套空间维数 m(称为窗口长度)构造一资料矩阵。当这一资料矩阵计算出明显成对的特征值，且相应的 EOF 几乎是周期性或正交时，通常就对应着信号中的振荡行为。可见，SSA 在数学上相应于 EOF 在延滞坐标上的表达，亦可以看作是 EOF 的一种特殊应用。分解的空间结构与时间尺度有关，可以有效地从一个有限的含有噪声的时间序列中提取信息。Broomhead 和 King 及 Freadrich 最先将 SSA 引入到非线性动力学研究，后来由 Vautard 和 Ghil[5~6]将其进行了一系列改进，并应用到研究气候序列的周期振荡现象中。SSA 的优点主要表现在两方面：一是它的滤波器不像通常的谱分析需要预先给定，而是根据资料自身最优确定。因此，它适合确定和寻找噪声系统中的弱信号。尤其它不需要作时间序列由不同频率正弦波叠加而成的假定，因而也就无需将一个本质上是非线性振荡的信号分解为大量正弦波之叠加来讨论。二是对嵌套空间维数 m 的

限定，可以使得对振荡的转换进行时间定位。SSA 是一种特别适合于识别隐含在气候序列中的弱信号，是一种研究周期振荡现象的新统计技术。

6.5.1 方法概述

给定一个间隔为 1 取样、样本量为 n_T、均值为 0 的时间序列 $x_i=x(t)$。再给定嵌套空间维数 m，$m\leqslant\frac{n_T}{2}$。那么，可将原时间序列 x_i 排列为 $m\times n$ 的资料矩阵：

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ x_2 & x_3 & \cdots & x_{n+1} \\ \vdots & \vdots & & \vdots \\ x_m & x_{m+1} & \cdots & x_{n_T} \end{bmatrix} \tag{6.5.1}$$

其中 $n=n_T-m+1$。(6.5.1)式阵的滞后自协方差则是一个 $m\times m$ 的矩阵，即

$$S_{ij}=\frac{1}{n}\sum_{t=1}^{n}x_t x_{t,j} \tag{6.5.2}$$

其中 j 为时间滞后步长，$j=1,2,\cdots,m$。显然，S_{ij} 为对称阵且主对角线为同一常数，称为 Toeplitz 矩阵。计算求出 S_{ij} 的特征值 λ_k 和相应的特征向量 φ_{kk}。特征值 λ_k 的开方值 σ_k 为奇异值。那么，滞后 j 的时间函数 $x(t+j)$ 的展开式表达为：

$$x(t+j)=\sum_{k=1}^{m}t_{kt}\varphi_{kj}\quad(t=1,2,\cdots,n) \tag{6.5.3}$$

由于资料阵是由嵌入时间滞后构成的，故 t_{kt} 为时间主分量，记为 T-PC，由下式求得：

$$t_{kt}=\sum_{j=1}^{m}x_{t+j}\varphi_{kj}\quad(t=1,2,\cdots,n) \tag{6.5.4}$$

φ_{kj} 称为时间经验正交函数，记为 T-EOF。和普通的 EOF 一样，T-PC 仍是时间 t 的函数。T-EOF 则是滞后时间步长的函

数，而不再是空间的函数。

棘手的问题是如何恰当地选取嵌套空间维数 m。从要求涵盖较多信息量的角度要选取大些的 m，从统计可信度考虑则 m 越小越好。通常视研究问题的时间尺度，选择适中的 m 为宜。

Vautard R. P 和 M Ghil 提出了几种分离振荡和噪音分量的方法。这里介绍其中一种。这种方法能够很好地从时间序列中分离出周期小于嵌套维数 m、谱宽小于 $1/m$ 的振荡。首先，估计特征值的误差

$$\delta\lambda_k = (2/n_d)^{1/2}\lambda_k \tag{6.5.5}$$

其中

$$n_d = (n_w/m) - 1 \tag{6.5.6}$$

式中 n_d 是给定窗口 m 的自由度个数。当一对 T--EOF 所对应的 λ_k，λ_{k+1} 满足

$$|\lambda_{k+1} - \lambda_k| \leqslant \min\{\delta\lambda_k, \delta\lambda_{k-1}\} \tag{6.5.7}$$

且这对 T-EOF 和 T-PC 是互相正交时，这一对 T-EOF 就代表系统的基本振荡。检验后者条件的办法是计算所给定的一对 T-PC 之间的滞后相关系数。如果存在很大的滞后相关系数，则表明对应的这对 T-PC 具有正交性。一般情况下，气候序列隐含的显著周期信号不止一个，因而有几对 T-PC 之间的滞后相关系数达到最大（通常大于 0.90）。滞后长度 j 作为推断显著振荡周期的依据。

6.5.2 计算步骤

用 SSA 提取气候时间序列基本周期的计算流程如下：

(1)将一维气候时间序列 $x(t)$ 按给定的嵌套空间维数构造形如(6.5.1)式的二维资料矩阵。

(2)计算资料矩阵的协方差矩阵 S_{ij}。

(3)利用Jacobi 方法求解协方差阵 S_{ij}的特征值 λ_k 和相应的特征向量。再利用(6.5.4)式求出时间主分量。

(4)将特征值按大小排序,并计算方差贡献:

$$\mathrm{Var}_k = \frac{\lambda_k}{\sum_{i=1}^{m}\lambda_i^2} \quad (k = 1,2,\cdots,m) \qquad (6.5.8)$$

(5)按(6.5.5)式计算各特征值的误差范围。若某对 *T-EOF* 所对应的特征值满足(6.5.8)式的条件,则进一步计算这对 *T-PC* 的滞后相关系数。将相关系数亦按大小排列,如果存在较大数值,表示这对 *T-PC* 代表系统的基本周期。

(6)计算周期。由于上述一对 *T-PC* 近于正交,j 滞后时间内相差 90°,一个周期 360°为 $4j$,将最大滞后相关系数对应的滞后时间长度 j 乘以 4,所得到的周期就是系统存在的显著周期。

(7)继续寻找其它对特征值 λ_k, λ_{k+1},若满足(6.5.7)式及 *T-PC* 互相正交条件,也视为统计的基本周期。但当满足条件的特征值出现在某个非线性动力系统的统计维数 S 之后就无需再考虑。当特征值 λ_k 随 k 的变化曲线斜率由明显的负值转化为近似 0 时,对应的 k 值就是统计维数 S。

计算过程中的一些具体问题,在应用实例中还会进一步说明。

6.5.3 计算结果分析

在气候研究中,SSA 主要用于对大气的年际和季节尺度的低频振荡进行分析。另外,它还可以对一维时间序列进行非线性吸引子的重建及对预报因子的信息压缩。SSA 用于气候诊断方面的用途可以作以下两方面分析:

(1)分析气候时间序列隐含的显著周期。这里需要强调的

是，显著周期长度与窗口长度 m 的选择密切相关。因此，选择恰当的 m 非常重要，它取决于讨论问题的时间尺度，用 SSA 研究长于窗口长度的周期是毫无意义的。

(2)分析前几个显著主分量所代表信号的趋势变化。

应用实例[6.4]：作为计算实例取 1952～1996 年赤道东太平洋(0～10°S，180～90°W)季平均标准化的海表温度资料作奇异谱分析。样本量为 180 个季，嵌套空间维数取 40。构造一新资料矩阵，用 Jacobi 方法求解特征值和特征向量，并计算出 T-PC。

图 6.3 为特征值随滞后长度 k 的变化曲线。由图可以看出，在 $k=17$ 时，特征值随 k 变化曲线斜率由负转为近似 0。因此，确定统计维数 $S=17$，亦即我们只需讨论 $k=17$ 之前的特征值。另外，图 6.3 还展现出存在三对 T-EOF 对应的特征值相近，即 T-EOF_1 和 T-EOF_2 对应的 λ_1 和 λ_2，T-EOF_3 和 T-EOF_4 对应的 λ_3 和 λ_4 及 T-EOF_5 和 T-EOF_6 对应的 λ_5 和 λ_6，这前 6 个 T-EOF 解释总方差的 63%。应该说，它们代表了赤道东太平洋海温的主要信息。

第一对 T-EOF(T-EOF_1，T-EOF_2)占总方差的 31%。图 6.4 为这对 T-EOF 的变化曲线。两个 T-EOF 曲线均呈现出十分明显的周期性。目测大约存在 16 个季的周期。T-EOF_1 和 T-EOF_2 变化趋势十分一致，T-EOF_2 比 T-EOF_1 滞后约 4 个季。特征值误差范围 $|\lambda_2-\lambda_1|=14.40$，$\min(\delta\lambda_1,\delta\lambda_2)=893.26$。因此，第一对特征值满足(6.5.7)式的条件。$T$-$PC_1$ 和 T-PC_2 之间的滞后相关中存在较大的相关系数。在滞后长度 $j=4$ 时，相关系数达最大为 0.94。次大相关系数为 0.89，出现在滞后 $j=3$ 时。表明 T-PC_1 和 T-PC_2 具有正交性。对应的周期为 $4\times j$，即存在长度为 $4\times4=16$ 个季(4 年)的主要

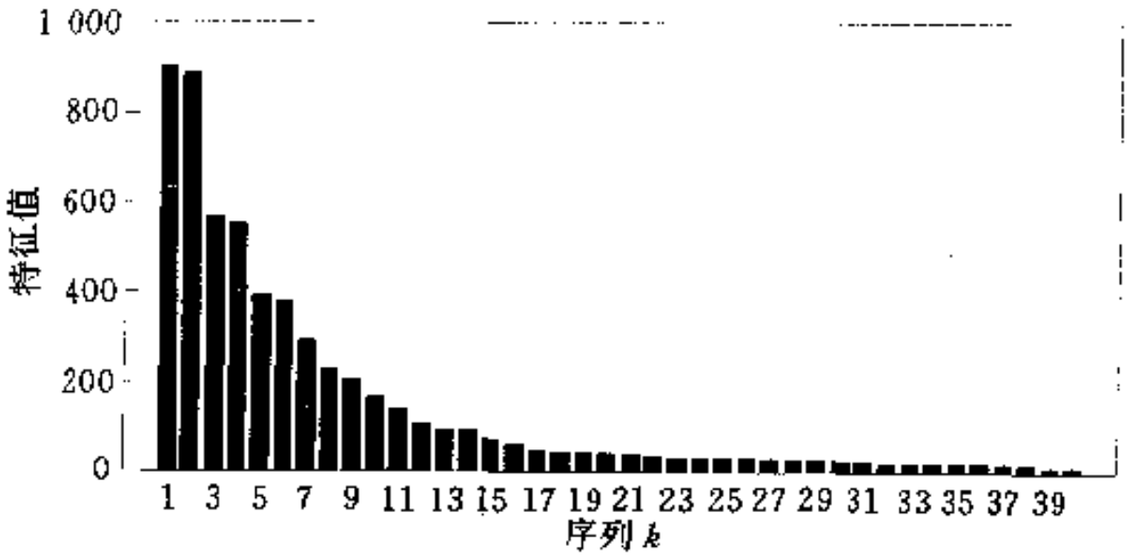


图 6.3　特征值 λ_k 随滞后长度 k 的变化曲线

周期，其次存在 12 个季(3 年)的周期。这种提取周期的方法所得到的结果与 $T\text{-}EOF_1$ 和 $T\text{-}EOF_2$ 曲线显示的周期是一致的。

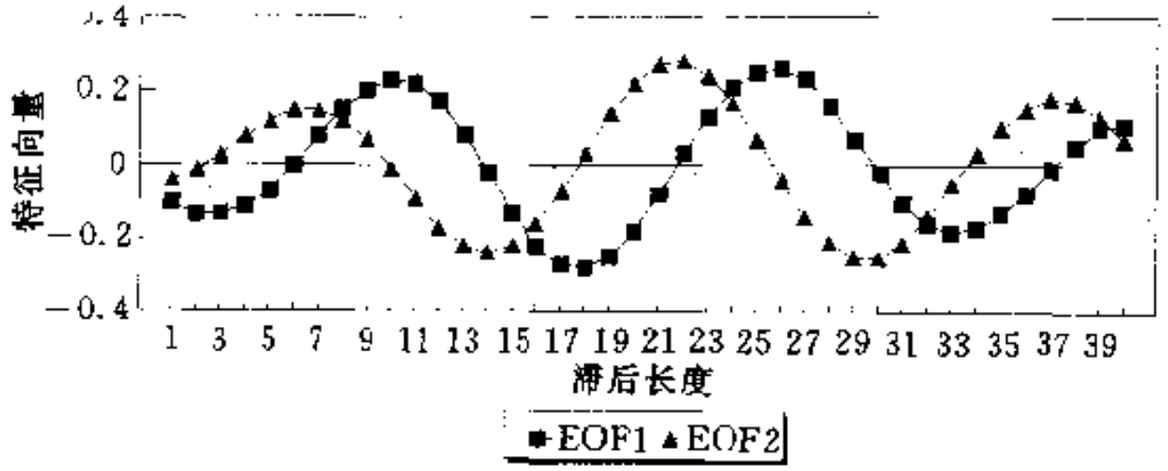


图 6.4　$T\text{-}EOF_1$ 和 $T\text{-}EOF_2$ 的变化曲线

第二对 $T\text{-}EOF$($T\text{-}EOF_3$ 和 $T\text{-}EOF_4$)占总方差的 19%。特征值误差范围 $|\lambda_4-\lambda_3|=14.70$，$\min(\delta\lambda_3,\delta\lambda_4)=552.66$，亦满足(6.5.7)式的条件。$T\text{-}PC_3$ 和 $T\text{-}PC_4$ 的滞后相关中亦有较大的相关系数。$j=4$ 时，相关系数为 0.95，$j=3$ 时为 0.91。表明这对 $T\text{-}PC$ 亦具正交性，且亦存在 16 个季(4 年)和 12 个季(3 年)的周期。

第三对 T-EOF(T-EOF_5 和 T-EOF_6)占总方差的 13%。特征值误差范围满足(6.5.7)式的条件。但是,T-PC_5 和 T-PC_6 的滞后相关系数没有较大数值,故认为不具正交性。

由 SSA 分析结果可知,赤道东太平洋海温存在 16 个季和 12 个季,即 3～4 年的显著周期。

§6.6 小波分析

小波分析(Wavelet Analysis)亦称多分辨分析(Multiresolution Analysis),是近几年国际上十分热门的一个前沿领域,被认为是傅里叶分析方法的突破性进展。1984 年法国地质学家 J. Morlet 在分析地震波的局部性质时,将小波概念引入到信号分析中。之后,理论物理学家 A. Grossman 和数学家 Y. Meyer 等人又对小波进行了一系列深入研究,使小波理论有了坚实的数学基础。进入 90 年代,小波分析成为众多学科共同关注的热点。在信号处理、图像处理、地震勘探、数字电路、物理学、应用数学、力学、光学等诸多科技领域得以广泛应用[7～8]。由于小波分析对信号处理的特殊优势,很快得到气象学家们的重视,将其应用于气象和气候序列的时频结构分析中,并已有不少引人注目的研究成果[9]。在气候诊断中,广泛使用的傅里叶变换可以显示出气候序列不同尺度的相对贡献,而小波变换不仅可以给出气候序列变化的尺度,还可以显现出变化的时间位置。后者对于气候预测是十分有用的[10～11]。需要指出的是,小波分析是一种基本数学手段,它可以应用在多种多样领域,可以从统计学角度研究,也可以应用在动力学乃至人工智能中。这里仅介绍用小波分析进行气候序列小波分解的具体方法及主要分析的内容。

6.6.1 方法概述

(1)小波分析的来源。经典的傅里叶分析的本质是将任意一个关于时间 t 的函数 $f(t)$变换到频域上，

$$F(w) = \int_R f(t)e^{iwt}dt \qquad (6.6.1)$$

其中 w 为频率；R 为实数域。$F(w)$确定了 $f(t)$在整个时间域上的频率特征。可见，经典的傅里叶分析是一种频域分析。对时间域上分辨不清的信号，通过频域分析便可以清晰地描述信号的频率特征。因此，从 1822 年傅里叶分析问世以来，得到十分广泛的应用。上面讲到的谱分析就是傅里叶分析方法。但是，经典的傅里叶变换有其固有缺陷，它几乎不能获取信号在任一时刻的频率特征。这里就存在时域与频域的局部化矛盾。在实际问题中，人们恰恰十分关心信号在局部范围内的特征。这就需要寻找时频分析方法。

1964 年 Gabor 引入了窗口傅里叶变换

$$\tilde{F}(w,b) = \frac{1}{\sqrt{2\pi}}\int_R f(t)\bar{\Psi}(t-b)e^{-iwt}dt \qquad (6.6.2)$$

其中函数 $\Psi(t)$是固定的，称为窗函数；$\bar{\Psi}(t)$是 $\Psi(t)$的复数共轭；b 是时间参数。由(6.6.2)式可知，为了达到时间域上的局部化，在基本变换函数之前乘上一个时间上有限的时限函数 $\Psi(t)$。这样 e^{-iwt}起到频限作用，$\Psi(t)$起到时限作用。随着时间 b 的变换，Ψ 确定的时间窗在 t 轴上移动，逐步对 $f(t)$进行变换。从(6.6.1)式中看出窗口傅里叶变换是一种窗口大小及形状均固定的时频局部分析，它能够提供整体上和任一局部时间内信号变化的强弱程度。像带通滤波就属于这类方法。由于窗口傅里叶变换的窗口大小及形状固定不变，因此局部化只是一次性的，不可能灵敏地反映信号的突变。事实上，反映

信号高频成分需用窄的时间窗，低频成分用宽的时间窗。在加窗傅里叶变换局部化思想基础上产生了窗口大小固定、形状可以改变的时频局部分析——小波分析。

(2)小波变换。若函数 $\Psi(t)$ 满足下列条件的任意函数

$$\int_R \Psi(t)\mathrm{d}t=0$$
$$\int_R \frac{|\hat{\Psi}(w)|^2}{|w|}\mathrm{d}w<\infty \tag{6.6.3}$$

其中 $\hat{\Psi}(w)$ 是 $\Psi(t)$ 的频谱。令

$$\Psi_{a,b}(t) = |a|^{-\frac{1}{2}}\Psi(\frac{t-b}{a}) \tag{6.6.4}$$

为连续小波，Ψ 叫基本小波或母小波，它是双窗函数，一个是时间窗，一个是频率谱。$\Psi_{a,b}(t)$ 的振荡随 $\frac{1}{|a|}$ 增大而增大。因此，a 是频率参数，b 是时间参数，表示波动在时间上的平移。那么，函数 $f(t)$ 小波变换的连续形式为：

$$w_f(a,b) = |a|^{-\frac{1}{2}}\int_R f(t)\overline{\Psi}(\frac{t-b}{a})\mathrm{d}t \tag{6.6.5}$$

由(6.6.5)式看到，小波变换函数是通过对母小波的伸缩和平移得到的。小波变换的离散形式为

$$w_f(a,b) = |a|^{-\frac{1}{2}}\Delta t\sum_{i=1}^{n} f(i\Delta t)\Psi(\frac{i\Delta t-b}{a}) \tag{6.6.6}$$

其中 Δt 为取样间隔；n 为样本量。离散化的小波变换构成标准正交系，从而扩充了实际应用的领域。

小波方差为：

$$\mathrm{Var}(a) = \sum (w_f)^2(a,b) \tag{6.6.7}$$

由连续小波变换下信号的基本特性证明，下面两个函数是母小波。

①Harr 小波：

$$\Psi(t)=\begin{cases}1 & 0\leqslant t<\frac{1}{2}\\ -1 & \frac{1}{2}\leqslant t<1\\ 0 & \text{其它}\end{cases}$$

②墨西哥帽状小波：

$$\Psi(t)=(1-t^2)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\quad -\infty<t<\infty$$

6.6.2 计算步骤

离散表达式的小波变换计算步骤如下：

(1)根据研究问题的时间尺度确定出频率参数 a 的初值和 a 增长的时间间隔。

(2)选定并计算母小波函数。

(3)将确定的频率 a，研究对象序列 $f(t)$ 及母小波函数 $\Psi(t)$ 代入(6.6.6)式，算出小波变换 $w_f(a,b)$，在编程计算 $w_f(a,b)$ 时，要做两重循环，一个是关于时间参数 b 的循环，另一个是关于频率参数 a 的循环。

6.6.3 计算结果分析

小波分析既保持了傅里叶分析的优点，又弥补了某些不足。原则上讲，过去使用傅里叶分析的地方，均可以由小波分析取代。从上面方法概述中可知，小波变换实际上是将一个一维信号在时间和频率两个方向上展开，这样就可以对气候系统的时频结构作细致的分析，提取有价值的信息。小波系数与时间和频率有关，因此，可以将小波变换结果绘制为二维图像。如图 6.5 所示，横坐标为时间参数 b，纵坐标为频率参数 a，图中数值为小波系数。这样将不同波长的结构进行了客观的分离，使波幅一目了然地展现在一张图上。当然，对结果的

分析还需凭借对所研究的系统的认识。根据作者个人的体会，对小波变换结果可以作以下几方面的分析：

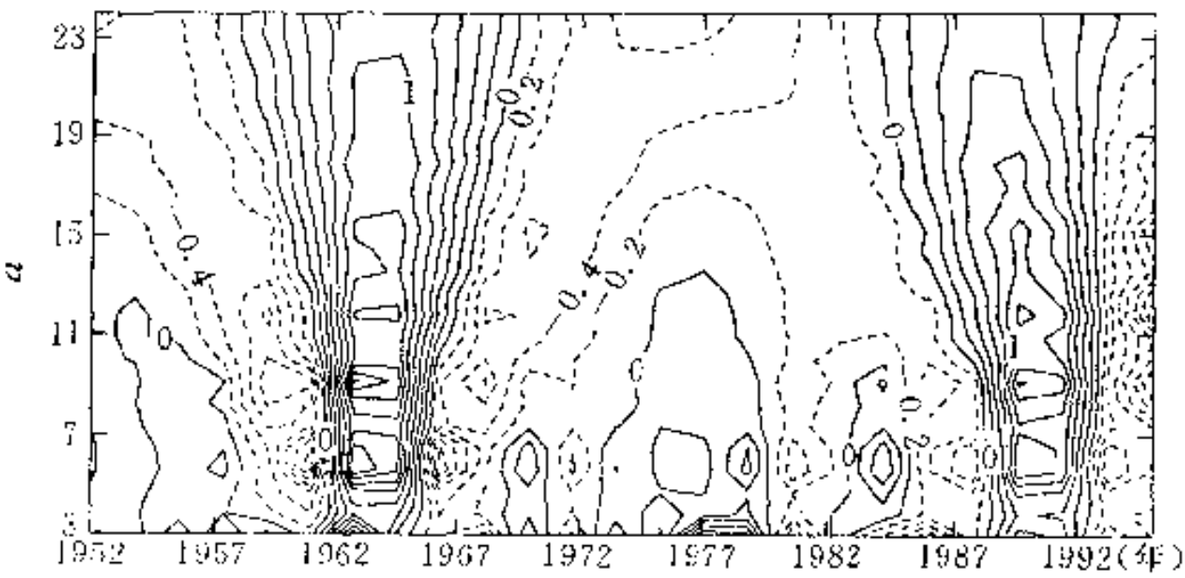


图 6.5　华北春季干旱指数小波变换

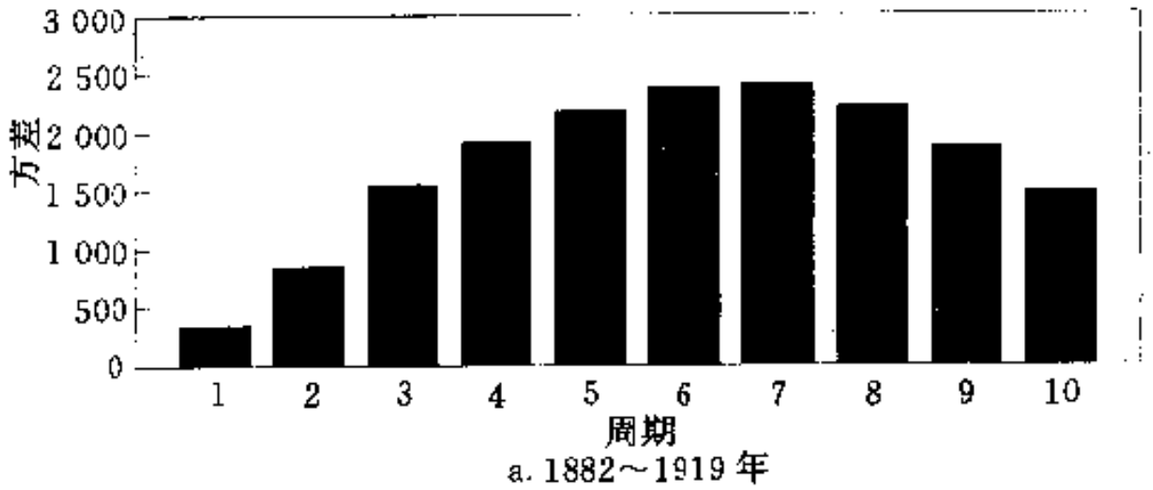
(1)利用分辨率是可调的这一特性，对我们感兴趣的细小部分进行了放大。从而可以十分细致地分析系统的局部结构、任一点附近的振荡特征，如分析某一波长振荡的强度等等。

(2)在小波系数呈现振荡之处分辨局地的奇异点，确定序列不同尺度变化的时间位置，提供突变信号，由此可以作序列的阶段性分析，并为气候预测提供信息。

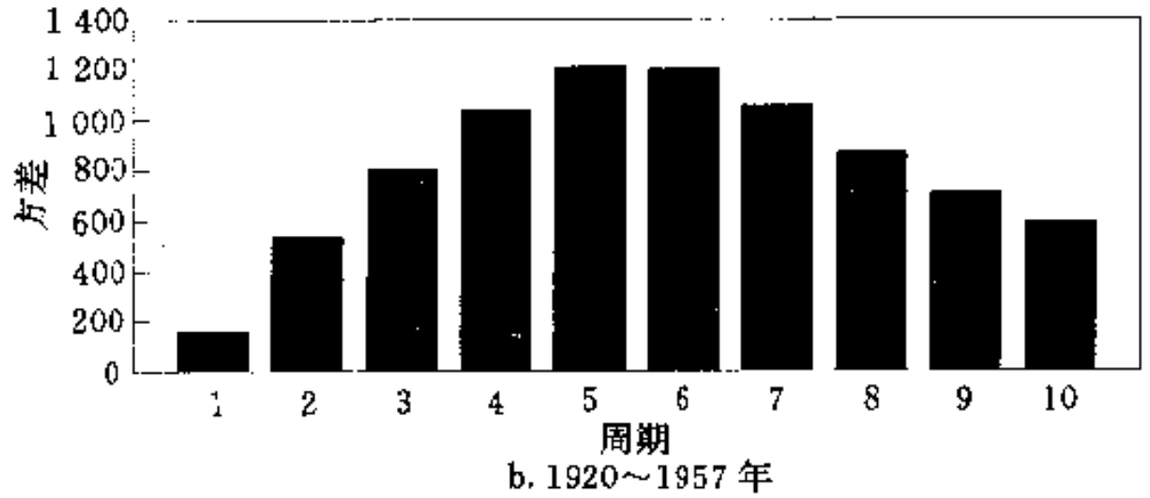
(3)从平面图上同时给出的不同长度的周期随时间的演变特征，认识不同尺度的扰动特性，由此判断序列存在的显著周期。

(4)利用小波方差可以更准确地诊断出多长周期的振动最强。另外，从分段的小波方差中推断某一时段内多长周期的振动最突出。

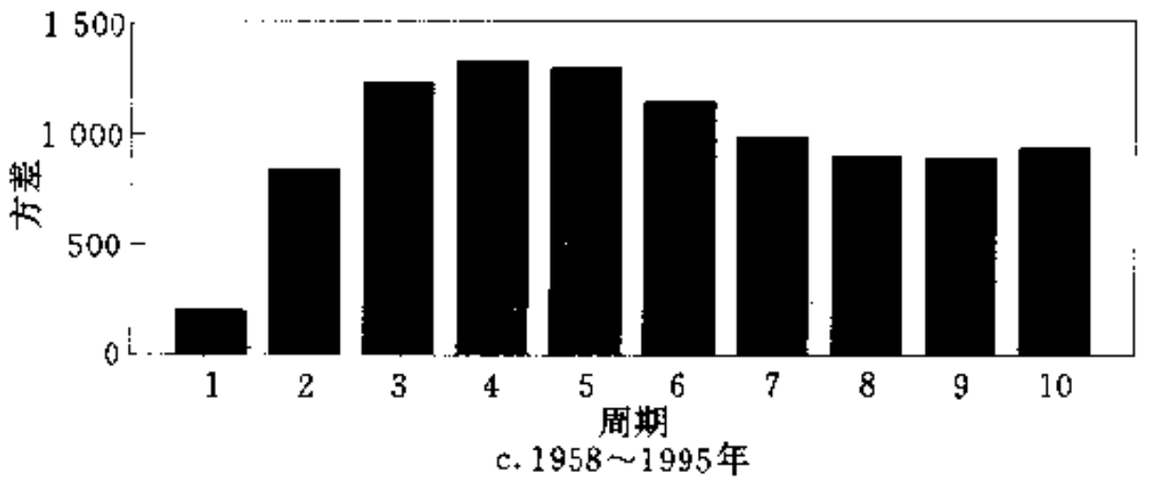
应用实例[6.5]：对 1952～1995 年华北春季干旱指数作小波分析[3]。这里 $n=44$，$a=3$，$b=24$。图 6.5 为小波变换平



a. 1882～1919 年



b. 1920～1957 年



c. 1958～1995年

图 6.6　南方涛动指数小波方差

面图。图的上半部分为低频，等值线相对稀疏，对应较长尺度周期的振荡。下半部分是高频，等值线相对密集，对应较短尺度周期的振荡。

从图 6.5 中呈现振荡之处可以分辨出奇异点，每个奇异点就是一次转折。在频率 $a=6$ 时的 1965 年处小波系数出现了最大值，表明 1965 年前后春季干旱指数发生了最强的振动。另外，图像呈现出明显的阶段性。就年代际尺度变化而言，1967～1986 年这 20 年华北春季干旱指数变化相对稳定，处在比较干旱时期。1966 年以前时段的变化结构与 1987 年以后时段的变化结构相似，变化均比较剧烈。

应用实例[6.6]：对 1882～1995 年 114 年南方涛动指数作小波变换。从小波变换呈现振动之处很容易分辨出厄尔尼诺与拉尼娜事件的转折点。厄尔尼诺与拉尼娜出现的周期振动是随时间变化的。在某一时段以某种周期为主，另一时段则另一长度周期占主导，从小波方差图（图 6.6）可以看得十分清楚。1882～1919 年 38 年中 7 年周期的振动最强；1920～1957 年的 38 年中 5 年周期振动最强；而最近的 38 年（1958～1995 年）则是 4 年周期振动最为突出。可见，近 30～40 年以来南方涛动的振荡比较频繁。

参考文献

[1]黄嘉佑，李黄．气象中的谱分析．北京：气象出版社，1984

[2]项静恬等．动态和静态数据处理——时间序列和数理统计分析．北京：气象出版社，1991. 579～588

[3]魏凤英．华北干旱不同时间尺度的变化特征．见：王馥棠等主编．华北农业干旱研究进展．北京：气象出版社，1997. 1～10

[4]海金，S. 茅于海等译．谱分析的非线性方法．北京：科学出版社，1986

[5] Vautard. SSA: A toolkit for noisy chaotic signals. physical, D58, 1992. 96—126

[6] Ghil and Vautard, Interdecadal oscillations and the warming trend in global temperature time series Nature, 1991 350: 324—327

[7] 崔锦泰[美]. 程正兴译. 小波分析导论. 西安: 西安交通大学出版社, 1994

[8] 秦前清, 杨宗凯. 实用小波分析. 西安: 西安电子科技大学出版社, 1994

[9] Hengyi Weng and K-M. Lau. Wavelets period doubling and time-frequency localization with application to organization of convection over the tropical western Pacific, Journal of the atmospheric sciences, 1994, 51(17): 2523—2541

[10] Arnedo A, Grasseau G, Holschneider M. Wavelet transform analysis of invariant measures of some dynamical system, Phys. Rev. Lett. 1988, 61: 2281

[11] Meyer Y et al. Wavelet and their application, Berlin: Springer-Verlag, 1992

第七章　气候变量场时空结构的分离

某一区域的气候变量场通常由许多个观测站点或网格点构成，这给直接研究其时空变化特征带来困难。如果能用个数较少的几个空间分布模态来描述原变量场，且又能基本涵盖原变量场的信息，是一个具有实用价值的工作，也就是寻找某种数学表达式将变量场的主要空间分布结构有效地分离出来。气候统计诊断中应用最为普遍的办法是把原变量场分解为正交函数的线性组合，构成为数很少的不相关典型模态，代替原始变量场，每个典型模态都含有尽量多的原始场的信息。其中经验正交函数(Empirical Orthogonal Function，EOF)分解技术就是这样一种方法。

EOF 最早是由统计学家 Pearson 在 1902 年提出来的[1]。50 年代中期 Lorenz 将其引入大气科学研究中。由于计算条件的限制，直至 70 年代初才在我国的气候研究领域中使用。70 年代中期以后，随着计算机技术的迅速发展。EOF 分解技术在气候诊断研究中得以充分应用。之所以被广泛使用，还由于它具有一系列突出的优点：第一，它没有固定的函数，不像有些分解需要以某种特殊函数为基函数，例如：球谐函数等。第二，它能在有限区域对不规则分布的站点进行分解。第三，它的展开收敛速度快，很容易将变量场的信息集中在几个模态上。第四，分离出的空间结构具有一定的物理意义。正因为如此，EOF 已成为气候科学研究中分析变量场特征的主要工具。以 EOF 为气候特征分析手段的研究成果十分丰硕，揭示出许多有价值的气候变化事实。

近10年来，气候统计诊断方法有了很大的进展，其中以EOF为基础的变量场分解方法的飞跃发展格外引人注目。针对气候变量场特征分析的需要，发展了揭示气象场空间结构和时间相关特征的扩展经验正交函数(Extended Empirical Orthogonal Function，EEOF)、着重表现空间的相关性分布结构的旋转经验正交函数(Rotated Empirical Orthogonal Function，REOF)、可以揭示空间行波结构的复经验正交函数(Complex Empirical Orthogonal Function，CEOF)和描述动力系统非线性变化特征的主振荡型(Principal Oscillation Patterns，POPs)。这些方法给气候统计诊断研究开阔了视野，使研究水平进入了一个更高层次。本章我们就以EOF为基础，介绍上述几种方法的特点、计算步骤及作者对计算结果分析的一些认识。

当然，EOF的应用范围远不止这一章所包含的内容。第六章叙述的奇异谱分析就是与EOF有联系的统计技术。尤其近年来，EOF分析方法在应用方面有十分迅速的发展。利用EOF是正交函数这一基本事实，发展了用EOF为基函数进行对强迫气候信号的检测和估计[2]、用于循环稳态型(Cyclostationary)气候时间序列信号的检测和估计[3]等技术。张邦林、丑纪范还提出了基于EOF的气候数值模拟及模式设计的新构思[4]。另外，EOF还被用来作为气候变量缺测资料插补的工具[5]。

§7.1 经验正交函数分解

经验正交函数(EOF)分解在数理统计学的多变量分析中称为主分量分析。是一种分解方法的两种提法。

由 m 个相互关联的变量，每个变量有 n 个样本构成矩阵形式 $X_{m\times n}$，对 X 进行线性变换，即由 p 个变量线性组合为一新变量：

$$Z_{p\times n} = A_{p\times m}X_{m\times n}$$

称 Z 为原变量的主分量，A 为线性变换矩阵。这一过程将原多个变量的大部分信息最大限度地集中到少数独立变量的主分量上。

将主分量分析在气候变量场上进行。将由 m 个空间点 n 次观测构成的变量 $X_{m\times n}$ 看作是 p 个空间特征向量和对应的时间权重系数的线性组合：

$$X_{m\times n} = V_{m\times p}T_{p\times n}$$

称 T 为时间系数，V 为空间特征向量。这一过程将变量场的主要信息集中由几个典型特征向量表现出来。

可见，主分量分析和经验正交函数分解是用两种形式推导出的同一方法。我们这里介绍的是在气候变量场上进行的经验正交函数分解。

7.1.1 方法概述

将某气候变量场的观测资料以矩阵形式给出

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mj} & \cdots & x_{mn} \end{bmatrix} \qquad (7.1.1)$$

其中 m 是空间点，它可以是观测站或网格点。n 是时间点，即观测次数。x_{ij} 表示在第 i 个测站或网格上的第 j 次观测值。

EOF 展开，就是将(7.1.1)式分解为空间函数和时间函

数两部分的乘积之和

$$x_{ij}=\sum_{k=1}^{m}v_{ik}t_{kj}=v_{i1}t_{1j}+v_{i2}t_{2j}+\cdots+v_{im}t_{mj} \quad (7.1.2)$$

写为矩阵形式为

$$X=VT \quad (7.1.3)$$

其中

$$V=\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mm} \end{bmatrix}$$

$$T=\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix}$$

分别称为空间函数矩阵和时间系数矩阵。根据正交性，V 和 T 应该满足下列条件

$$\begin{cases} \sum_{i=1}^{m}v_{ik}v_{il}=1 & \text{当 } k=l \text{ 时} \\ \sum_{j=1}^{n}t_{kj}t_{lj}=0 & \text{当 } k\neq l \text{ 时} \end{cases} \quad (7.1.4)$$

若 X 为距平资料矩阵，则可以对(7.1.3)式右乘 X'，即

$$XX'=VTX'=VTT'V' \quad (7.1.5)$$

XX' 是实对称阵。上标"′"表示矩阵转置。根据实对称分解定理一定有：

$$XX'=V\Lambda V' \quad (7.1.6)$$

其中 Λ 为 XX' 矩阵的特征值构成的对角阵。由(7.1.5)和

(7.1.6)式可知

$$TT' = \Lambda \tag{7.1.7}$$

由特征向量性质可知，$V'V$ 是单位矩阵，即满足(7.1.4)式的要求。可见，空间函数矩阵可以由 XX' 中的特征向量求出。V 得出后，即可得到时间系数

$$T' = V'X \tag{7.1.8}$$

当气候变量场的空间点数 m 大于样本量 n 时，采用所谓时空转换方案，可以减少许多计算机内存单元和计算时间。为叙述方便，这里暂且记 XX' 的特征向量为 V_N，记 $X'X$ 的特征向量为 V_R。根据特征向量的性质有：

$$X'XV_R = \Lambda V_R \tag{7.1.9}$$

对上式左乘 X 有：

$$XX'XV_R = \Lambda XV_R \tag{7.1.10}$$

记为

$$V = XV_R \tag{7.1.11}$$

则 V 为矩阵 XX' 的特征向量，有：

$$XX'V = \Lambda V \tag{7.1.12}$$

说明 $X'X$ 与 XX' 具有相同的非零特征值。但是，V 不是标准化的，它的模是：

$$V'V = V_R X'XV_R = V'_R \Lambda V_R = \Lambda \tag{7.1.13}$$

并不满足 $V'V=1$。因此，标准化的特征向量 V_N 为：

$$V_N = \frac{1}{\sqrt{\Lambda}}V \tag{7.1.14}$$

可以证明 $V'_N V_N = 1$。

可见，时空转换就是先求出 $X'X$ 的特征值和特征向量，借此求出 XX' 阵的特征向量。

7.1.2　计算步骤

EOF 的一般计算步骤如下：

(1)对原始资料矩阵 X 作距平或标准化处理。然后计算其协方差矩阵 $S=XX'$，S 是 $m\times m$ 的实对称阵。

(2)用求实对称矩阵的特征值及特征向量方法(最常用的是 Jacobi 方法)求出 S 阵的特征值 Λ 和特征向量 V。

(3)矩阵 Λ 为对角阵，对角元素即为 XX'的特征值 $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_m)$。将特征值按大小排列为：

$$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant 0$$

(4)利用(7.1.8)式求出时间系数矩阵 T。

(5)计算每个特征向量的方差贡献：

$$R_k=\lambda_k\Big/\sum_{i=1}^{m}\lambda_i\quad[k=1,2,\cdots,p(p<m] \tag{7.1.15}$$

及前 p 个特征向量的累积方差贡献：

$$G=\sum_{i=1}^{P}\lambda_i\Big/\sum_{i=1}^{m}\lambda_i\quad(p<m) \tag{7.1.16}$$

如果空间点数大于样本量，则用 EOF 的时空转换过程计算：

(1)对原始资料矩阵 X 作预处理后，计算协方差矩阵 $S=X'X$。

(2)求出 S 阵的特征值和特征向量 V_R。

(3)利用(7.1.11)和(7.1.14)式求出特征向量 V_N，即 XX'的特征向量。

(4)与一般 EOF 步骤 3～5 计算相同。

7.1.3　显著性检验

分解出的经验正交函数究竟是有物理意义的信号还是毫无意义的噪音，应该进行显著性检验，特别是当变量场空间点数 m 大于样本量时，显著性检验尤其重要。这一点常常被忽

视。目前，常用的检验方法有：

(1)特征值误差范围，用 North 等人提出的计算特征值误差范围来进行显著性检验[6]。特征值 λ_j 的误差范围为：

$$e_j = \lambda_j\left(\frac{2}{n}\right)^{\frac{1}{2}}$$

n 为样本量。当相邻的特征值 λ_{j+1} 满足

$$\lambda_j - \lambda_{j+1} \geqslant e_j$$

时，就认为这两个特征值所对应的经验正交函数是有价值的信号。

(2)Monte Carlo 技术。Preisendorfer R. W 和 T. P. Barnett 最早将 Monte Carlo 技术用于经验正交函数的显著性检验[7]。首先按(7.1.15)式计算观测变量场特征值的方差贡献 $R_k, k=1,2,\cdots,p$。利用随机数发生器产生高斯分布的随机序列资料矩阵，矩阵也由 m 个空间点，n 个样本量构成。对这一矩阵进行模拟经验正交函数计算，对模拟计算的特征值 δ_k 排序。这样的过程共重复 100 次，每次亦计算方差贡献

$$U_k^r = \delta_k^r \Big/ \sum_{i=1}^{m} \delta_i^r$$

$$(k = 1,2,\cdots,p \quad r = 1,2,\cdots,100)$$

将 U_k^r 排序

$$U_k^1 \leqslant U_k^2 \leqslant \cdots \leqslant U_k^{100} \quad (k = 1,2,\cdots,p)$$

如果

$$R_k > U_k^{95}$$

就认为第 k 个特征向量在 95%置信水平下具有统计显著性，有分析价值。

7.1.4 变量场资料的预处理

EOF 分解实际上就是求矩阵 XX' 的特征值和特征向量

过程。求 XX' 时，使用变量场 X 的数据形式不同，得到的结果就不同。

变量场无外乎三种形式——原始变量场、变量的距平场和变量的标准化场。当用原始场计算时，XX' 就是原数据交叉乘积，得到的第一特征向量代表了平均状况，其权重很大。对于不存在季节变化的变量场来说，它的分解结果物理意义直观。作者在制作全国汛期降水预报时，用全国 6～8 月降水总量作 EOF 分解，分离出的特征向量十分典型，物理意义十分清楚，在下面有关预报方法的章节中将作详细介绍。但是，对于以分析变量场特征为主要目的的研究，所用的变量场大多存在季节变化，平稳性很差，造成经验正交函数不稳定。当用距平场计算时，XX' 是协方差矩阵，从分析的意义来讲，分离出的特征向量的气象学意义比较直观，经验正交函数在一定时效内具有稳定性。当用标准化场计算时，XX' 是相关系数矩阵，分离出的特征向量代表的是变量场的相关分布状况，更适合作分类分型分析。由此可见，在使用 EOF 分解时，可以根据需要，采用不同的资料形式，但对特征向量所代表的物理含义应该有明确的认识。

另外，对某一区域的变量场进行 EOF 展开时，选择观测站点要注意其均匀性，以免造成结果失真[8]。

7.1.5　计算结果分析

凭借气候学知识对前几项有意义的特征向量及所对应的时间系数作分析。

(1)通过显著性检验的前几项特征向量最大限度地表征了某一区域气候变量场的变率分布结构。它们所代表的空间分布型式是该变量场典型的分布结构。如果特征向量的各分量均为同一符号的数，那么这一特征向量所反映的是该区域

变量变化趋势基本一致的特征，数值绝对值较大处则为中心。如果某一特征向量的分量呈正、负相间的分布型式，这一特征向量则代表了两种分布类型。图 7.1a 是用 1951～1996 年中国 160 个站夏季(6～8 月)降水量作 EOF 展开的第二特征向量。由图可看出，江淮流域大范围为正值，黄河流域及华南地区为负值。这一特征向量代表江淮流域降水趋势与黄河流域华南地区为相反的两种分布型式，即江淮流域降水多，黄河流域及华南降水少的分布型式或江淮流域降水少，黄河流域及华南降水多的分布型式。

(2)特征向量所对应的时间系数代表了这一区域由特征向量所表征的分布型式的时间变化特征。系数数值绝对值越大，表明这一时刻(月、年等)这类分布型式越典型。例如：图 7.1a 特征向量所对应的时间系数序列(图 7.1b)代表的是中国夏季降水年际趋势变化。某年的时间系数为正值，则代表该年呈江淮流域降水偏多，黄河流域和华南地区降水偏少的分布型式。若时间系数为负值，则表明该年呈相反的降水分布型式。系数绝对值越大，这类分布型式就越显著。

(3)从特征值的方差贡献和累积方差贡献了解所分析的特征向量的方差占总方差的比例及前几项特征向量总共占总方差的比例。

§7.2 扩展经验正交函数分解

用经验正交函数分解我们可以得到气候变量场空间上的分布结构，是固定时间形式的空间分布结构，它不能得到扰动的时间上移动的空间分布结构。然而，气候变量场在时间上存在显著的自相关及交叉相关。扩展的经验正交函数(EEOF)

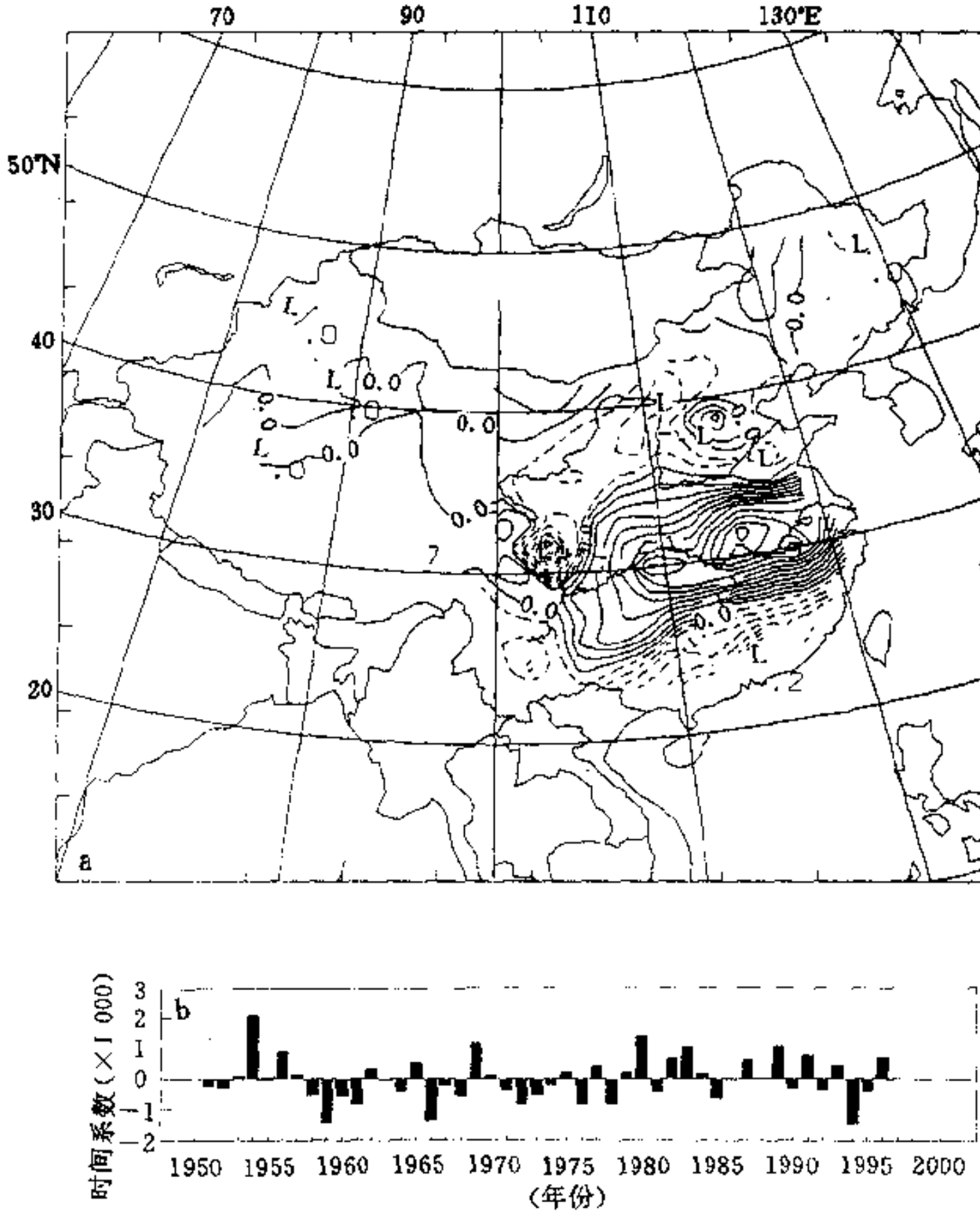


图 7.1　中国夏季降水量的第二特征向量(a)及时间系数(b)

充分利用了变量场时间上的这种联系，因而可以得到变量场的移动性分布结构。这一方法是 1982 年 Weare 和 Nasstrom 提出的[9]。

EEOF 的基本方法与 EOF 相似。这里主要介绍计算步骤。关键是计算协方差矩阵的资料矩阵是由几个连续的时间上的观测值构成的，相当于构造一个比 EOF 扩大了几倍的资料矩阵，因而计算机容量要求大，收敛速度也较慢。

7.2.1 计算步骤

EEOF 的计算步骤如下：

7.2.1.1 构造资料矩阵

对于一个有 m 个空间点数、时间取样为 n 的变量场，首先建立一个新的资料矩阵。例如：建立滞后 2 个时次的资料矩阵，它的形式为

$$\underset{3m\times(n-2j)}{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n-2j} \\ \vdots & \vdots & & \vdots \\ x_{mj} & x_{m2} & \cdots & x_{mn\ \ 2j} \\ x_{1j+1} & x_{1j+2} & \cdots & x_{1n-j} \\ \vdots & \vdots & & \vdots \\ x_{mj-1} & x_{mj-2} & \cdots & x_{mn-j} \\ x_{12j+1} & x_{12j+2} & \cdots & x_{1n} \\ \vdots & \vdots & & \vdots \\ x_{m2j+1} & x_{m2j+2} & \cdots & x_{mn} \end{bmatrix} \tag{7.2.1}$$

由(7.2.1)式可见，新的资料矩阵由原时刻资料矩阵、滞后一个时次和滞后二个时次的资料矩阵构成。式中 j 为滞后时间长度。j 的选取视研究的具体问题而定。例如：欲研究气候变量场准两年振荡，那么 j 取为 4 个月，这时资料矩阵由滞后 0，4 和 8 个月构成。

7.2.1.2　计算协方差矩阵

计算资料阵(7.2.1)式的协方差矩阵 S，这时 S 是 $3m\times 3m$ 阶的实对称矩阵。

7.2.1.3　求解特征值和特征向量

用 Jacobi 方法求出 S 阵的特征值 λ 和特征向量 V。这时有 $3m$ 个特征值和 $3m$ 个特征向量。尤其要注意的是，每个特征向量又均包括 $3m$ 个空间点。例如：得到的第一特征向量是

$$V_1=(V_1,V_2,\cdots,V_m,V_{m+1},V_{m+2},\cdots,V_{2m},V_{2m+1},\cdots,V_{3m}) \tag{7.2.2}$$

$V_1,V_2,\cdots,V_m$ 是滞后 0 时次的特征向量，$V_{m+1},V_{m+2},\cdots,V_{2m}$ 是滞后 1 个时次的特征向量，$V_{2m+1},V_{2m+2},\cdots,V_{3m}$ 是滞后 2 个时次的特征向量。可见，一个特征向量包括三个时次的空间分布结构。另外，应该注意：这时的特征向量正交性是某一特征向量三个时次的特征向量之和与另一特征向量三个时次特征向量之和的正交。

7.2.1.4　计算时间系数

与 EOF 一样，用(7.1.8)式计算时间系数矩阵 T。得到的 T 矩阵是 $3m$ 行，$n-2j$ 列。应该注意每个特征向量的时间系数所对应的时刻。前 m 个特征向量的时间系数对应的时刻是 $1\sim n-2j$；$m+1,\cdots,2m$ 个特征向量的时间系数对应的时刻是 $j+1\sim n-j$；$2m+1,\cdots,3m$ 个特征向量的时间系数对应的时刻是 $2j+1\sim n$。

7.2.1.5　计算方差贡献和累积方差贡献

用(7.1.15)和(7.1.16)式计算特征向量的方差贡献和累积方差贡献。在计算两式分母部分时要注意，这里是计算 $3m$ 个特征值之和。

7.2.2 Batterworth 带通滤波

在气候诊断分析中，常常用 EEOF 来作某气候变量场的准周期振荡演变特征分析。这时，需要预先将变量场的特定周期分量分离出来，然后用分离后的资料再作 EEOF 分解，可以用对每一测站或网格点的数据进行一阶 Batterworth 带通滤波来实现。

设某一测站或格点观测值序列为 $x_0, x_2, \cdots, x_{n-1}$，带通滤波计算公式为：

$$y_k = a(x_k - x_{k-2}) - b_1 y_{k-1} - b_2 y_{k-2} \qquad (7.2.3)$$

由(7.2.3)式可见，在做 k 时刻滤波时用到过去前 2 个时刻的数据和 k 时刻以前 2 个时刻的滤波结果。问题归纳为确定系数 a, b_1 和 b_2。

首先，需要先给出三个频率：ω_0, ω_1 和 ω_2。ω_0 是带通滤波器的中心频率，ω_1 和 ω_2 是 ω_0 两边的两个频率值。系数 a, b_1 和 b_2 由下式计算

$$a = \frac{2\Delta Q}{4 + 2\Delta Q + Q_0^2}$$

$$b_1 = \frac{2(Q_0^2 - 4)}{4 + 2\Delta Q + Q_0^2}$$

$$b_2 = \frac{4 - 2\Delta Q + Q_0^2}{4 + 2\Delta Q - Q_0^2}$$

其中

$$\Delta Q = 2\left|\frac{\sin\omega_1\Delta T}{1 + \cos\omega_1\Delta T} - \frac{\sin\omega_2\Delta T}{1 + \cos\omega_2\Delta T}\right|$$

$$Q_0^2 = \frac{4\sin\omega_1\Delta T\sin\omega_2\Delta T}{(1 + \cos\omega_1\Delta T)(1 + \cos\omega_2\Delta T)}$$

ΔT 为取样时间间隔。

三个频率 ω_0,ω_1 和 ω_2 根据具体滤波的周期长度来确定。例如:要研究准 3.5 年的周期振荡,那么就需要作 30～60 个月的带通滤波。这时取 $\omega_1=\frac{2\pi}{30}$,$\omega_2=\frac{2\pi}{60}$,$\omega_0=\sqrt{\omega_1\times\omega_2}$或取 $\omega_0=\frac{2\pi}{40}$,$\omega_1=\frac{2\pi}{30}$,$\omega_2=\frac{\omega_0^2}{\omega_1}$。

7.2.3 计算结果分析

(1)如果计算的是滞后 2 个时次的 EEOF,那么一个特征向量就得到 3 张空间分布结构图。根据这些图可以分析空间系统的移动方向、强度变化等特征。这些变化特征是一般 EOF 得不到的。但是,遇到本身时间的持续性较差的变量场时,得到的空间分布结构往往难以解释。

(2)根据特征向量对应的时间系数可以分析准周期的振幅变化及不同滞后长度之间振幅的位相差。

应用实例(7.1):为了研究中国降水的准 3.5 年周期各位相的演变特征[10],取中国东部 90 个站各月降水量距平先作各站 30～60 个月的带通滤波,然后建立滞后 2 个时次的资料矩阵。滞后时间长度 j 取为 5 个月,这样使一个完整的循环接近 3.5 年。降水准 3.5 年周期振荡的 EEOF 展开收敛很快,前两个特征向量已解释总方差的 99.3%。前两个特征向量的时间系数变化振幅相近;只是第一特征向量时间系数超前第二特征向量时间系数约 10 个月左右位相。第一特征向量滞后 10 个月的分布结构与第二特征向量滞后 0 个月的分布结构相似,第二特征向量滞后 10 个月的分布结构与第一特征向量滞后 0 个月的分布结构相似,说明第一特征向量和第二特征向量可以共同描述 3.5 年周期中不同位相降水的异常分布。

§7.3 旋转经验正交函数分解

从7.1节的介绍中看到，EOF展开得到的前几个特征向量，可以最大限度地表征气候变量场整个区域的变率结构。但是，EOF也有其局限性，即分离出的空间分布结构不能清晰表示不同地理区域的特征。另外，进行EOF展开时，所取区域范围不同，例如：取整个区域和分块区域，得到的特征向量空间分布图形亦会不同，这就给进行物理解释带来困难。再者，计算EOF取样大小不同，对反映真实分布结构的相似度也会有不同，即存在一定的取样误差。EOF上述的局限性，使用旋转经验正交函数(REOF)可以得到克服。其实，REOF分解并不是新分析方法，它与因子分析中的旋转主因子分析无本质区别。只是近年来将其用于变量场的分析越来越多。旋转后的典型空间分布结构清晰，不但可以较好地反映不同地域的变化，还可以反映不同地域的相关分布状况。REOF比EOF在取样误差上也小得多。因此，REOF愈来愈受到人们的重视，且成为分离变量场典型空间结构的一种新倾向。

7.3.1 方法概述

REOF与因子分析中旋转主因子分析是一种方法的两种提法。这里暂且用因子分析替代EOF，二者只略有差异。对于一标准化的(注意：这里一定是标准化的)含有 m 个变量，n 次观测样本的资料阵 $X_{m\times n}$ 可以表示为公共因子矩阵 $T^*_{p\times n}$ ($p<m$)和因子荷载阵 $V_{p\times p}$ 的乘积及特殊因子 $U_{m\times n}$ 之和的形式

$$X = VT^* + U \tag{7.3.1}$$

特殊因子仅与 X 有关，它与 EOF 的差别仅在于此。若忽略了 U，则与EOF一致。

公共因子是标准化变量，各公共因子均是均值为 0、方差为 1 的独立变量。公共因子之间协方差阵为单位矩阵，即

$$T^*T^{*\prime} = I \tag{7.3.2}$$

在 EOF 中，时间系数矩阵 T 满足

$$TT' = \Lambda \tag{7.3.3}$$

Λ 为相关矩阵 XX' 的特征值。因而，我们得到

$$T^* = \Lambda^{-\frac{1}{2}}T \tag{7.3.4}$$

由 EOF 可知，

$$T = V'X \tag{7.3.5}$$

因此，

$$T^* = \Lambda^{-\frac{1}{2}}V'X \tag{7.3.6}$$

从而，

$$X = V\ \Lambda^{\frac{1}{2}}\ \Lambda^{-\frac{1}{2}}T \tag{7.3.7}$$

如果把 p 个公共因子看成由 p 个因子空间构成的坐标基，因子荷载就视为 p 个变量在这个坐标基上的投影。公共因子坐标轴的旋转过程就是作线性变换的过程。新的公共因子坐标基表示为：

$$\overline{T} = GT^* \tag{7.3.8}$$

其中 G 为线性变换矩阵，原因子荷载阵 V 可通过线性变换矩阵 A 变为新的因子荷载阵 $\overline{V}$

$$\overline{V} = VA \tag{7.3.9}$$

如果 $\overline{V}$ 和 $\overline{T}$ 满足(7.3.7)式，则有

$$X = \overline{VT} \tag{7.3.10}$$

类似，相关阵有

$$R = \overline{VTT'V'} \tag{7.3.11}$$

用(7.3.8)和(7.3.9)式代入(7.3.11)式得，

$$R = V(AG)(AG)'V' \qquad (7.3.12)$$

若令

$$(AG)(AG)' = I \qquad (7.3.13)$$

则满足旋转前(7.3.7)式变量相关阵的结构。

在因子轴转动过程中，要求矩阵(AG)必须是正交的。但不一定要求 A 和 G 阵正交。如果要求新因子轴也是正交，则要求

$$\overline{TT}' = GT^*T^{*\prime}G' = GG' = I \qquad (7.3.14)$$

即要求 G 是正交，这样导致 A 阵也要是正交的，则

$$(AG)(AG)' = AGG'A' = AA' = I \qquad (7.3.15)$$

这时可简单取 $G=A^{-1}$。如果不要求新因子轴是正交的，即

$$\overline{TT}' \neq I$$

这种旋转称为仿射旋转或斜交旋转。

现在的问题是如何实现因子坐标轴的旋转。一般分为正交旋转与斜交旋转两种方式。极大方差旋转是正交旋转，是气候诊断分析中最常使用的旋转方法。这种方法的实质，是将各因子轴旋转到某个位置，使每个变量在旋转后的因子轴上极大、极小两极分化，从而使高荷载只出现在少数变量上，即在旋转因子矩阵中，少数变量有高荷载，其余均接近 0。使因子荷载矩阵结构简化，满足了旋转因子轴"简单结构解"的要求。从变量场的角度解释，经过极大方差旋转，使分离出的典型空间模态上只有某一较小区域上有高荷载，其余区域均接近 0，使得空间结构简化、清晰。

新因子上的因子荷载阵 $\overline{V}$，其元素为 $\overline{v}_{ij}$，欲使新因子上少数变量有高荷载，而同时其余接近 0，就要使新的因子荷载元素的方差

$$S^2 = \frac{m\sum_{j=1}^{p}\sum_{i=1}^{m}(\bar{v}_{ij}^2/h_i^2)^2 - \sum_{j=1}^{p}(\sum_{i=1}^{m}(\bar{v}_{ij}^2/h_i^2)^2}{m^2} \tag{7.3.16}$$

达到极大。其中

$$h_i^2 = \sum_{j=1}^{p}\bar{v}_{ij}^2 \tag{7.3.17}$$

表示第 i 个变量由公共因子解释的方差。为了使(7.3.16)式达到极大，连续使用因子轴的转动角的三角函数变换矩阵来极大化方差。每次从要旋转的 p 个因子中选两个进行正交旋转，使它们的因子荷载满足(7.3.16)式的判据。再用其中一个新因子与另外一个原因子进行旋转，满足(7.3.16)式判据，这样共进行 $p(p-1)/2$ 次旋转，就完成了一次旋转循环，重新进行循环，直至所有要旋转的因子对均满足(7.3.16)式判据为止。

对于第 k 个及第 q 个新因子的荷载应满足

$$\begin{aligned}\bar{v}_{ik} &= v_{ik}\cos\theta + v_{iq}\sin\theta \\ \bar{v}_{iq} &= -v_{ik}\sin\theta + v_{iq}\cos\theta\end{aligned} \tag{7.3.18}$$

将(7.3.18)式代入(7.3.16)式，并令 $\frac{\partial S}{\partial\theta}=0$，即

$$\tan 4\theta = \frac{2\sum_{i=1}^{m}u_i w_i - 2\sum_{i=1}^{m}u_i\sum_{i=1}^{m}w_i/m}{\sum_{i=1}^{m}(u_i^2-\omega_i^2) - \left[(\sum_{i=1}^{m}u_i)^2 - (\sum_{i=1}^{m}\omega_i)^2\right]/m} \tag{7.3.19}$$

其中

$$\begin{aligned}u_i &= (v_{ik}^2 - v_{iq}^2)/h_i^2 \\ \omega_i &= 2v_{ik}v_{iq}/h_i^2\end{aligned} \tag{7.3.20}$$

由上述公式计算出 θ 角，进行一次旋转。

7.3.2 计算步骤

对于一个变量场的 REOF 具体计算步骤，可以简单地以流程图(图 7.2)表示。

7.3.3 旋转经验正交函数个数的确定

旋转经验正交函数个数 p，可以用下述办法确定：

(1)由经验正交函数的累积方差贡献来确定。一般可取累积方差贡献达 85%为标准来确定旋转特征向量的个数 p。方差贡献百分率根据具体问题适当增减。

(2)通过特征值对数曲线变化来确定旋转特征向量的个数。如果某个特征值之后的直线的斜率明显变小，即以该点特征值的个数作为旋转特征向量的个数 p。

(3)用 North 特征值误差范围来确定旋转特征向量的个数 p。这一方法已在 7.1 节经验正交函数的显著性检验中介绍。

7.3.4 计算结果分析

REOF 计算结果的物理含义与 EOF 有所不同，可以作以下几方面的分析：

(1)REOF 得到的空间模态是旋转因子荷载向量。因此，每个向量代表的是空间相关性分布结构。经历了旋转过程，高荷载集中在某一较小区域上，其余大片区域的荷载接近 0。如果某一向量的各分量符号均一致，则它代表了这一区域的气候变量变化一致，且以高荷载地域为中心的空间分布结构。如果某一向量在某一区域的分量符号为正，而在另一区域的分量符号为负，高荷载集中在正区域或负区域，它代表了这两区域变化趋势相反，且以高荷载所在区域为中心的分布结构。通过空间分布结构，不仅可以分析气候变量场的地域结构，也可

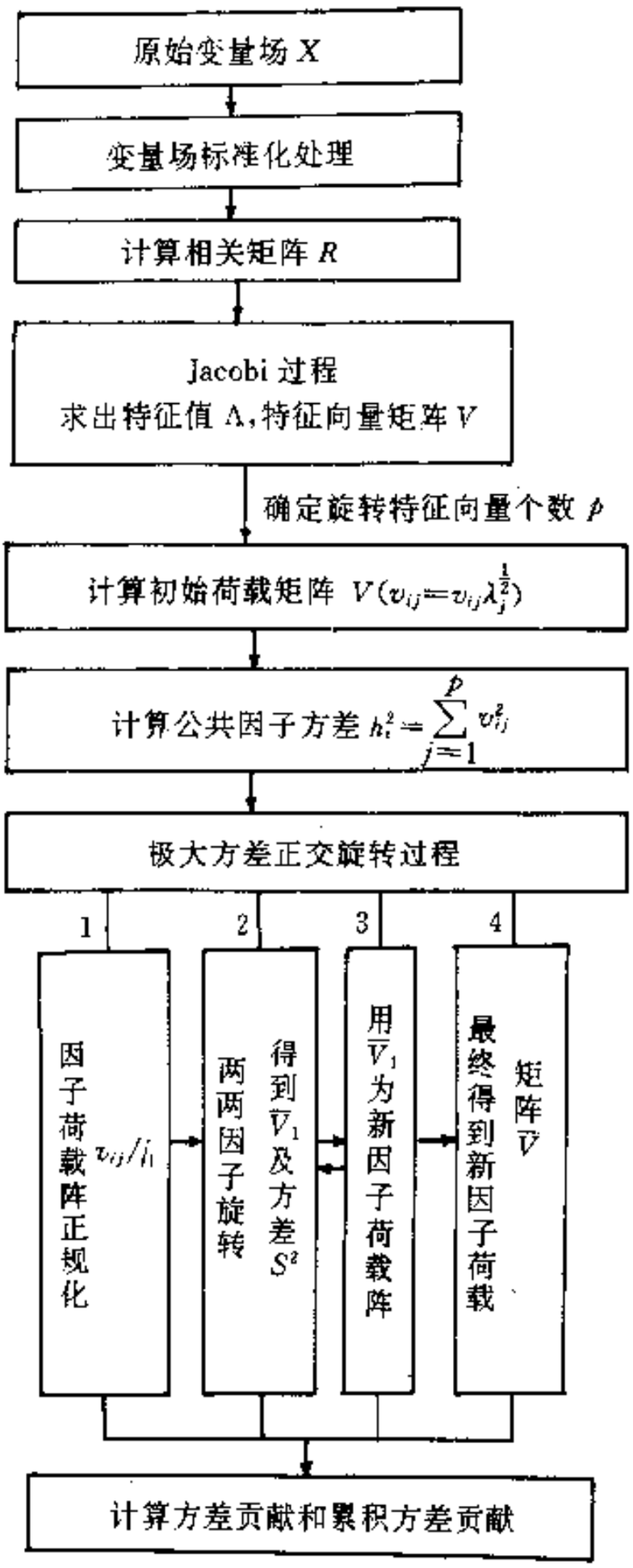


图 7.2　REOF 计算流程图

以通过各向量的高荷载区域对气候变量场进行区域和类型的划分等研究。

(2)通过旋转空间模态对应的时间系数，可以分析相关性分布结构随时间的演变特征。时间系数的绝对值越大，表明这一时刻(年、月等)的这种分布结构越典型，极大值中心亦越明显。

(3)旋转后方差贡献要比 EOF 均匀分散。通过它们可以了解旋转的特征向量解释总方差的比例。

应用实例[7.2]：取青藏高原 30 个测站 1961～1995 年冬季(11～3 月)积雪日数作 REOF，分析这一区域冬季积雪的地域分布结构及时间变化特征。首先对资料矩阵作标准化处理。作 EOF 分析后，根据对数特征值图确定旋转前 5 项特征向量，计算荷载向量矩阵，进一步作方差极大正交旋转。图 7.3a 为第一旋转空间模，它占总方差的 22.5%。这一空间模展示了青藏高原冬季积雪的一个明显中心，位置在高原东北部巴颜喀拉山区。这是青藏高原冬季积雪的一个典型的分布结构。从这一空间模对应的时间系数距平变化曲线(图 7.3b)看出，这种分布结构的冬季积雪具有明显的年际振荡和年代际变化特征。

§7.4 复经验正交函数分解

1981 年 Rasmusson 等人将复经验正交函数(CEOF)分解的经典思想引进气象研究中[11]。之后，Barnett 将其精炼并作了进一步发展[12]。目前流行使用的计算格式大都是由 Barnett 针对气象变量场分析特点给出的。CEOF 是一种能够分离出气候变量场空间尺度行波分布结构及位相变化的方法。

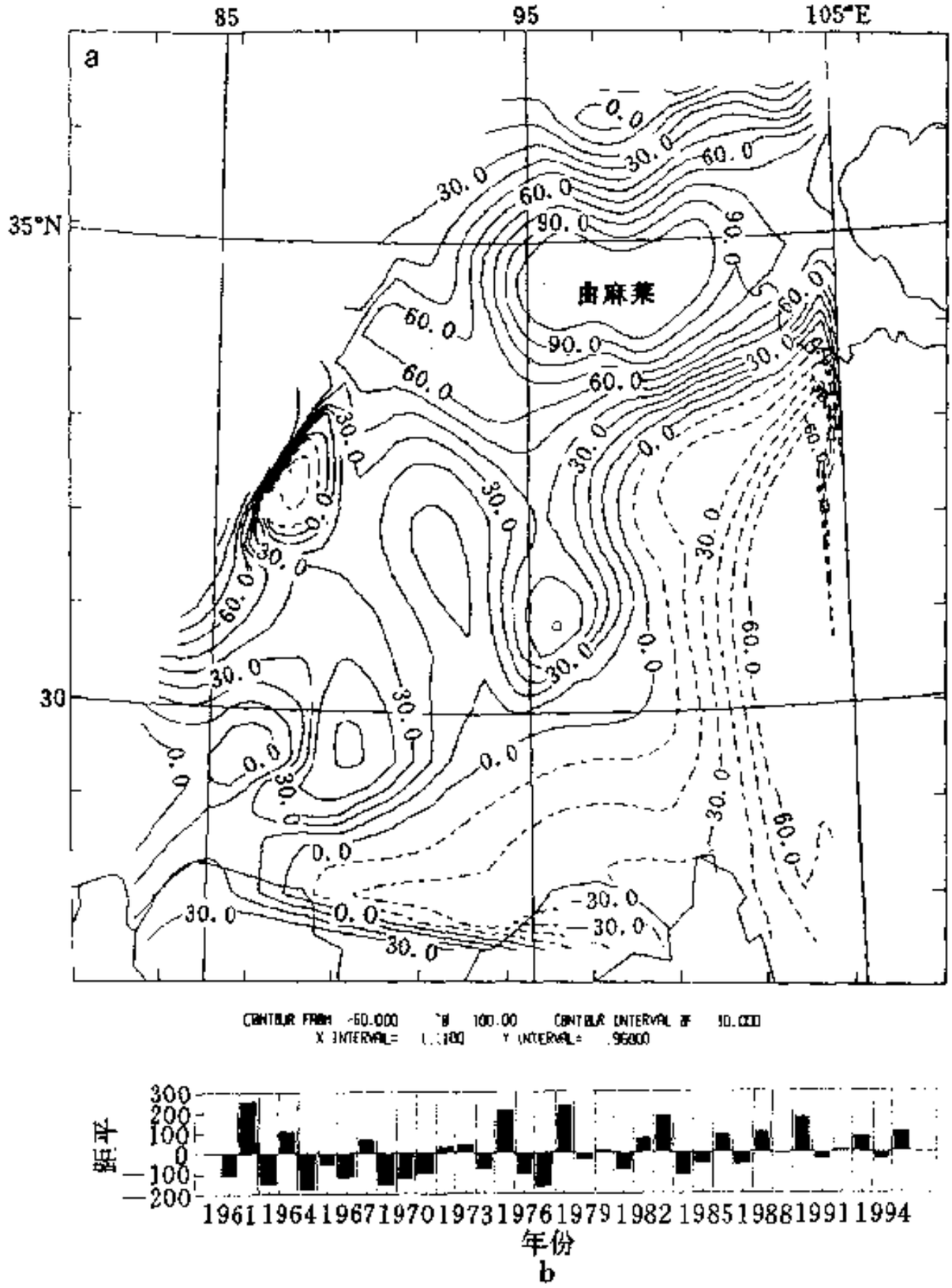


图 7.3 青藏高原积雪日数的第一旋转空间模(a)及其时间系数距平变化曲线(b)

传统的 EOF 分离出的仅是空间驻波振动分布结构。CEOF 的这一特殊功效为气候变量场的诊断研究提供了更有效的工具，揭示出一些用原有方法不曾得到的有价值的信息。CEOF 实质上是将一个标量场通过变换，构造一个同时含有实部和虚部的复数矩阵——Hermite 矩阵进行分解。

7.4.1 Hermite 矩阵及其构造方法

CEOF 是在构成的 Hermite 复数矩阵基础上进行的。这里有必要先对 Hermite 复数矩阵的有关基础知识及变换构成方法进行复习，以便更清楚地了解 CEOF 方法。

7.4.1.1 复数

复数又称虚数，不是计数和测量时使用的初等意义的数。复数单位为 i，i 满足下列关系式：

$$\begin{cases} i^2 = (-i)^2 = -1 \\ i = \sqrt{-1} \\ -i = -\sqrt{-1} \end{cases}$$

每个复数可以表示为实数 a 与一个虚数 ib 之和：

$$C = a + ib$$

其中实数 $a=\mathrm{Re}(C)$ 和 $b=\mathrm{Im}(c)$ 称为复数 C 的实部和虚部。

7.4.1.2 共轭复数

两个有相同实数部分和相反虚数部分的复数

$$\begin{cases} C = a + ib \\ C^* = a - ib \end{cases}$$

称为共轭复数。

7.4.1.3 Hermite 矩阵

由复数元素构成的矩阵称为复数矩阵。若满足

$$(U^*)' = U \text{ 或 } \overline{U} = U$$

则称 U 为 Hermite 矩阵。其中“ * ”表示复数共轭,“'”表示转置,“—”表示转置共轭。

例如:

$$U=\begin{bmatrix} 5 & 1-2i \\ 1+2i & 1 \end{bmatrix}\text{其共轭阵}\ U^{*}=\begin{bmatrix} 5 & 1+2i \\ 1-2i & 1 \end{bmatrix}$$

$$(U^{*})'=\begin{bmatrix} 5 & 1-2i \\ 1+2i & 1 \end{bmatrix}=U$$

7.4.1.4 酉矩阵

若 $\bar{U}U=I$ 且 $U\bar{U}=I$,则称 U 为一个酉矩阵或复正交阵。

例如:

$$U=\begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}\qquad \bar{U}=\begin{bmatrix} 0 & -i \\ -i & 0 \end{bmatrix}$$

$$\bar{U}U=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}=I$$

7.4.1.5 Hermite 矩阵的特征值和特征向量

若 U 为 $m\times m$ 阶的 Hermite 矩阵时,其特征值 Λ 为实数,且存在一酉矩阵 B,使得

$$\bar{B}UB=\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}$$

B 为 Hermite 矩阵的特征向量。

7.4.1.6 构造 Hermite 阵方法

CEOF 过程是在一个复数矩阵上进行的。因此,预先要将一个标量(实数)序列 $u_j(t)$ 变换为一个同时含有实部 $u_j(t)$ 和

虚部 $\hat{u}_j(t)$ 的复数序列。构造复数序列的虚部方法常用的有两种：

(1)滤波。生成一个与实数序列 $u_j(t)$ 相正交的序列

$$\hat{u}_j(t) = \sum_{l=-L}^{L} u_j(t-l)h(l)$$

其中 L 为滤波器长度，一般 L 取 7～25，L 取得太大会损失过多信息，$h(l)$ 为权重系数

$$h(l) = \begin{cases} \dfrac{2}{l\pi}\sin^2\left(\dfrac{\pi l}{2}\right) & l \neq 0 \\ 0 & l = 0 \end{cases}$$

可以证明，这一变换相当于 $\dfrac{\pi}{2}$ 位相差的滤波过程。

(2)傅里叶变换。原实数序列 $u_j(t)$ 可以分解为傅里叶形式

$$u_j(t) = \sum_{\infty} [a_j(\omega)\cos\omega t + b_j(\omega)\sin\omega t]$$

这里

$$a_j(\omega) = \frac{1}{T}\int_0^T u_j(t)\cos\omega t\,dt$$

$$b_j(\omega) = \frac{1}{T}\int_0^T u_j(t)\sin\omega t\,dt$$

式中 $\omega = \dfrac{2\pi}{T}$ 为圆频率；T 为周期；$a_j(\omega)$；$b_j(\omega)$ 为傅里叶系数。可以生成一个与 $u_j(t)$ 相正交的序列

$$\hat{u}_j(t) = \sum_{\infty} [a_j(\omega)\cos(\omega t + 90°) + b_j(\omega)\sin(\omega t + 90°)] = \sum_{\infty} [b_j(\omega)\cos\omega t - a_j(\omega)\sin\omega t]$$

变换出虚部后，就可以构造出 Hermite 复数矩阵。

7.4.2 方法概述

设有一实资料序列 $u_j(t)$，构造其复资料序列

$$U_j(t) = u_j(t) + i\hat{u}_j(t) \qquad (7.4.1)$$

对于 m 个空间点，n 个观测样本的气候变量场的复资料阵为

$$U_{m\times n}=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{bmatrix} \tag{7.4.2}$$

分解复资料阵

$$U = BP \tag{7.4.3}$$

其中 B 为 $m\times m$ 阶复空间函数矩阵，P 为 $m\times n$ 阶复时间函数矩阵。

根据 Hermite 矩阵特征值和特征向量性质可知，复空间函数矩阵由复协方差阵 $U\overline{U}$ 不同特征值 λ_k 的特征向量构成，$\overline{U}$ 为 U 的转置共轭。复时间函数阵

$$P = \overline{B}U \tag{7.4.4}$$

在复空间函数和时间函数基础上求出表征振荡和移动特征的空间振幅函数 $S_k(x)$，空间位相函数 $Q_k(x)$，时间振幅函数 $S_k(t)$ 和时间位相函数 $Q_k(t)$。

$$S_k(x) = [B_k(x)B_k^*(x)]^{1/2} \tag{7.4.5}$$

$$Q_k(x) = \arctan\left[\frac{\mathrm{Im}B_k(x)}{\mathrm{Re}B_k(x)}\right] \tag{7.4.6}$$

$$S_k(t) = [P_k(t)P_k^*(t)]^{1/2} \tag{7.4.7}$$

$$Q_k(t) = \arctan\left[\frac{\mathrm{Im}P_k(x)}{\mathrm{Re}P_k(t)}\right] \tag{7.4.8}$$

式中 x 表示空间点；t 为时间点；k 为特征向量序号；$B_k(x)$ 表示第 k 个特征值对应的特征向量；$B_k^*(x)$ 是 $B_k(x)$ 的共轭向量；$P_k^*(t)$ 表示 $P_k(t)$ 的共轭。

7.4.3 计算步骤

CEOF 的计算步骤为：

(1)用滤波或傅里叶变换方法建立一实数矩阵的 Hermite 复数矩阵。

(2)计算 Hermite 复数矩阵的协方差矩阵 $S=U\overline{U}$，S 也为 Hermite 复数矩阵。

(3)根据 Hermite 矩阵的分解定理，计算出特征值 λ_i 及对应的特征向量 B。特征值 λ_i 为实数，特征向量为复数矩阵。

(4)用(7.4.4)式计算复时间系数矩阵 P。

(5)用(7.4.5)～(7.4.8)式计算有关移动特性的 4 个量。

(6)用(7.1.15)和(7.1.16)式计算特征向量的方差贡献及累计方差贡献。

7.4.4 计算结果分析

对于 CEOF 的计算结果，主要分析在复空间函数和复时间函数基础上得到的表示移动特征的 4 个振幅和位相函数。对于某一气候变量场的移动特性要从振幅和位相两方面考虑。因此，合理地解释 CEOF 的结果并非易事。

(1)通过空间振幅函数，分析气候变量场的空间分布结构，寻找变化强度的中心。根据空间位相函数分析波的传播方向。

(2)时间振幅函数反映变化强度随时间的变化，由时间位相函数分析波的传播速度。

应用实例[7.3]：选取 1850～1991 年我国东部 25 个站旱涝百分率资料作 CEOF 分析[13]。将 25 个站 142 年旱涝百分率资料矩阵用滤波变换扩充到酉空间，建立复数矩阵进行分解。其中滤波器长度 L 取为 7。利用(7.4.5)～(7.4.8)式计算出空间振幅函数 $S_k(x)$，空间位相函数 $Q_k(x)$，时间振幅函数 $S_k(t)$和时间位相函数 $Q_k(t)$。前 3 个复特征向量描述了总方差变化的 68%。相同资料的普通 EOF，前 3 个特征向量仅占

总方差的 36%。可见，中国东部近百年旱涝场中波动特征非常明显。图 7.4 表示的是第 2 个空间模的空间传播特征。空间振幅有两个高值中心(图 7.4a)，一个在黄河流域，另一个在江南。该模的空间位相(图 7.4b)，从北向南具有显著变化。黄河流域与长江以南位相差为 180°。表明这个模态反映的旱涝异常信息具有明显的传播特征。

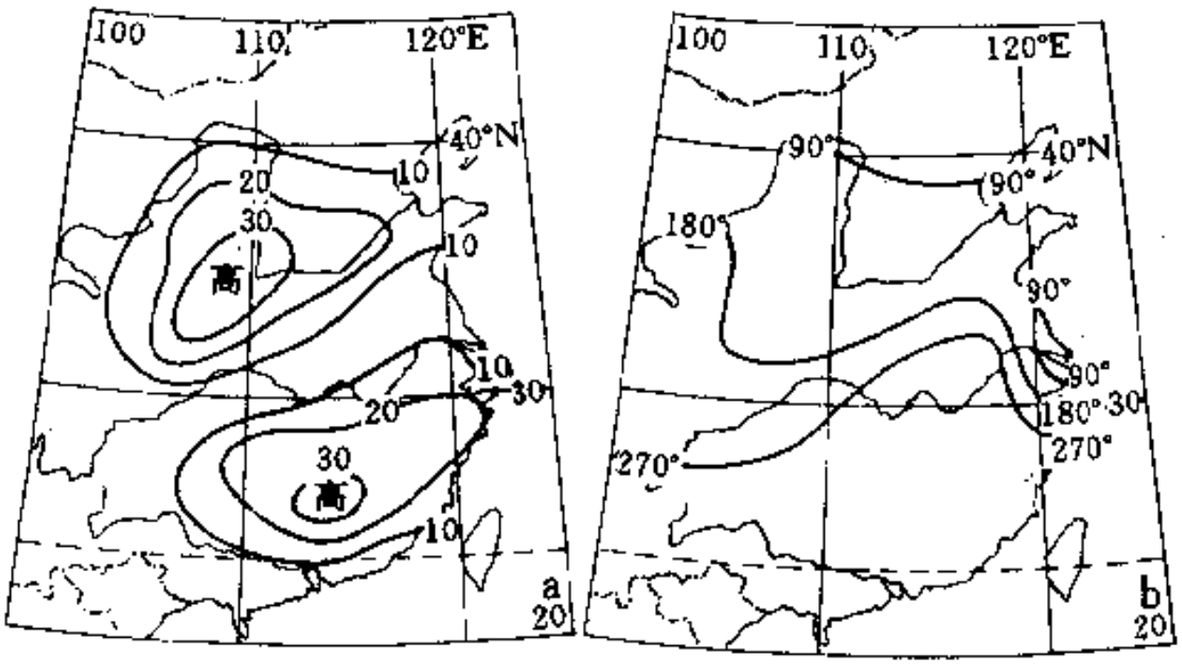


图 7.4　中国东部旱涝的模 2 空间振幅(a)和空间位相(b)

§7.5　主振荡型分析

主振荡型(POPs)技术的基本思想是 Hasselmann 在 1988 年提出来的[14]。其基本概念是由主相互作用型(Principal Interaction Patterns，PIPs)推导出来的，与此同时，Storch 等人由一个线性系统的标准模态来定义 POP，对这一新技术进行了系列卓有成效的完善，并在气候系统的低频变化、QBO 及 ENSO 的诊断和预测研究中加以应用[15~16]。后来，

Storch 等人专门对 POP 的概念、应用及拓展方法作了详尽的综述[27]。应用实例证明，POP 是一种识别复杂气候系统时空变化特征的多变量分析新技术。

POP 分析最本质的论点认为，气候系统的主要过程是由一阶马尔科夫过程所描述的线性动力过程。其它次要过程被认为是一种随机噪声强迫。具体实施时，用 EOF 展开一个气候变量场进行截断，提取主要过程，用自回归滑动平均技术构造系统动力模型。因此，从这个角度上讲，POP 分析技术亦可以看作是常规 EOF 和自回归方法的联合和拓展。我们知道，EOFs 在给定时间内可以生成变量场协方差结构的一个最优表达式，但却不能揭露系统的时间演变结构或其内部动力过程。7.2 节中讲到的用同一变量场不同滞后时刻构造协方差阵的 EEOF，可以提供时间演变的空间结构，但它不能与谱结构相联合。对于上述功能 POP 兼而有之。7.4 节讲的 CEOF 在功能上与 POP 类似。二者最主要的差别在于，CEOF 是在解释方差最大和相互正交的约束条件下构造的，所分析的时空移动特征不是由 CEOF 直接给出的，而是由时间系数推导计算出来的。POP 与动力方程相联系，得到的时空演变特征是计算的直接输出。

7.5.1 方法概述

这里叙述的是 Storch 等人由一个线性系统的标准模态推导出来的 POP 方法。

7.5.1.1 EOF 展开

设有 m 个测站的一个系统 $X(r,t)$，r 是空间函数，$1 \leqslant r \leqslant m$，$t$ 是时间函数。那么，这个系统就有 m 个标准模态。对于一个大的系统，这些标准模态就代表大量的资料，而其重要特征很难提取出来。因此，将资料转换为 EOF 空间。EOFs 定义

为 $X(r,t)$协方差阵的特征向量。如果用 $V_i(r)$代表 EOFs，那么资料矩阵 $X(r,t)$可以展开为：

$$X(r,t) = \sum_{i=1}^{k} Z_i(t)V_i(r) \qquad (7.5.1)$$

这里 $Z_i(t)$是时间系数。

我们将 EOF 空间截断为 k 维($k \leqslant m$)。这样做的目的是希望将我们感兴趣的那部分信号保留下来，舍掉噪声成分。经过压缩后有：

$$X(r,t) = \sum_{i=1}^{k} Z_i(t)V_i(r) \qquad (7.5.2)$$

7.5.1.2　主振荡型(POP)

设 $X(r,t)$第 t 时刻的列向量 $x(t)$为一个线性离散化的实数系统，其标准模态

$$x(t+1) = Ax(t) \qquad (7.5.3)$$

是变换矩阵 A 的特征向量 V。通常情况下，A 是非对称的常数矩阵，其全部或某些特征值 λ 和特征向量 V 是复数，即有共轭复数特征值 $\lambda^r + i\lambda^i$ 及特征向量 $V^r + iV^i$。由于 A 是一实矩阵，共轭复数 λ^* 和 V^* 亦满足

$$AV^* = \lambda^* V^* \qquad (7.5.4)$$

任何时刻 t 的状态 X 均可以用特征向量来表示：

$$X = \sum_{j=1}^{k} Z_j V_j \qquad (7.5.5)$$

共轭复数特征向量对的系数也是共轭复数。将(7.5.5)式代入(7.5.3)式中，耦合的系统(7.5.3)式变为不耦合的，产生出 k 个单一方程，其中 k 是过程 x 的维数

$$Z(t+1)V = \lambda Z(t)V \qquad (7.5.6)$$

如果 $Z(0)=1$，则有

$$Z(t)V = \lambda^t V \tag{7.5.7}$$

复数共轭对 V, V^* 对过程 $X(t)$ 的贡献为

$$P(t) = Z(t)V + [Z(t)V]^* \tag{7.5.8}$$

记 $V=V^r+iV^i$，$2Z(t)=Z^r(t)-iZ^i(t)$，(7.5.8)式可以写为

$$\begin{aligned} P(t)&=Z^r(t)V^r+Z^i(t)V^i= \\ &\rho^t[\cos(\omega t)V^r-\sin(\omega t)V^i] \end{aligned} \tag{7.5.9}$$

系统的振荡及传播模型是由(7.5.9)式来描述的。这时特征值 $\lambda=\rho\exp(-i\omega)$，且取 $Z(0)=1$。

振动模态表现为特征向量 V^r 和 V^i 之间，以下面顺序交替出现

$$\cdots \to V^r \to -V^i \to -V^r \to V^i \to V^r \to \cdots \tag{7.5.10}$$

称特征向量 V 为主振荡型(POP)。实的特征向量 V^r 称为实POP，它描述系统的驻波振荡。复的特征向量 V^i 称为复POP，它描述系统振荡的传播。系数 Z 称为POP系数，其时间演变由(7.5.6)式给出。

由上述过程可以得到两个重要参数：

①振荡周期。

$$T = \frac{2\pi}{\omega} \tag{7.5.11}$$

它是完成(7.5.10)式一个完整循环所需的时间。其中 $\omega=\arctan^{-1}|\lambda^i/\lambda^r|$。

②振幅 e-折度时间.

$$\tau = -1/\ln(\rho) \tag{7.5.12}$$

它是初始振幅由 $|Z(0)|=1$ 降至 $|Z(\tau)|=\frac{1}{e}$ 所需要的时间。

7.5.1.3　变换矩阵 A 的估计

矩阵 A 可以由下式得到

$$A = S_1 S_0^{-1} \tag{7.5.13}$$

其中 S_0 和 S_1 分别是滞后时刻为 0 和 1 的协方差矩阵

$$\begin{cases} S_0 = \langle x(t)x'(t) \rangle \\ S_1 = \langle x(t+1)x'(t) \rangle \end{cases} \tag{7.5.14}$$

其中〈　〉表示求平均。

在动力学理论中，(7.5.3)式为离散化线性差分方程。在 POP 分析中，关系式

$$x(t+1) = Ax(t) - \text{noise} \tag{7.5.15}$$

是给定的。这时，变换矩阵 A 则可以用在(7.5.15)式右边乘以 $x'(t)$，并取数学期望的方法得到，即

$$A = \frac{\mathrm{E}[x(t+1)x'(t)]}{\mathrm{E}[x(t)x'(t)]} \tag{7.5.16}$$

在计算方法上，(7.5.16)式与(7.5.13)式是一样的。

7.5.1.4　半随相关型

在描述气候系统 $x(t)$ 的主振荡过程时，在另一系统 $y(t)$ 中存在与之相伴随的振荡过程，其表达式为：

$$y(t) = Z(t)q(t) - Z^*(t)q^*(t) + e(t) \tag{7.5.17}$$

其中 $Z(t)$ 是 $x(t)$ 主振荡型时间系数；$Z^*(t)$ 是 $Z(t)$ 的复数共轭；$Z^*(t) = Z^r(t) - iZ^i(t)$；$q(t)$，$q^*(t)$ 是复伴随相关型；$e(t)$ 是由 $q(t)$，$q^*(t)$ 描述 $y(t)$ 产生的误差。

如果在实际应用时仅关心含有一个单独变量复数共轭 POP 对时，复伴随相关型 q 就可以简化为由下式求得：

$$q = \frac{\langle |Z(t)|^2 \rangle \langle Z^*(t)y(t) \rangle \quad \langle Z^{*2}(t) \rangle \langle Z(t)y(t) \rangle}{\langle Z^*(t)Z(t) \rangle^2 - \langle Z^2(t) \rangle \langle Z^{*2}(t) \rangle} \tag{7.5.18}$$

7.5.2 计算步骤

POP 的计算存在一定的难度，且因研究目的的不同而有所不同。下面给出用 POP 作气候诊断的最基本步骤。

(1)一般情况下，对原始气候变量场预先进行 EOF，截取占 80%以上方差的前 k 个 EOF 的时间系数序列 $x(t)$进行 POP。

(2)如果将资料中所要分析的信号预先确定在某一频率段内，则应对 $x(t)$进行带通滤波。

(3)用(7.5.13)和(7.5.14)式计算系数矩阵 A。

(4)求实非对称矩阵 A 的特征值 λ 及其共轭 λ^* 和对应的特征向量 V 及其共轭向量 V^*。

(5)将特征值和特征向量代入(7.5.6)式，递推求出 *POP* 系数 Z。

(6)计算每对特征向量占总方差的百分比。

(7)用(7.5.11)和(7.5.12)式算出(7.5.9)式描述的振荡模态的振荡周期和振幅衰减时间。

(8)如果需要考虑另一系统与本系统的关系，则利用(7.5.18)式计算伴随相关。

7.5.3 计算结果分析

对于 POP 计算结果的理解和解释并不是十分容易的事。有时根据需要对 POP 分析的结果可以再作些处理，以便使结果清晰、直观，便于分析和解释。如将 POP 模态由波动的振幅

$$A^2(r) = [V^r(r)]^2 + [V^i(r)]^2 \qquad (7.5.19)$$

和相对位相

$$\Psi(r) = \tan^{-1}[V^i(r)/V^r(r)] \qquad (7.5.20)$$

的空间分布图来表示。再如：计算 POP 系数 Z^i 和 Z^r 序列的交叉谱，分析其时间的振动特征。下面结合应用实例对计算结

果的分析简略加以说明。

(1)从 T 值来分析振荡 POP 对的振荡周期大约是多少。例如:章基嘉等人对热带太平洋地区(122.5°E～87.5°W,27.5°N～27.5°S)的月平均海温距平场作 POP 分析[18]。第一对 POP 振荡周期 T 大约为 39 个月。

(2)根据特征值 λ 的大小,对 POP 描述的波动进行分类:

①当 $|\lambda|<1,\tau>0$ 时,振幅随时间而减弱,称为衰减波动。

②当 $|\lambda|=1,\tau=\infty$ 时,振幅不随时间而变,称为中性波动。

③当 $|\lambda|>1,\tau<0$ 时,振幅随时间而增大,称为增长波动。

文献[18]中的例子,第一对 POP 描述的即为衰减波动。

(3)对特征向量 V^r,V^i 表征的变量场的空间传播特征进行分析。分析时注意振荡模态的交替出现。在 $t=0$ 时,海温距平场第一对 POP 的 V^r 型在热带太平洋的东部和中部为正值,其余区域为弱的负值区,它代表厄尔尼诺现象的成熟位相。那么 $t=\frac{\tau}{4}$ 时,即大约 9 个月以后,由 $-V^i$ 型替代,东太平洋变为弱的正值,中太平洋则为较强的负值中心,代表厄尔尼诺衰减位相。依…→V^r →$-V^i$ →$-V^r$ →V^i →V^r →…顺序,$t=\frac{2\tau}{4}$ 时,即大约 18 个月以后为 $-V^r$ 型,热带太平洋的东部和中部变为负值,厄尔尼诺完全消失或出现拉尼娜成熟位相。到 $t=\frac{3\tau}{4}$ 时,即 27 个月以后为 V^i 型,中太平洋开始出现正值,即呈厄尔尼诺开始发展。在 $t=\tau$ 时,又重复出现 V^r 型。可见,POP 型描述了厄乐尼诺的演变过程。

(4)由 POP 系数 Z^r,Z^i 分析振荡随时间的演变特征。文献[18]中所举的例子 Z^r 极大值对应历次厄尔尼诺事件,历史上最强的厄尔尼诺对应于最大的正振幅。Z^r 极小值对应拉尼

娜事件。Z^i 则基本上与 Z^r 相反，但是数值没有 Z^r 那么大。

(5)利用伴随相关型来分析 POP 与另一变量场的关系。q^r 型对应 V^r，q^i 对应 V^i。例如：文献[18]中，对上述海温距平场的 POP 系数与 850kPa 上风场距平求伴随相关型。q^r 型，在中太平洋，沿赤道有一支强西风气流，恰与厄尔尼诺事件海温赤道东暖西冷的分布相配合。

(6)通过对传播型的 V^r 型时间系数 Z^r 和 V^i 型时间系数 Z^i 序列分别作功率谱分析或作两序列的交叉谱分析，用通过显著性检验的周期来验证 POP 分析得到的振荡周期是否可信。段安民、吴洪宝曾作了这方面的分析[19]。

§7.6 循环稳态主振荡型分析

7.5 节介绍的 POP 是假定气候变量在 Stationay(定常)条件下进行的。定常的含义是一个变量的各种统计量不依赖时间。然而，许多气候变量是循环稳态(Cyclostationary)的，即气候过程及其变率是由若干个显著时间尺度表征的。例如：地面温度不仅呈现年际变化，而且还呈现更高频率的波动。这种波动依赖于时间，也就是存在显著的固有循环。例如：年际和日的循环。固有循环描述了许多气候过程及其变率特征。因此，一个循环稳态过程的各种统计量依赖于固有循环的一个特定位相。循环稳态变量的最初构思是 1985 年 Hasselmann 在一个未发表的手稿中提出来的。1991 年由 Blumenthal 给出了如何实施循环稳态 POP 分析的具体方案[20]。

7.6.1 方法概述

7.6.1.1 Cyclostationary 过程的描述

给定一对整数 t，T。其中 t 是气候时间序列含有的循环次

数。例如：对于 40 年的序列，年循环次数 $t=40$，T 是季节日期次数，一个循环内含有的时间步长 $T=1,2,\cdots,n$。例如：一个年循环内，月的时间步长为 $T=1,2,\cdots,12$。

不失一般性，序列的循环稳态表示为：

$$(t,T+n)=(t+1,T) \tag{7.6.1}$$

那么，气候变量的循环稳态过程可以写为：

$$x(t,T+1)=A(T)x(t,T)+\text{noise} \tag{7.6.2}$$

且

$$\begin{aligned} x(t,T+n)&=x(t+1,T)\\ A(T+n)&=A(T) \end{aligned} \tag{7.6.3}$$

连续使用 n 次(7.6.2)式，并利用下式

$$B(T)=A(T+n-1)A(T+n-2),\cdots,A(T+1)A(T) \tag{7.6.4}$$

可以得到：

$$x(t+1,T)=B(T)x(t,T)+\text{noise} \tag{7.6.5}$$

由于存在周期性，那么就有 n 个形如(7.6.5)式的模型。这样对于每个模型可以使用常规的 POP 分析。

7.6.1.2　特征值和特征向量的确定

对于(7.6.5)式中的常数矩阵 $B(T)$ 的特征值 λ^T 和特征向量 V^T

$$B(T)V^T=\lambda^T V^T \tag{7.6.6}$$

且有 $\overline{V}^T V^T=1$。对于 n 个不同 $B(T)$ 模型 λ^T 均相同。$\overline{V}^T$ 表示 V^T 的转置共轭。

循环稳态系统存在以下关系：

①$B(T+1)$与 $B(T)$具有相同的特征值 λ^T。

②若 V^T 是 $B(T)$的特征向量，则 $A(T)V^T$ 就是 $B(T+1)$的特征向量。

特征向量可以由下式递推出来

$$V^{T+1} = CA(T)V^T \tag{7.6.7}$$

其中C为任一复数

$$C = r_T^{-1}\exp i\varphi_T \tag{7.6.8}$$

式中r_T为衰减率。如果选择模$|C| = r_T^{-1}$，则有：

$$\overline{V}^T V^T = 1$$

对某一时间步长T而言，若满足归一化条件(7.6.9)式，且如果

$$r_T = \| A(T)V^T \| \tag{7.6.10}$$

则

$$\overline{V}^{T+1}V^{T+1} = 1 \tag{7.6.11}$$

由于满足周期性条件$V^{T+n} = V^T$，对于所有的T均可以确定出角度φ_T

$$\varphi_T = \omega/n \tag{7.6.12}$$

这里$\lambda = \rho\exp(-i\omega)$，$\rho = \prod_{k=0}^{n-1} r_{T+k}$。

对于一个循环内，POP随因子ρ而衰减，随角度$-\omega$而旋转。因此，为了保证$V^{T+n} = V^T$，在每一时间步上振荡型由r_T来控制，由ω/n来向后旋转。

7.6.1.3 POP系数的确定

由下列递推公式导出POP系数：

$$Z(t, T+1) = r_T\exp(-i\omega/n)Z(t,T) + \text{noise} \tag{7.6.13}$$

重复使用(7.6.13)式，就可以得到常规POP模型

$$Z(t+1,T) = \left(\prod_{k=0}^{n-1} r_{T+k}\right)\exp(-i\omega)Z(t,T) + \text{noise} = \lambda Z(t,T) + \text{noise} \tag{7.6.14}$$

7.6.1.4　常数矩阵 $A(T)$的估计

对每一循环内的 $T=1,2,\cdots,n$，可以使用常规的 POP 估计 $A(T)$的办法计算，即

$$A(T) = S_{1,T}S_{0,T}^{-1} \qquad (7.6.15)$$

其中 $S_{0,T}$和 $S_{1,T}$是滞后时刻为 0 和 1 的协方差矩阵。

7.6.2　计算结果分析

对于存在明显季节循环的气候变量场可以考虑作 Cyclostationary POP 分析。同时可以作常规的 POP 分析，两种结果进行比较。除了分析与常规 POP 相同的内容外，对 Cyclostationary POP 还可以分析如下内容：

(1)分析衰减率 r_T 随季节的变化。从中可以了解气候过程中哪几个月在加强，其中最强发生在哪个月；哪几个月在衰减，最弱的出现在哪个月。另外，分析最强与最弱滞后的时间。

(2)由 Cyclostationary POP 的特征向量和 POP 系数可以分析气候变量场随季节而变化的空间分布特征及时间演变特征。当然，也可以分析年平均的时空分布特征。

§7.7　复主振荡型分析

复主振荡型(Complex Principal Oscillation Patterns，CPOP)需要先通过 Hibert 变换构造一个复数时间序列，在此基础上作 POP 分析。有关 CPOP 的计算，Storch 等人在文献[17]中作了简单介绍。

如果原实数序列 $x(t)$可以分解为傅里叶形式，那么就可以利用 7.4 节 CEOF 中叙述的傅里叶变换办法构造复数矩阵的虚数部分 $\hat{x}(t)$。复数向量表示为：

$$w(t) = x(t) + i\hat{x}(t) \qquad (7.7.1)$$

复数 POP 可以通过一阶线性模型得到，即

$$w(t+1) = Aw(t) + \text{noise} \tag{7.7.2}$$

系数矩阵 A 的估计为：

$$A = S_1 S_0^{-1} \tag{7.7.3}$$

其中

$$\begin{aligned} S_0 &= \langle w(t)\overline{w}(t)\rangle \\ S_1 &= \langle w(t+1)\overline{w}(t)\rangle \end{aligned} \tag{7.7.4}$$

这里 $\overline{w}(t)$表示 $w(t)$的转置共轭。A 的特征向量就是 CPOP。由于 A 是复数，特征向量与实数或标准的 POP 不同，它们不是以共轭复数对的形式出现，CPOP 的个数等于(7.7.2)式过程的维数。

复数状态 $w(t)$的展开式

$$w(t) = \sum_j Z_j(t)V_j \tag{7.7.5}$$

对任何给定时刻 t，特征向量 V 对 $w(t)$的贡献由

$$P(t) = Z(t)V \tag{7.7.6}$$

给出或用

$$\begin{aligned} P &= P^r + iP^i \\ V &= V^r + iV^i \\ Z &= Z^r - iZ^i \end{aligned} \tag{7.7.7}$$

表示，即

$$P^r(t) = Z^r(t)V^r - Z^i(t)V^i \tag{7.7.8}$$

$$P^i(t) = Z^r(t)V^i - Z^i(t)V^r \tag{7.7.9}$$

实部 $P^r(t)$描述的是 $x(t)$空间的信号，虚部 $P^i(t)$描述的是动量 $\hat{x}(t)$空间的信号。无噪音的 CPOP 系数的时间演变由

$$Z(t+1) = \lambda Z(t) \tag{7.7.10}$$

给出。这里 $\lambda = \rho\exp(-i\omega)$。那么(7.7.8)和(7.7.9)式可以表

示为：

$$P^r(t) = \rho^l[\cos(\omega t)V^r - \sin(\omega t)V^l] \quad (7.7.11)$$

$$P^l(t) = \rho^l[\cos(\omega t)V^l + \sin(\omega t)V^r] \quad (7.7.12)$$

对于实部空间的振动模态的按常规 POP(7.5.10)式的顺序交替出现。而虚部空间的振动模态按下列顺序交替出现：

$$\cdots \to V^l \to V^r \to -V^l \to -V^r \to V^l \to \cdots \quad (7.7.13)$$

参考文献

[1]Pearson k. On lines and plans of closest fit tc system of points in space philos. Mag. 1902,6:559—572

[2]North G. R. K. Y. Kim S. S. P. Shen and J. W. Hardin. Detection of forced climatic signals, Part I: Theory. J. Climate, 1995, 8:401—408

[3]Kim K. Y, G. R. North and Jing-Ping Huang, EOFs of one-dimensional cyclostationary time series: computation, examples, and stochastic modelling, J. Atmos. Sci, 1996, 53:1007—1017

[4]张邦林，卫纪范．经验正交函数在气候数值模拟中的应用．中国科学(B 辑)，1991，4：442～448

[5]孙照勃，袁建强，张邦林．用逐步迭代法插补海表温度的研究．南京气象学院学报，1991，14(2)：143～149

[6]North G. R. T. Bell. R. Cahalan and F. J. Moeng, Sampling errors in the estimation of empirical orthogonal function. Mon. Wea. Rev. 1982, 110:699—706

[7]Preisendorfer R. W and T. P. Barnett. Significance test for empirical orthogonal function conference on probability and statistics in atmospheric science, Las Vegas, 1977. 169—172

[8]丁裕国，江志红．非均匀站网 EOFs 展开的失真性及其修正．气象学报，1995，53(2)：247～253

[9]Weare B. C. and T. S. Nasstrom. Examples of extended empirical orthogonal function analyses, Mon. Wea. Rev. 1982, 110(6):481—485

[10]张先恭,魏凤英．太平洋海表温度与中国降水准3.5年周期变化．见:长期天气预报理论和方法的研究课题组编."八五"长期天气预报理论和方法的研究．北京:气象出版社,1996.169～175

[11]Rasmusson E.M.P.A.Arkin and W.Y.Chen. Biennial Variation in surface temperature over the United States as revealed by singular decomposition. Mon. Wea. Rev. 1981,109:587--598

[12]Barnett T.P. Interaction of the monsoon and Pacific trade wind systems at interannual time scales, Part I: The equatorial Zone. Mon. Wea. Rev. 1983,111:756—773

[13]魏凤英,张先恭,李晓东.用CEOF分析近百年中国东部旱涝的分布及其年际变化特征,应用气象学报,1995,6(4):454～460

[14]Hasselmann K. PIP and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns, J. Geophys. 1988,93:11015—11021

[15]Storch H.V. T. Bruns, I. F. Bruns and Hasselmann K. Principal oscillation patterns analysis of 30 to 60 day oscillation in a GCM, J. Geophys, 1988,93:11022 -11036

[16]Xu J.S. On the relationship between the stratospheric QBO and the tropospheric SO. J. Atmos. 1992,49:725--734

[17]Storch H. V. Burger, G. Schnur, R. and J. S Storch. Principal Oscillation patterns: A review, J. climate, 1995,8:377—400

[18]章基嘉,丁锋,王盘兴．大尺度海气异常关系的主振荡型分析．应用气象学报,1993,4(增刊):1～9

[19]段安民,吴洪宝．全球热带海表温度异常的主振荡型分析．南京气象学院学报,1998,21(1):61～69

[20]Blumenthal B. Predictability of a coupled ocean-atmosphere model, J. Climate. 1991,4:766—784

第八章　两气候变量场相关模态的分离

在气候变化研究中，存在着大量两个变量场之间的相关问题，即研究两个场之间相关系数的空间结构和它们各自对相关场的贡献。对于这类研究，计算普通 pearson 相关系数难以奏效，因为 pearson 相关系数是一种点相关，无法得到两场相关的整体概念，也不能分离出两变量场的空间相关模态。目前，已有多种分离两场相关模态的方法，常用的有：①联合 EOF 分析。将两变量场的资料合并为一个矩阵，然后执行 EOF 过程，提取两场耦合关系的模态；②相关场 EOF 分析。假设一变量场 $X_{p\times n}$，另一变量场 $Y_{q\times n}$，计算它们的两两相关系数，构造出相关场 $R_{p\times q}$，然后再作 EOF 分析。由于 R 一般不是对称方阵，求解 R 的特征值和特征向量时不能再使用 Jacobi 方法，而需要用求解实非对称矩阵的奇异值分解式；③典型相关分析。这一方法是将两变量场转化为几个典型变量，通过研究典型变量之间的相关系数来分析两变量场的相关，这一方法可以有效地分离两场的最大线性相关模态；④BP 典型相关分析。它是典型相关的一种新形式。它的基本思路是讨论两场主分量的关系；⑤奇异值分解。它的出发点与典型相关相同，但是计算要简便得多。从统计学角度来讲，典型相关条件强，推理严谨，而奇异值分解需要一定的使用条件。这里列出的第 1 种方法，在合并两个变量场资料之后，计算上与普通 EOF 无差别。所不同的是，要根据联合特征向量的各分量权重提取两场耦合相关信息，这里不作叙述。第 2 种方法，在

计算上与普通 EOF 的区别在于，相关阵需要使用矩阵理论中的奇异值分解定理求解特征值和特征向量，这里也不作单独介绍。本章将重点讨论后三种方法。有关奇异值求解形式及其计算在介绍的三种方法中均要涉及到，因此在第一节中先作一简单介绍。

§8.1 奇异值分解式定理及其计算

求解矩阵的特征值和特征向量是多元分析中十分重要，且是基本的计算。在 EOF 分解中，变量场的协方差阵是实对称矩阵，可以采用 Jacobi 方法求解矩阵的特征值和特征向量。在本章中，涉及到的两变量场的交叉协方差阵常常是实非对称阵。对于这类矩阵需要利用奇异值分解式求解其特征值和特征向量[3]。

8.1.1 奇异值分解式定理

设两变量场的交叉协方差阵 $V_{p\times q}$ 为任意非零矩阵，则有

$$V_{p\times q}=L_{p\times p}\Lambda_{p\times q}R'_{q\times q} \tag{8.1.1}$$

为 V 的奇异值分解式。可以证明，任意非零矩阵必存在(8.1.1)奇异值分解式。式中 L 和 R 分别为 p 和 q 阶正交方阵，$q>p$，Λ 为 $p\times q$ 阶对角阵，Λ 的对角线元素为：

$$\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p) \tag{8.1.2}$$

称 Λ 为 V 的奇异值。

由于

$$V'V=R\Lambda' L'L\Lambda R'=R\mathrm{diag}(\lambda_1^2,\lambda_2^2,\cdots,\lambda_p^2)R' \tag{8.1.3}$$

因此，$\lambda_1^2,\lambda_2^2,\cdots,\lambda_p^2$ 是 $V'V$ 的特征值，且 $\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_p^2$，$R$ 是 $V'V$ 对应于 λ 的特征向量。

同理，由 VV' 可知，$\lambda_1^2,\lambda_2^2,\cdots,\lambda_p^2$ 亦是 VV' 的特征值，L 则是 VV' 对应于 λ_i 的特征向量。可见 VV' 和 $V'V$ 有相同的非负特征值。

8.1.2　奇异值分解的计算

设 V 的秩（这里用 rk 表示）为 $m(m\leqslant\min(p,q))$，$rk(V)-rk(V'V)=m$，

因此有

$$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_m^2>0,\lambda_{m+1}^2=\cdots\lambda_p^2=0$$

这里记 $R_1=(r_1,r_2,\cdots,r_m)$，$R=(R_1R_2)$，$\Lambda_1=\mathrm{diag}(\lambda_1^2,\lambda_2^2,\cdots,\lambda_m^2)$ 则

$$V'V=(R_1R_2)\begin{bmatrix}\Lambda_1 & 0\\ 0 & 0\end{bmatrix}\begin{pmatrix}R_1'\\ R_2'\end{pmatrix}=$$

$$R_1\Lambda_1R_1'=R_1\Lambda_1^{1/2}\Lambda_1^{1/2}R_1' \qquad (8.1.4)$$

式中 $\Lambda_1^{1/2}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_m)$

显然，有

$$(VR_1\Lambda_1^{1/2})'(VR_1\Lambda_1^{1/2})=I \qquad (8.1.5)$$

如果设

$$L_1=VR_1\Lambda_1^{-1/2} \qquad (8.1.6)$$

那么，

$$L_1'L_1=I \qquad (8.1.7)$$

说明 L_1 是 $p\times m$ 列正交阵。

由于 $R_2=(R_{m+1},R_{m+2},\cdots,R_q)$

$$R'R=I \qquad (8.1.8)$$

因此

$$R_1R_1'+R_2R_2'=I \qquad (8.1.9)$$

$$VR_2=0 \qquad (8.1.10)$$

因而有

$$L_1 \Lambda_1^{1/2} R_1' = V R_1 R_1' = V(I - R_2 R_2') = V - 0 = V \tag{8.1.11}$$

若将 L_1 扩充为正交方阵 $L=(L_1 L_2)$ 则有

$$V = L \begin{bmatrix} \Lambda_1^{1/2} & 0 \\ 0 & 0 \end{bmatrix} R' \tag{8.1.12}$$

简单归纳起来，矩阵 V 的奇异值分解的计算步骤为：

8.1.2.1 计算

$$V'V = R \begin{bmatrix} \lambda_1^2 & & & & \\ & \ddots & & & \\ & & \lambda_m^2 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} R'$$

这里 $V=(V_1 V_2)$，V_1 为 $p \times m$ 阶矩阵。

8.1.2.2 计算

$$L_1 = V R_1 \Lambda_1^{-1/2}$$

$$\Lambda_1^{1/2} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m)$$

8.1.2.3 利用(8.1.11)式形成 V 的奇异值分解

$$V = L_1 \Lambda_1^{1/2} R_1'$$

§8.2 典型相关分析

1936 年 Hotelling 在研究两组变量间的相关关系时，引进了典型相关和典型变量的概念[2]。将原来较多的变量转化为少数几个典型变量，通过研究典型变量之间的相关系数，分析两组变量间的相关关系。到了本世纪 60 年代，典型相关分析(Canonical Correlation Analysis，CCA)作为一种分析手段在社会科学研究领域得到广泛应用。1968 年 Glahn 首次将

CCA 使用在统计天气预报中[3]。从此，CCA 开始在两气象变量组或两变量场的相关研究中应用[4-5]。应用实践表明，CCA 是一种具有坚实数学基础，推理严谨，能够有效提取两组变量或两变量场相关信号的有用工具。分离两变量场的相关结构，就是将两变量场的每个测站或网格点资料视为变量，这样研究对象仍归结为两组变量。

8.2.1 方法概述

假设我们研究的两组变量或两个变量场，一组变量或一个场 X 有 p 个变量或空间点，样本量为 n；另一组变量或一个场 Y 有 q 个变量或空间点，样本量亦为 n。这里要求 $n>p,q$，变量场 X 资料矩阵为：

$$X=\begin{bmatrix}x_1\\x_2\\\vdots\\x_p\end{bmatrix}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\\vdots&\vdots&&\vdots\\x_{p1}&x_{p2}&\cdots&x_{pn}\end{bmatrix} \tag{8.2.1}$$

变量场 Y 资料矩阵为：

$$Y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_q\end{bmatrix}=\begin{bmatrix}y_{11}&y_{12}&\cdots&y_{1n}\\y_{21}&y_{22}&\cdots&y_{2n}\\\vdots&\vdots&&\vdots\\y_{q1}&y_{q2}&\cdots&y_{qn}\end{bmatrix} \tag{8.2.2}$$

其中 $x_k(k=1,2,\cdots,p)$和 $y_k(k=1,2,\cdots,q)$均为含 n 次观测的向量：

$$\begin{aligned}x_k&=(x_{k1}\quad x_{k2}\quad\cdots\quad x_{kn})\\y_k&=(y_{k1}\quad y_{k2}\quad\cdots\quad y_{kn})\end{aligned} \tag{8.2.3}$$

8.2.1.1 协方差阵

变量场 X 的协方差阵为：

$$S_{xx}=\frac{1}{n}XX'=\frac{1}{n}\begin{bmatrix}x_1\\x_2\\\vdots\\x_p\end{bmatrix}[x_1',x_2',\cdots,x_p'] \quad (8.2.4)$$

变量场 Y 的协方差阵为：

$$S_{yy}=\frac{1}{n}YY'=\frac{1}{n}\begin{bmatrix}y_1\\y_2\\\vdots\\y_q\end{bmatrix}[y_1',y_2',\cdots,y_q'] \quad (8.2.5)$$

两场之间协方差阵为：

$$S_{xy}=\frac{1}{n}XY'=\frac{1}{n}\begin{bmatrix}x_1y_1' & x_1y_2' & \cdots & x_1y_q'\\x_2y_1' & x_2y_2' & \cdots & x_2y_q'\\\vdots & \vdots & & \vdots\\x_py_1' & x_py_2' & \cdots & x_py_q'\end{bmatrix} \quad (8.2.6)$$

$$S_{yx}=\frac{1}{n}YX'=\frac{1}{n}\begin{bmatrix}y_1x_1' & y_1x_2' & \cdots & y_1x_p'\\y_2x_1' & y_2x_2' & \cdots & y_2x_p'\\\vdots & \vdots & & \vdots\\y_qx_1' & y_qx_2' & \cdots & y_qx_p'\end{bmatrix} \quad (8.2.7)$$

显然

$$S_{xy}=S'_{yx}$$

将两个场组合为一个 $p+q$ 个变量的向量，$p+q$ 个变量的协方差阵为：

$$S=\begin{bmatrix}S_{xx} & S_{xy}\\S_{yx} & S_{yy}\end{bmatrix} \quad (8.2.8)$$

8.2.1.2 典型变量与典型相关系数

典型相关的基本思想是，对两组变量分别作线性组合构成新的一对变量 u_1, v_1，使得它们之间有最大相关系数。再分别作与 u_1, v_1 正交的线性组合 u_2, v_2，使它们之间有其次大的相关系数。如此进行下去，直至认为合适为止，$u_i, v_i, i=1,2\cdots$ 就称为典型变量。

变量场 X 的原 p 个变量线性组合为一新变量：

$$u_1 = c_{11}x_1 + c_{21}x_2 + \cdots + c_{p1}x_p = c_1'X \quad (8.2.9)$$

其中

$$c_1' = (c_{11}\ c_{21}\ \cdots\ c_{p1}) \quad (8.2.10)$$

变量场 Y 的原 q 个变量线性组合为一新变量

$$v_1 = d_{11}y_1 + d_{21}y_2 + \cdots + d_{q1}y_q = d_1'Y \quad (8.2.11)$$

其中

$$d_1' = (d_{11}\ d_{21}\ \cdots\ d_{q1}) \quad (8.2.12)$$

称 u_1, v_1 为典型变量，$c_{11}, c_{21}, \cdots, c_{p1}$ 和 $d_{11}, d_{21}, \cdots, d_{q1}$ 为典型荷载特征向量。

为使线性组合后的新变量具有数学期望等于 0，方差等于 1，即对 u_1 变量有

$$\frac{1}{n}u_1u_1' = c_1'S_{xx}c_1 = 1 \quad (8.2.13)$$

同理，对 v_1 变量有：

$$\frac{1}{n}v_1v_1' = d_1'S_{yy}d_1 = 1 \quad (8.2.14)$$

上述一对典型变量之间的相关系数在两个变量场所有线性组合而成的典型变量中最大，即要求相关系数

$$r_1 = \frac{1}{n}u_1v_1' = c_1'S_{xy}d_1 \quad (8.2.15)$$

最大。称 r_1 为典型相关系数。

再作线性组合 u_2, v_2，在与 u_1, v_1 线性无关情况下，满足在剩余方差中，它们之间相关系数

$$r_2 = \frac{1}{n} u_2 V_2' = c_2' S_{xy} d_2 \tag{8.2.16}$$

达到极大，且 u_2, v_2 方差为 1。

如此继续下去，依次有第三对典型变量 u_3, v_3……可以证明，典型变量的对数等于两个变量场协方差阵 S_{xy} 的秩数，对气候场即为空间点数 p, q 中最小的数。这里假定可以找到 q 对典型变量。

8.2.1.3　典型荷载特征向量

在约束条件(8.2.13)和(8.2.14)式下，满足(8.2.15)式协方差极大原则。由拉格朗日乘法求函数

$$Q = c_1' S_{xy} d_1 - \frac{\nu_1}{2}(c_1' S_{xx} c_1 - 1) - \frac{\nu_2}{2}(d_1' S_{yy} d_1 - 1) \tag{8.2.17}$$

的极大值。其中 ν_1, ν_2 为拉格朗日乘数。函数 Q 的极值问题归结为

$$\frac{\partial Q}{\partial c_1} = 0 \tag{8.2.18}$$

$$\frac{\partial Q}{\partial d_1} = 0 \tag{8.2.19}$$

将(8.2.17)式代入(8.2.18)和(8.2.19)式有：

$$\begin{cases} S_{xy} d_1 - \nu_1 S_{xx} c_1 = 0 & ① \\ S_{yx} c_1 - \nu_2 S_{yy} d_1 = 0 & ② \end{cases} \tag{8.2.20}$$

分别左乘 c_1' 和 d_1' 有

$$\begin{cases} c_1' S_{xy} d_1 - \nu_1 c_1' S_{xx} c_1 = 0 & ① \\ d_1' S_{yx} c_1 - \nu_2 d_1' S_{yy} d_1 = 0 & ② \end{cases} \tag{8.2.21}$$

将约束条件(8.2.13)和(8.2.14)式分别代入(8.2.21)式中的

①和②得：

$$\begin{cases} c_1' S_{xy} d_1 = \nu_1 \\ d_1' S_{yx} c_1 = \nu_2 \end{cases} \tag{8.2.22}$$

由于 $c_1' S_{xy} d_1$ 的矩阵乘积是一个数，且 $S_{xy}' = S_{yx}$

$$c_1' S_{xy} d_1 = d_1' S_{yx} c_1 \tag{8.2.23}$$

因此

$$c_1' S_{xy} d_1 = \nu_1 = \nu_2 \tag{8.2.24}$$

又由于

$$r_1 = c_1' S_{xy} d_1$$

即

$$r_1 = \nu_1 = \nu_2$$

对(8.2.20)式中①左乘 $S_{yx} S_{xx}^{-1}$ 则有

$$S_{yx} S_{xx}^{-1} S_{xy} d_1 - \nu_1 S_{yx} S_{xx}^{-1} S_{xx} c_1 = 0 \tag{8.2.25}$$

即

$$S_{yx} c_1 = \frac{1}{\nu_1} S_{yx} S_{xx}^{-1} S_{xy} d_1$$

将(8.2.25)式代入(8.2.20)式中②得：

$$\frac{1}{\nu_1} S_{yx} S_{xx}^{-1} S_{xy} d_1 - \nu_1 S_{yy} d_1 = 0 \tag{8.2.26}$$

即

$$(S_{yx} S_{xx}^{-1} S_{xy} - \nu_1^2 S_{yy}) d_1 = 0$$

令 $\lambda_1 = \nu_1^2$，则有

$$(S_{yx} S_{xx}^{-1} S_{xy} - \lambda_1 S_{yy}) d_1 = 0 \tag{8.2.27}$$

对上式左乘 S_{yy}^{-1} 得：

$$(S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} - \lambda_1 I) d_1 = 0 \tag{8.2.28}$$

由(8.2.28)式可知，问题归结为求 λ_1 和 d_1。也就是求矩阵 $S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy}$ 的特征值 λ_1 及其对应的特征向量。对于实非对

称矩阵的特征值和特征向量，用 8.1 节中介绍的奇异值分解形式求解。

求出 λ_1 和典型荷载特征向量 d_1 后，可以很容易地求出系数 c_1，利用(8.2.20)式中①得到：

$$\begin{cases} S_{xy}d_1 = \nu_1 S_{xx}c_1 \\ c_1 = S_{xx}^{-1}S_{xy}d_1/\nu_1 \end{cases} \tag{8.2.29}$$

利用 λ_1 和 ν_1 的关系，可得：

$$c_1 = \frac{S_{xx}^{-1}S_{xy}d_1}{\sqrt{\lambda_1}} \tag{8.2.30}$$

求出荷载特征向量 c_1 和 d_1，就可以得到第一对典型变量 u_1 和 v_1。

第一对典型变量的典型相关系数为：

$$r_1 = \nu_1 = \sqrt{\lambda_1} \tag{8.2.31}$$

类似地，可以求出 X 和 Y 的第二对典型变量 u_2, v_2，其方差为 1，且与 u_1, v_1 不相关，它们具有其次大的相关系数。依次进行下去，求出 q 个特征根 $\lambda_i (i=1,\cdots,q)$ 及相应的荷载特征向量，并由 $c_i, d_i (i=1,\cdots,q)$ 得到 q 对典型变量及其相关系数。

8.2.1.4　典型相关系数的显著性检验

对于 q 对典型相关变量，其相关性是否显著需要进行检验。将问题化为典型相关系数为 0 的假设检验。采用 Bartlett 关于大样本的 χ^2 检验。将特征根按由大到小排列 $\lambda_1^2 \geqslant \lambda_2^2 \geqslant \cdots \geqslant \lambda_q^2$。令

$$L_1 = (1-\lambda_1^2)(1-\lambda_2^2),\cdots,(1-\lambda_q^2) \tag{8.2.32}$$

对于较大的样本量 n，在两组变量总体不相关的假设下，统计量

$$\chi_1^2 = -\left[(n-1) - \frac{1}{2}(p+q+1)\right]\ln L_1 \quad (8.2.33)$$

近似地遵从自由度为 $p \times q$ 的 χ^2 分布。选定显著性水平 α，查 χ^2 分布表，若 $\chi_1^2 > \chi_\alpha^2$ 则认为第一个典型相关系数是显著的。表明第一对典型变量显著相关。

减去第一个典型相关系数 λ_1^2，这时令

$$L_2 = (1-\lambda_2^2)(1-\lambda_3^2),\cdots,(1-\lambda_q^2) \quad (8.2.34)$$

统计量

$$\chi_2^2 = -\left[(n-2) - \frac{1}{2}(p+q+1)\right]\ln L_2 \quad (8.2.35)$$

近似地遵从自由度为 $(p-1)(q-1)$ 的 χ^2 分布。若 $\chi_2^2 > \chi_\alpha^2$ 则第二个典型相关系数是显著的，也就是说第二对典型变量显著相关。依次进行下去，这样可以找到反映两组变量相互联系的 k 对典型变量。

8.2.1.5 典型回归

通过显著性检验的典型变量代表了两变量场之间的线性协方差关系的主要信息，且又相互独立。利用这种典型相关，可以建立两变量场的典型回归方程。

假设找到 k 对典型变量，那么，变量场 Y 的典型变量矩阵 $V_{k\times n}$ 与变量场 X 的典型变量矩阵 $U_{k\times n}$ 满足

$$V = \Lambda^{\frac{1}{2}} U \quad (8.2.36)$$

其中

$$\Lambda^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_k} \end{bmatrix}$$

$$U = C'X \tag{8.2.37}$$

其中

$$C' = (c_1\ c_2\ \cdots\ c_k)$$

$$V = D'Y \tag{8.2.38}$$

其中

$$D' = (d_1\ d_2\ \cdots\ d_k)$$

将(8.2.37)和(8.2.38)式代入(8.2.36)式得：

$$D'Y = \Lambda^{\frac{1}{2}}C'X \tag{8.2.39}$$

对(8.2.39)式两边左乘 D 再求解：

$$Y = (DD')^{-1}D\ \Lambda^{\frac{1}{2}}C'X \tag{8.2.40}$$

由于

$$D'S_{yy}D = I \tag{8.2.41}$$

对(8.2.41)式左乘 D 右乘 D' 得：

$$DD'S_{yy}DD' = DD'$$

$$S_{yy} = (DD')^{-1} \tag{8.2.42}$$

将(8.2.42)式代入(8.2.40)式得到：

$$Y = S_{yy}D\ \Lambda^{\frac{1}{2}}C'X \tag{8.2.43}$$

8.2.2 计算步骤

典型相关分析的计算步骤如下：

(1)对变量场 X 和 Y 进行标准化预处理。

(2)计算标准化后的变量场 X 的协方差阵 S_{xx}，变量场 Y 的协方差阵 S_{yy}和两变量场交叉协方差阵 S_{xy}。

(3)解方程：

$$(S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy} - \lambda S_{yy})d = 0$$

用奇异值分解计算方法求出 $S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}$矩阵的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_q$ 及对应的荷载特征向量 $d_1, d_2, \cdots, d_q$。

(4)利用特征值 λ_i 和荷载特征向量 d_i 求 c_i

$$c_i = \frac{S_{xx}^{-1}S_{xy}d_i}{\sqrt{\lambda_i}} \quad (i = 1,2,\cdots,q)$$

(5)计算典型变量：

$$\begin{aligned} U_i &= c_i' X \\ V_i &= d_i' Y \end{aligned} \quad (i=1,2,\cdots,q)$$

(6)求典型相关系数：

$$r_i = \sqrt{\lambda_i} \quad (i = 1,2,\cdots,q)$$

(7)对典型相关系数进行显著性检验。

(8)如果需要，利用(8.2.43)式建立典型回归方程。

8.2.3 计算结果分析

对于典型相关计算结果大致可以作以下几方面的分析：

(1)典型相关系数。典型相关系数反映了两典型变量场之间的相关程度。通过显著性检验的典型相关系数越大，表明两典型变量场之间的相关越密切。

(2)典型荷载特征向量。变量场经过标准化处理，典型荷载特征向量的元素 $c_{11},c_{21},\cdots,c_{p1},\cdots,d_{11},d_{21},\cdots,d_{q1}\cdots$ 就是相应变量的权重系数。由一对典型变量的特征向量构成两变量场的一对典型场。通过荷载特征向量各分量的数值和符号分析两典型场之间同时或滞后的相关关系。权重绝对值大的空间区域有可能提供有价值的信号。

(3)典型变量序列。将两变量场 n 次观测标准化资料逐一代入(8.2.9)和(8.2.11)式，可以得到典型变量的时间序列。通过典型变量时间序列，分析其随时间的演变特征和规律。

(4)分别计算显著典型变量与对应的原变量场的相关系数。这样得到两变量场的空间相关分布模态。相关分布型在

一定程度上反映了两变量场的遥相关特征。以研究两变量场相关结构为主要目的，就是以这种相关模态为分析对象，从相关模态中可以检测出显著典型变量反映两变量场相互作用的敏感区域。

应用实例[8.1]：作为计算实例，这里给出简单两组变量的典型相关计算结果。第一组变量 Y 取长江流域 1951～1996 年 6,7,8 月降水量，记为 y_1, y_2, y_3。第二组变量 X 取 1951～1996 年 5 月西太平洋副高面积指数、副高脊线和副高西伸脊点，记为 x_1, x_2, x_3。这里 $q=3, p=3, n=46$。

变量标准化处理后，实施上述计算步骤，得到荷载典型变量特征向量：

$$U=\begin{bmatrix} -0.3852 & -1.1230 & 0.8323 \\ -0.5153 & 0.4398 & 0.7946 \\ 0.5024 & -0.6354 & 1.2229 \end{bmatrix}=[u_1\ u_2\ u_3]$$

$$V=\begin{bmatrix} -0.8099 & 0.6228 & 0.3898 \\ 0.1000 & -0.9999 & 0.5212 \\ -0.6457 & -0.1571 & -0.8006 \end{bmatrix}=[v_1\ v_2\ v_3]$$

典型相关系数：$r_1=0.5251, r_2=0.2693, r_3=0.0126$ 经显著性检验第一典型相关系数 r_1 是显著的。那么，第一对典型变量为：

$$u_1=-0.3852x_1-0.5153x_2+0.5024x_3$$

$$v_1=-0.8099y_1+0.1000y_2-0.6457y_3$$

由上式可以得到，第一典型变量逐年(1951～1996 年)值的序列，列于表 8.1。

计算第一典型变量与原两组变量的相关系数：

$UR_1=-0.71$　　$UR_2=-0.58$　　$UR_3=0.86$

$VR_1=-0.78$　　$VR_2=-0.39$　　$VR_3=-0.63$

从相关系数可以看出，在前期西太平洋副高与长江流域夏季降水的关系中，西伸脊点与夏季降水有较大的正相关，说明它在二者的关系中起主要作用。另外，副高面积指数与夏季降水有较明显的负相关。副高与长江流域夏季降水的关系主要反映在 6 和 8 月，即与这两个月的降水关系密切。

表 8.1　第一典型变量逐年(1951～1996 年)值

			u_1				
0.336	0.338	0.405	−1.343	−1.194	−0.972	2.354	−0.608
−0.418	−0.277	−0.853	−0.879	0.210	−0.730	−0.172	0.877
2.306	2.401	−0.734	−0.248	0.257	0.521	−0.547	2.401
0.569	2.259	−0.358	−0.112	−0.232	−0.537	−0.260	−0.680
−0.849	−0.077	0.545	1.151	−0.381	−0.260	0.097	0.202
−0.924	−0.049	−0.902	−1.282	−0.808	−0.545		
			v_1				
0.881	0.420	−0.082	−2.446	−1.641	−0.667	0.226	1.074
0.523	0.682	0.550	−0.786	0.966	0.046	0.020	1.442
0.973	1.651	−1.029	0.327	−0.008	0.797	−0.266	0.729
−1.140	0.499	−1.196	1.662	0.220	−2.561	1.348	−0.313
−0.677	−0.477	1.620	0.417	0.172	−0.870	−0.350	0.091
0.511	0.429	−0.903	−1.043	−1.298	−0.523		

§8.3　BP 典型相关分析

在使用典型相关分析时，要求样本量 n 大于两组变量的个数或两个变量场空间点数 p,q，以保证典型变量的稳定性。采取限制变量个数或空间点数的方式，有可能损失有价值的信息。1987 年，Barnett，T. P. 和 R. Preisendorfer 提出了一种典型相关分析的新计算格式[6]，既解决了上述问题，又使计算

得到了简化，人们将此方法称为 BPCCA。

利用典型相关分析两变量场相关关系时，首先构造各组变量的组合变量，然后再讨论两场组合变量之间的关系。而 BPCCA 直接用两变量场的主分量来讨论它们之间的关系。

8.3.1 方法概述

假设有标准化处理后的两组变量或两个变量场，一组变量或一个场 X 有 p 个变量或空间点，n 个样本量；另一组变量或另一个场 Y 有 q 个变量或空间点，n 个样本量。两场 EOF 分解的线性表达式分别为：

$$x_i(t) = \sum_{j=1}^{p_1} \lambda_j^{1/2} a_j(t) v_j(i)$$

$$(i = 1,2,\cdots,p \quad t = 1,2,\cdots,n) \qquad (8.3.1)$$

$$y_i(t) = \sum_{k=1}^{q_1} \mu_k^{1/2} b_k(t) e_k(i)$$

$$(i = 1,2,\cdots,q \quad t = 1,2,\cdots,n) \qquad (8.3.2)$$

式中，$\lambda_j, v_j(i)$ 分别是 X 场协方差阵 S_{xx} 的前 p_1 个特征值及其对应的特征向量；$\mu_k, e_k(i)$ 则分别是 Y 场协方差阵 S_{yy} 的前 q 个特征值及其对应的特征向量。$a_j(t)$ 和 $b_k(t)$ 分别为 X 和 Y 的主分量。

X 场主分量 $a_j(t)$ 是由特征值 λ_j 对应的时间系数 $t_j(t)$ 得到的，

$$a_j(t) = \frac{1}{\sqrt{\lambda_j}} t_j(t) \qquad (8.3.3)$$

类似地，Y 场主分量 $b_k(t)$ 是由特征值 μ_k 对应的时间系数 $f_k(t)$ 得到的，

$$b_k(t) = \frac{1}{\sqrt{\mu_k}} f_k(t) \qquad (8.3.4)$$

截取前 p_1 和前 q_1 个主分量构成变量场的两组新变量进行典型相关分析。这样，就能满足 $n > p_1, q_1$ 的条件，且达到样本量 n 与变量数 p_1, q_1 之比大于 2 的标准。

这时，变量场 X 的主分量资料阵为：

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{p1} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{p_1 1} & a_{p_1 2} & \cdots & a_{p_1 n} \end{bmatrix} \qquad (8.3.5)$$

变量 Y 的主分量资料阵为：

$$B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{q1} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{q_1 1} & b_{q_1 2} & \cdots & b_{q_1 n} \end{bmatrix} \qquad (8.3.6)$$

与典型相关一样，X 场主分量矩阵 A(8.3.5)式的协方差阵为 S_{aa}，Y 场主分量矩阵 B(8.3.6)式的协方差阵为 S_{bb}，两场主分量的交叉协方差阵为 S_{ab}。

变量场 X 的 p_1 个主分量的 BP 线性组合为：

$$u_1 = c_{11}a_1 + c_{21}a_2 + \cdots + c_{p_1 1}a_{p1} \qquad (8.3.7)$$

其中，

$$C_1 = (c_{11}\, c_{21} \cdots c_{p_1 1})$$

变量场 Y 的 q_1 个主分量的 BP 线性组合为：

$$v_1 = d_{11}b_1 + d_{21}b_2 + \cdots + d_{q_1}b_{q_1} \qquad (8.3.8)$$

其中

$$D_1 = (d_{11}\, d_{21} \cdots d_{q_1 1})$$

与典型相关基本思想一致，由线性组合构成的新的一对变量 u_1, v_1，它们之间有最大相关系数。u_1, v_1 被称为第一对 BP 典型变量，$c_{11}, c_{21}, \cdots, c_{P_1 1}$ 和 $d_{11}, d_{21}, \cdots, d_{q_1 1}$ 为相应的 BP 典型变量的荷载特征向量。

BP 典型相关系数为：

$$r_1 = \frac{1}{n} u_1, v^{1\prime} = c_1' S_{ab} d_1 \qquad (8.3.9)$$

与典型相关分析相同，选取最优化特征的主分量的典型变量是求解极值问题。按照 8.2 节中叙述的典型相关分析的一系列步骤，依次得到 $u_1, v_1, u_2, v_2 \cdots$ 对相互独立的典型变量。

8.3.2　计算步骤

归纳起来 BP 典型相关分析包括以下步骤：

(1)对变量场 X 和 Y 作标准化预处理。

(2)对标准化后的两个变量场分别进行 EOF 分析。将变量场 X 投影到前 p_1 个 EOF 上，将变量场 Y 投影到前 q_1 个 EOF 上。这样分别截取到 p_1 和 q_1 个特征值及相应的特征向量及时间系数。

(3)利用(8.3.3)和(8.3.4)式分别计算出两变量场的主分量，构造出主分量矩阵 A 和 B。

(4)分别计算主分量矩阵 A 的协方差阵 S_{aa}，主分量矩阵 B 的协方差阵 S_{bb}，两主分量矩阵的交叉协方差阵 S_{ab}。

(5)执行典型相关计算步骤中的 3～7 步。

8.3.3　计算结果分析

BPCCA 主要用于研究大尺度的气候变量场的耦合特征，尤其适用于样本量小于空间点数的变量场。它计算简便、物理含义清楚，且可以得到稳定的典型变量。因此，BPCCA

在气候诊断研究中更具实用性。但是应注意，其计算结果可能会与普通CCA略有不同。CCA以考查两个变量场的整个交叉协方差结构为出发点，BPCCA则是从两变量场主要特征的协方差结构中提取其典型相关的。因此，结果会有差别。当然，如果取两变量场所有的主分量时，即取的特征值个数与空间点数相等时，A 和 B 是方阵，可以证明，BPCCA与CCA计算结果完全一致。

计算典型变量与对应变量场之间的相关系数阵，以此分析两变量场的遥相关特征。当然，也可以由物理因子序列构成因子变量组，另一个是某一区域变量场，用BPCCA研究变量场与因子变量组之间的相互关系，检测各个因子对变量场变化的影响程度及敏感区域[7]，这也是BPCCA在气候变化成因分析中常用的一种方式。

§8.4 奇异值分解

1976年Prohaska提出将EOF分析技术直接用于两气象场的交叉相关系数场的分解计算方案[3]，旨在最大限度地分离出两场的高相关区，以此了解成对变量场之间相关系数场的空间结构及各自对相关场的贡献。其基本做法如下：对标准化的空间大小及样本量形如(8.2.1)和(8.2.2)式的变量场 X 和 Y，计算交叉相关系数：

$$s_{ij}=\frac{1}{n}\sum_{t=1}^{n}x_i(t)y_j(t)$$

$$(i=1,2,\cdots,p \quad j=1,2,\cdots,q) \qquad (8.4.1)$$

由上式得到 p 个站点的标准化时间序列与 q 个站点标准化时间序列之间的相关系数。假定 $q>p$，计算平均乘积

$$h_{lm} = \frac{1}{q}\sum_{k=1}^{q} s_{lk}s_{mk} \quad (l,m = 1,2,\cdots,p) \qquad (8.4.2)$$

由元素 h_{lm} 构成的矩阵 H 为对称阵，其特征向量形成正交系。这时，有矩阵方程：

$$(H - \lambda_k I)U = 0 \qquad (8.4.3)$$

其中 I 为单位矩阵；λ 和 U 分别为特征值和特征向量。求解方程(8.4.3)式得到特征值 λ_k 和特征向量 U。由特征向量和原相关系数矩阵定义为：

$$v_{kj} = \sum_{i=1}^{p} s_{ij}u_{ik} \qquad (8.4.4)$$

因此，原相关系数矩阵被分离为两部分的线性组合

$$s_{ij} = \sum_{k=1}^{p} v_{kj}u_{ik} \qquad (8.4.5)$$

空间分布型式 U 表示 X 场对相关系数场的贡献，另一空间分布型式 V 表示 Y 场对相关系数场的贡献。

Prohaska 将这一方案用于诊断美国月平均地面气温与北太平洋海平面气压之间的关系。1982 年徐瑞珍、张先恭参考 Prohaska 的做法分析了 500hPa6～8 月平均高度与同期中国东部地面温度场的相关[9]，之后相继见到将这一做法使用到两变量场关系分析的工作中[10]。1992 年 Bretherton 和 Wallace 等人从矩阵理论中的奇异值分解定理出发，较系统地描述了这一方法的原理和计算[11～12]，并将这种分析方法冠以“奇异值分解”(Singular Value Decomposition，SVD)的名称，并对两变量耦合场分解的几种方法作了比较。此后，SVD 在气候诊断方面的应用显著增多。同时，有关这一方法的使用条件，其结果的真实性及它与 EOF，CCA 关联的讨论也异常

活跃[13~16]。这里给出以奇异值分解定理为依据的计算格式。

8.4.1 方法概述

设有两个变量场，这里不妨称一个场为左场，由 p 个空间点构成，样本量为 n，记为矩阵形式 X。另一个场称为右场，由 q 个空间点构成，样本量亦为 n，记为 Y。X,Y 中元素均已作过标准化处理。

8.4.1.1 奇异值分解

假设两场之间交叉协方差矩阵为 $S_{p\times q}$。对任何一个 $p\times q$ 阶实非对称矩阵 S 的奇异值分解，都可以得到：

$$S = L\begin{bmatrix} \Lambda_m & 0 \\ 0 & 0 \end{bmatrix}R' \tag{8.4.6}$$

(8.4.6)式的分量形式：

$$S = \sum_{k=1}^{m}\lambda_k l_k r_k' \qquad m \leqslant \min(p,q) \tag{8.4.7}$$

这里向量 l_k 有 m 个，相互正交，称为左奇异向量，向量 r_k 有 m 个，亦相互正交，称为右奇异向量。

SVD 的目的就是要寻找两变量场的线性组合，即由左、右两场分别构造两个矩阵

$$U = L'X \tag{8.4.8a}$$

$$V = R'Y \tag{8.4.8b}$$

为了惟一地分解(8.4.6)式，令 L,R 为正交化向量的条件，即

$$LL' = I \tag{8.4.9a}$$

$$RR' = I \tag{8.4.9b}$$

同时使矩阵 U,V 之间有极大化协方差

$$\mathrm{COV}(U,V) = L'SR = \max \tag{8.4.10}$$

根据条件极值求解，可以推导出

$$\begin{aligned} S'L &= \lambda_k R \\ SR &= \lambda_k L \end{aligned} \quad (k=1,2,\cdots,m) \qquad (8.4.11)$$

(8.4.11)式写成矩阵形式，即为(8.4.6)式。非负值 λ_k 为特征值，在奇异值分解中称为奇异值，$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m > 0$。实对称矩阵 $S'S$ 的特征值和特征向量为 λ_k 和 R，SS' 的特征值和特征向量为 λ_k 和 L。两个矩阵的特征值 λ_k 是相同的，即

$$\begin{cases} (SS' - \lambda_k I)L = 0 \\ (S'S - \lambda_k I)R = 0 \end{cases} \qquad (8.4.12)$$

用 L 左乘(8.4.8a)式，并运用(8.4.9a)式导出左变量场展开式

$$X = LU \qquad (8.4.13)$$

其中 U 为左场的时间系数矩阵，记为向量形式为：

$$U(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_m(t) \end{bmatrix} \qquad (8.4.14)$$

同理，右变量场展开式

$$Y = RV \qquad (8.4.15)$$

其中 V 为右场的时间系数矩阵，记为向量形式为：

$$V(t) = \begin{bmatrix} v_1(t) \\ v_2(t) \\ \vdots \\ v_m(t) \end{bmatrix} \qquad (8.4.16)$$

由(8.4.13)和(8.4.15)式可见，SVD 相当于将左、右变量场分解为左、右奇异向量的线性组合。每一对奇异向量和相应的时间系数确定了一对 SVD 模态。

8.4.1.2 方差贡献

每对奇异向量方差贡献为：

$$SCF_k = \lambda_k^2 \Big/ \sum_{i=1}^{m} \lambda_i^2 \tag{8.4.17}$$

前 k 对奇异向量累积方差贡献为：

$$CSCF_k = \sum_{i=1}^{k} \lambda_i^2 \Big/ \sum_{i=1}^{m} \lambda_i^2 \tag{8.4.18}$$

8.4.1.3 相关系数

已由 SVD 得到时间系数矩阵 U，V，就可以定义每对奇异向量的时间系数 U 和 V 之间的相关系数

$$r_k(U,V) = \frac{\mathrm{E}[u_k(t)v_k(t)]}{\mathrm{E}[u_k(t)]^{1/2}\mathrm{E}[v_k(t)]^{1/2}} \tag{8.4.19}$$

$r_k(U,V)$表示每对奇异向量之间线性组合相关关系的密切程度，与 CCA 中的典型相关系数类似，反映的是典型变量场总体相关状况。

左变量场 X 与右奇异向量的时间系数 V 之间的相关系数为：

$$r_k(X,V) = \frac{\mathrm{E}[x_i(t)v_k(t)]}{\mathrm{E}[x_i^2(t)]^{1/2}\mathrm{E}[v_k^2(t)]^{1/2}} = \frac{\lambda_k l_k}{\mathrm{E}[v_k^2(t)]^{1/2}} \tag{8.4.20}$$

右变量场 Y 与左奇异向量的时间系数 U 之间的相关系数为：

$$r_k(Y,U) = \frac{\mathrm{E}[y_i(t)u_k(t)]}{\mathrm{E}[y_i^2(t)]^{1/2}\mathrm{E}[u_i^2(t)]^{1/2}} = \frac{\lambda_k r_k}{\mathrm{E}[u_k^2(t)]^{1/2}} \tag{8.4.21}$$

(8.4.20)和(8.4.21)式相关系数分布型代表两变量场相互关系的分布结构，显著相关区则是两变量场相互作用的关键区域。通常将 $r_k(X,V)$，$r_k(Y,U)$称为异性相关系数。

同样，可以定义同性相关系数

$$r_k(X,U) = \frac{\mathrm{E}[x_i(t)v_k(t)]}{\mathrm{E}[x_i^2(t)]^{1/2}\mathrm{E}[u_k^2(t)]^{1/2}} = \frac{\lambda_k l_k}{\mathrm{E}[u_k^2(t)]^{1/2}} \tag{8.4.22}$$

$$r_k(Y,V)=\frac{\mathrm{E}[y_i(t)v_k(t)]}{\mathrm{E}[y_i^2(t)]^{1/2}\mathrm{E}[v_k^2(t)]^{1/2}}=\frac{\lambda_k r_k}{\mathrm{E}[v_k^2(t)]^{1/2}} \tag{8.4.23}$$

由于数据是经标准化处理的，因此每对奇异向量的时间系数与该场之间的相关分布，就是该对向量的空间分布型，它们在一定程度上代表了两变量场的遥相关型。

8.4.1.4 显著性检验

从统计意义上讲，SVD 像 CCA 一样，要求样本量 n 大于两变量场的空间点数 p,q，以得到有统计意义的 SVD 模态。但是，在气候研究问题中，往往变量场的空间点数比样本量大得多，计算结果究竟是信号还是噪音需要进行显著性检验。通常采用 Monte Carlo 技术检验 SVD 模态的显著性。

假设实测资料计算出的奇异值 λ_k 均是按大小排序的，即 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$。计算它们的方差贡献：

$$C_k=\frac{\lambda_k}{\sum_{i=1}^{m}\lambda_i}$$

根据左变量场空间点数 p、右变量场空间点数 q 及其样本量 n，利用随机数发生器生成高斯分布随机序列的两个资料矩阵，进行 100 次模拟 SVD 计算。每次模拟后均用奇异值 δ_k 计算方差贡献：

$$U_k^r=\delta_k^r/\sum_{i=1}^{m}\delta_i^r \quad (k=1,2,\cdots,m \quad r=1,2,\cdots,100)$$

将 U_k^r 排序

$$U_k^1\leqslant U_k^2\leqslant\cdots\leqslant U_k^{100} \quad (k=1,2,\cdots,m)$$

如果

$$C_k>U_k^{95}$$

则认为第 k 对 SVD 模态在 95%显著性水平上是显著的。

8.4.2 计算步骤

SVD 的计算过程如图 8.1 所示。

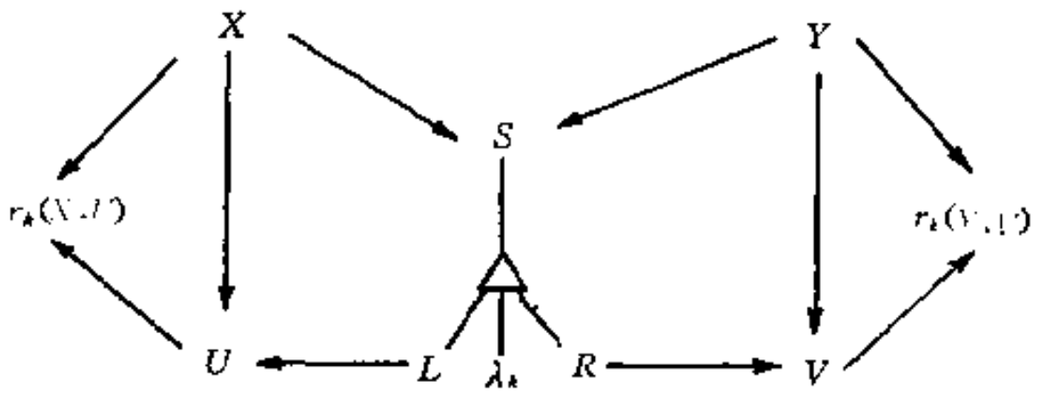


图 8.1 SVD 计算流程图

SVD 的具体计算步骤为：

(1)由左变量场 X 和右变量场 Y 计算交叉协方差阵 S。

(2)对实非对称阵作奇异值分解，得到奇异值 λ_k 及左奇异向量 L 和右奇异向量 R。

(3)根据左变量场 X 及左奇异向量 L 算出时间系数矩阵 U。由右变量场 Y 及右奇异向量 R 算出时间系数矩阵 V。

(4)利用(8.4.19)式计算奇异向量的时间系数 U 和 V 之间的相关系数。利用(8.4.20)～(8.4.23)式分别计算同性相关系数和异性相关系数。

(5)计算每对奇异向量的方差贡献及累积方差贡献。

(6)用 Monte Carlo 技术对奇异向量作显著性检验。

8.4.3 计算结果分析

对 SVD 的计算结果主要作如下的分析：

(1)从奇异向量的方差贡献及累积方差贡献了解某一对显著 SVD 模态及前几对显著 SVD 模态所占的方差比例。

(2)由奇异向量时间系数之间的相关系数 $r_k(U,V)$ 了解两变量场的显著空间分布型总体的相关程度。

(3)分析异性相关系数场,寻找一个场对另一个场相互影响的关键区。

(4)由于成对奇异向量是由两变量场的交叉协方差阵求出的,其中一个场的奇异向量对应的时间系数包容了另一个场的信息。因此,成对奇异向量的时间系数与该场的同性相关分布代表了两场耦合相关的空间结构。若两场线性相关的地域分布达到一定显著性水平,就表示两个变量在这一区域有遥相关特征。当然,由 SVD 方法得到的仅仅是一些统计事实,表征出的遥相关所包含的更深刻的学术蕴示及物理机制还需另作研究。

应用实例[8.2]:SVD 是研究两个气象变量场相关结构的诊断技术,由于计算简便,近来已被广泛应用于气候诊断研究中[17~20],得到一些有益的研究结果。在文献[19]中,取英国哈特莱气候中心 1951～1990 年夏季(6～8 月)季平均北美 2.5°～82.5°N,47.5°～157.5°W 范围 5°×5°陆地/海洋温度格点资料为左场,中国夏季(6～8 月)降水量为右场,其中 $p=391$,$q=160$,$n=40$。两变量场标准化处理后进行 SVD 计算。第一对空间分布型在统计意义上是显著的,其解释总方差的 14.75%。这对空间分布型时间系数之间的相关系数 $r_k(U,V)$ 为 0.78。表明这对空间分布型有着密切的关系。图 8.2 为中国夏季降水与夏季北美陆温的第一对空间分布型。

在中国夏季降水空间分布型(图 8.2a)中,内蒙、西北大部和华南为负相关区,但相关系数不大。东北、华北及黄河至江南的大范围地区为正相关区,高相关中心位于长江中下游地区,相关系数在 0.50 以上,达到 0.001 显著性水平。北美陆

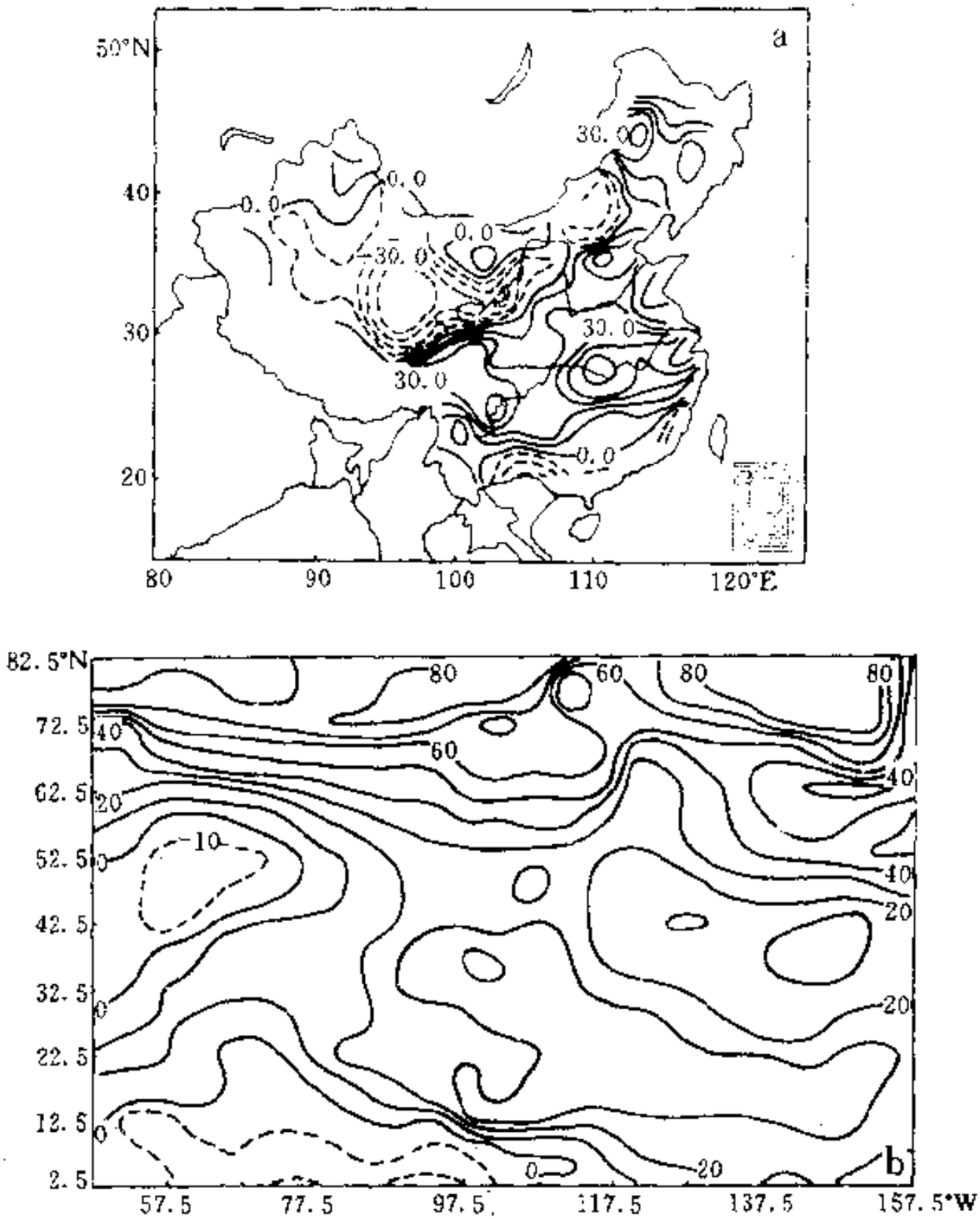


图 8.2　中国夏季降水(a)与北美夏季陆温(b)第一空间分布型

温空间分布型(图 8.2b)以正相关为主，高相关区位于 60°N 以北的高纬地区，相关系数均在 0.40 以上。70°N 以北相关系

数达 0.80 以上。这对空间分布型表明这样的遥相关特征，即当北美高纬地区的夏季陆地气温增高时，我国长江中下游地区夏季降水偏多，反之亦然。

§8.5 SVD 与 CCA 及有关问题的讨论

8.5.1 SVD 与 CCA

SVD 在大气科学领域是相对比较新的分析方法，而在社会科学领域中这一方法早已为人所熟知，并曾引起过讨论。SVD 的目的与 CCA 一样，是要寻找两组变量的线性组合。选取最优化特征的线性组合的准则是

$$\mathrm{COV}(U,V) = \max$$

即在 U,V 数学期望为 0、方差为 1 的条件下，使它们之间具有最大可能的协方差。SVD 的计算十分简便，直接对交叉协方差阵实施奇异值分解，即可得到非零的、按大小排序的奇异值及对应的左、右奇异向量。最大奇异值对应的奇异向量及时间系数被确定为最佳线性组合模态，视为分析两变量主要相关特征的依据。次大奇异值对应的奇异向量及时间系数被定为第二线性组合模态。因此，有人将 SVD 称为“典型协方差分析”[21]。

CCA 选取最优化线性组合的准则，是在 U,V 数学期望为 0、方差 1 的条件下，使典型变量在所有线性组合而成的典型变量中具有最大的相关系数，即

$$r(U,V) = \max$$

在剩余方差中，再寻找与第一对典型变量相独立的第二对典型变量，要求具有次大相关系数。依次进行下去……当变量个数较多时，计算过程比较繁杂。

可见，从统计学角度讲，CCA 的求解准则比 SVD 更合理。由于严格的求解条件，确保了前几对典型变量真实反映两变量场的主要耦合特征。SVD 没有这种保证，甚至有时可以产生虚假的相关。研究表明，特别是当空间点数大于样本量的情况下，产生虚假相关的可能性很大。可以采用两种途径解决这个问题：①避免使用小样本，且对 SVD 模态进行显著性检验；②使用 CCA 的另一种形式——BPCCA 进行分析。BPCCA 具有与 CCA 一样严格的求解条件，计算比 CCA 简便，且更适合分析空间点数大于样本量的变量场的相关结构。

8.5.2 有关 SVD 的讨论

SVD 作为寻找两组变量线性组合的方法，在社会科学研究领域早就开始使用，并且早在 70 年代初就有人对此方法的有关问题展开了讨论。1971 年 Van de Geer 在“模态匹配”范围内讨论了 SVD[22]。1982 年 Müller 在一篇未发表的手稿中进一步简述了有关问题，并将 SVD 称为“典型协方差分析”。1995 年 Newman 等人就 SVD 的限制进行了详细的论证，并提出关于 SVD 方法的警告[14]。Cheng X 和 Dunkerton 由 SVD 导出了正交旋转的 SVD 形式[23]。1996 年 Cherry 使用设计的 7 个模拟计算例子对 SVD 模态的真实性提出了疑问[21]。有关详细推导、论证可参阅上述文献。归纳起来，主要有以下几方面内容：

(1)只有在两变量场 X 和 Y 满足以下两个条件时，才能得到真实的 SVD 模态，即

①X 和 Y 两个场必须互为正交变换。

②两场之一的协方差阵为单位矩阵。否则只能得到近似的 SVD 耦合相关模态。

(2)SVD 希望得到的结果具有一定的物理意义。然而，如

何分析 SVD 模态蕴含的物理含义是十分困难的。有工作证明，SVD 分解的基础——两场交叉协方差阵不一定是所研究过程的唯一特征。至少在理论上存在无穷个具有相同交叉协方差结构、不同场内协方差结构的过程。所有过程具有相同组奇异值和奇异向量。所设计的 SVD 计算流程中，定义了异性相关系数和同性相关系数，想像由此来帮助解释 SVD 的计算结果。但是，事实上，它们只是对感兴趣的过程提供了一个连接，解释上仍存在问题，因为它们对奇异向量的时间系数求相关矩阵，不能保证使相关系数达最大。因而，无法保证分离的耦合相关结构是两变量场具有的真实特征。

(3)按 SVD 的求解条件及做法，它可以被看作是从协方差阵中寻找主成分权重的正交变换方法，或可以看作是用前 k 对空间分布型解释最大累积方差的一种方法。

(4)有推导证明，旋转的 SVD 方法可以在很大程度上改善两个场之间的线性相关，或者采用将 X 和 Y 场分别先投影到 EOF 空间，再进行 SVD，然后再转变到原空间上，以此办法来改善两个场之间的线性相关。

(5)由于 SVD 在理论上存在争议，在使用 SVD 进行气候诊断分析时，除了避免使用小样本和对 SVD 模态进行显著性检验外，在分析其结果时要格外小心，特别在做两变量场耦合相关的结论时，更要十分谨慎，以免得出虚假的结论。

参考文献

[1]程兴新，曹敏．统计计算方法．北京：北京大学出版社，1989.109～118

[2]Hotelling H. Relations between two sets of Variates, Biometrika. 1936, 28: 139—142

[3]Glahn H. R. Canonical Correlation and its relationship to discriminant analysis and multiple regression, J. Atmos. Sci. 1968, 25: 23—31

[4]Barnett T. P. Interaction of the monsoon and Pacific trade wind systems at interannual time scales, Part I: The equatorial Zone. Mon. Wea. Rev. 1983, 111: 756—773

[5]Nicholls N. The use of canonical correlation to study teleconnections, Mon. Wea. Rev. 1987, 115: 393—399

[6]Barnett T. P. and R. Preisendorfer, Origins and levels of monthly and seasonal forecast skill for united states surface air temperature determined by canonical correlation analysis. Mon. Wea. Rev. 1987, 115: 1825—1850

[7]江志红，丁裕国，金莲姬．中国近百年气温场变化成因的统计诊断分析．应用气象学报，1997，8(2)：175～185

[8]Prohaska J. A technique for analyzing the linear relationships between two meteorological fields, Mon. Wea. Rew. 1976, 104: 1345—1353

[9]徐瑞珍，张先恭．经验正交函数在两个气象场相关分析中的应用．气象学报，1982，40(1)：117～122

[10]Lanzante J. R. A rotated eigenanalysis of the correlation between 700mb heights and sea surface temperatures in the Pacific and Atlantic, Mon. Wea. Rev. 1984, 112: 2270—2280

[11]Wallace J. M. C. Smith and C. S. Bretherton Singular Value decomposition of sea-surface temperature and 500mb height anormalies, J. climate, 1992, 5: 561—576

[12]Bretherton C. S., C. Smith and J. M. Wallace. An intercomparison of methods for finding ccupled patterns in climate data, J. climate, 1992, 5: 541—562

[13]Shen S. and K. M. Lau . Biennial oscillation associated with the east asian summer monsoon and tropical sea surface tempera-

tures, J. Meteor. Soc. Japan, 1995, 73: 105—124

[14]M. A. Newman, P. D. Sardeshmakh. A caveat concerning singular value decomposition. J. climate, 1995, 8(2): 352—360

[15]施能. 气候诊断研究中的 SVD 显著性检验. 气象科技, 1996, 4: 5～6

[16]施能. 气象学中应用 SVD 方法的一些问题. 气象科技, 1997, 4: 8～12

[17]江志红, 丁裕国. 我国夏半年降水距平与北太平洋海温异常的奇异值分解法分析. 热带气象学报, 1995, 11(2): 133～138

[18]魏凤英, 张先恭. 北太平洋海表温度与中国夏季气温的耦合特征. 曹鸿兴等主编. 我国气候变化与成因研究. 北京: 气象出版社, 1996, 67～74

[19]魏凤英, 曹鸿兴. 奇异值分解及其在北美陆地气温与我国降水遥相关中的应用. 高原气象, 1997, 16(2): 174～182

[20]魏凤英. 全球海表温度变化与中国夏季降水异常分布. 应用气象学报, 1998, 9(增刊): 1～9

[21]Cherry S. Singular value decomposition analysis and canonical correlation analysis, J. Climate, 1996, 9: 2003—2009

[22]Van de Geer J. P. Introduction to multivariate analysis for the social sciences, W. H. Freeman, 1971

[23]Cheng X. and T. J. Dunkerton. Orthogonal rotation of spatial patterns derived from singular value decomposition analysis, J. Climate, 1995, 8(11): 2631—2643

第九章　最优回归预测模型

回归分析是气候预测中应用最为广泛的统计方法。它是处理随机变量之间相关关系的一种有效手段。通过对大量历史观测数据的分析、计算，建立一个变量（因变量）与若干个变量（自变量）间的多元线性回归方程。经过显著性检验，若回归效果显著，则可将所建立的回归方程用于预测。

在气候预测中应用回归分析的目的是建立方程。在建立预测方程过程中的一个重要问题是，如何从众多备选自变量中进行筛选，建立最优回归方程。所谓“最优”回归方程有两个含意。一是预报准确。希望在最终预测方程中包含尽可能多的自变量，尤其不能遗漏对因变量有显著作用的自变量。回归方程中包含的自变量越多，回归平方和就越大，剩余平方和就越小，剩余方差一般就小。二是为了应用方便，又希望预测方程中含尽量少的变量。因此，最优回归方程应包含对因变量有显著作用的自变量，而不包含不显著的变量。目前选择最优回归筛选方程的方法主要有：前向筛选法、后向筛选法、逐步筛选等。逐步筛选是气候预测中应用最普遍的方法。从一个自变量开始，按自变量对因变量作用的显著程度从大到小依次逐个引入回归方程。当先引入的变量由于后面变量的引入变得不显著时，则将其剔除。这一方法的优点是计算量较小。已有理论和实践证明[1]，上述三种方法均只能得到近似最优回归方程。欲想在某种变量选择准则下得到最优回归，就必须比较所有可能的变量组合的回归，按给定准则确定出最优回归子集。本章将介绍确定最优回归子集的具体计算方法。

在多元线性回归方程的参数估计中，最小二乘法是最常用的方法。但是，由于数据收集的局限性，使得自变量之间客观上存在近似线性关系，即存在复共线性关系。这种关系导致最小二乘法估计效果不稳定，甚至出现回归系数符号与实况相反的情况。对此，许多学者提出了改进的办法[2~3]。例如：本章将要介绍的消除自变量之间复共线性的主成分回归和特征根回归、直接降低回归系数均方误差的岭回归就是常用的办法。

§9.1 多元线性回归的基本方法

设因变量 y 与自变量 $x_1, x_2, \cdots, x_m$ 有线性关系，那么建立 y 的 m 元线性回归模型：

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \qquad (9.1.1)$$

其中 $\beta_0, \beta_1, \cdots, \beta_m$ 为回归系数；ε 是遵从正态分布 $N(0, \sigma^2)$ 的随机误差。

在实际问题中，对 y 与 $x_1, x_2, \cdots, x_m$ 作 n 次观测，即 y_t，$x_{1t}, x_{2t}, \cdots, x_{mt}$，即有：

$$y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_m x_{mt} + \varepsilon_t \qquad (9.1.2)$$

建立多元回归方程的基本方法是：

(1)由观测值确定回归系数 $\beta_0, \beta_1, \cdots, \beta_m$ 的估计 $b_0, b_1, \cdots, b_m$，得到 y_t 对 $x_{1t}, x_{2t}, \cdots, x_{mt}$ 的线性回归方程：

$$\hat{y}_t = b_0 + b_1 x_{1t} + \cdots + b_m x_{mt} + e_t \qquad (9.1.3)$$

其中 $\hat{y}_t$ 表示 y_t 的估计；e_t 是误差估计或称为残差。

(2)对回归效果进行统计检验。

(3)利用回归方程进行预报。

9.1.1 回归系数的最小二乘法估计

根据最小二乘法，要选择这样的回归系数 $b_0, b_1, \cdots, b_m$，使

$$Q = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 =$$

$$\sum_{t=1}^{n} (y_t - b_0 - b_1 x_{1t} - \cdots - b_m x_{mt})^2$$

(9.1.4)

达到极小。为此，将 Q 分别对 $b_0, b_1, \cdots, b_m$ 求偏导数，并令 $\frac{\partial Q}{\partial b_i} = 0$，经化简整理可以得到 $b_0, b_1, \cdots, b_m$，必须满足下列正规方程组：

$$\begin{cases} S_{11}b_1 + S_{12}b_2 + \cdots + S_{1m}b_m = S_{1y} \\ S_{21}b_1 + S_{22}b_2 + \cdots + S_{2m}b_m = S_{2y} \\ \quad \vdots \qquad \vdots \qquad\qquad \vdots \\ S_{m1}b_1 + S_{m2}b_2 + \cdots + S_{mm}b_m = S_{my} \end{cases} \tag{9.1.5}$$

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2, \cdots, - b_m\bar{x}_m \tag{9.1.6}$$

其中

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t \tag{9.1.7}$$

$$\bar{x}_i = \frac{1}{n}\sum_{t=1}^{n} x_{it} \quad i = 1, 2, \cdots, m \tag{9.1.8}$$

$$S_{ij} = S_{ji} = \sum_{t=1}^{n} (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j) = \sum_{t=1}^{n} x_{it}x_{jt} -$$

$$\frac{1}{n}\left(\sum_{t=1}^{n} x_{it}\right)\left(\sum_{t=1}^{n} x_{jt}\right) \quad (i = 1, 2, \cdots, m) \tag{9.1.9}$$

$$S_{iy} = \sum_{t=1}^{n} (x_{it} - \bar{x}_i)(y_t - \bar{y}) = \sum_{t=1}^{n} x_{it}y_t -$$

$$\frac{1}{n}\left(\sum_{t=1}^{n} x_{it}\right)\left(\sum_{t=1}^{n} y_t\right) \quad (i = 1, 2, \cdots, m) \tag{9.1.10}$$

解线性方程组(9.1.5),即可求得回归系数 b_i,将 b_i 代入(9.1.6)式可求出常数项 b_0。

一般情况下,用矩阵来研究多元线性回归更便利,令

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

$$\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \qquad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad b=\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix}$$

多元线性回归模型(9.1.1)式可以写为矩阵形式:

$$Y=X\beta+\varepsilon \tag{9.1.11}$$

正规方程组(9.1.5)式的矩阵形式则为:

$$(X'X)b=X'Y \tag{9.1.12}$$

因而回归系数的最小二乘法估计为:

$$b=(X'X)^{-1}X'Y \tag{9.1.13}$$

回归系数向量 b 的数学期望为:

$$\mathrm{E}(b)=\beta \tag{9.1.14}$$

回归系数向量 b 的协方差阵为:

$$\mathrm{E}[(b-\beta)(b-\beta)']=\sigma^2(X'X)^{-1} \tag{9.1.15}$$

可见,估计值 b 是参数 β 的无偏估计。

9.1.2 回归问题的统计检验

如前所述,我们是在假定 y 与 $x_1,x_2,\cdots,x_m$ 具有线性关系的条件下建立线性回归方程的。究竟 y 与 x_i 之间的线性关系是否显著?所建立的回归方程效果如何?这些需进行统计

检验来回答。

9.1.2.1　回归方程效果的检验

(1)方差分析及 F 检验。检验回归方程效果的优劣及其预测精度可以通过方差分析来实现。

将 y 的总离差平方和 S_{yy} 分解为：

$$S_{yy} = U + Q \tag{9.1.16}$$

其中

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{9.1.17}$$

$$U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{9.1.18}$$

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{9.1.19}$$

其中 U 称为回归平方和，在与误差相比意义下，它的大小反映了自变量的重要程度；Q 称为残差平方和，它的大小反映了试验误差对结果的影响。它们的自由度分别为 $f_{yy}=n-1$，$f_U=m$，$f_Q=n-m-1$。可以利用 U 和 Q 的相对大小来衡量回归效果，即检验所建回归方程是否有意义。

构造统计量：

$$F = \frac{U/m}{Q/(n-m-1)} \tag{9.1.20}$$

原假设　$H_0: \beta_1=\beta_1=\cdots=\beta_m=0$

若 H_0 成立，则认为回归方程无意义。可以证明，当 H_0 为真时，统计量 F 遵从自由度为 m 和 $n-m-1$ 的 F 分布。给定显著性水平 α，若计算值 $F>F_\alpha$，则在显著性水平 α 上拒绝原假设，认为回归方程有显著意义。

(2)复相关系数。回归方程效果的好坏亦可通过复相关系

数来进行衡量。一个变量 y 与若干变量 x_i 之间的线性关系可以由一个多元线性回归方程表示，因此，复相关系数是衡量 y 与估计值 $\hat{y}$ 之间线性关系的一个量。复相关系数表示为：

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} \qquad (9.1.21)$$

可以证明，复相关系数等于

$$R = 1 - \frac{Q}{S_{yy}} \qquad (9.1.22)$$

R 的绝对值越大，表示回归效果越好。

9.1.2.2 自变量作用的检验

上述利用方差分析和复相关系数检验回归方程的总体效果，并不能说明每个自变量 x_i 都有效果。检验各个自变量对 y 的作用是否显著，需要逐一对自变量进行检验。

原假设 H_0：

$$\beta_i = 0 \qquad (i = 1,2,\cdots,m)$$

构造统计量：

$$F_i = \frac{b_i^2/C_i}{Q(n-2)} \qquad (i = 1,2,\cdots,m) \qquad (9.1.23)$$

其中，

$$C_i = \left[\sum_{i=1}^{n}(x_{ii} - \bar{x}_i)^2\right]^{-1} \qquad (9.1.24)$$

Q 为残差平方和，由(9.1.19)式求出。统计量 F_i 遵从分子自由度为1，分母自由度为 $n-2$ 的 F 分布。若 $F_i > F_a$，则拒绝原假设，认为 x_i 在显著性水平 a 上对 y 的作用是显著的。

9.1.3 利用回归方程进行预测

将给定的样本值 $x_{1t+1}, x_{2t-1}, \cdots, x_{mt+1}$ 代入回归方程，即可得到一步预测：

$$\hat{y}_{t+1} = b_0 + b_1 x_{1t+1} + \cdots + b_m x_{mt+1} \quad (9.1.25)$$

实际使用时，应该给出 $\hat{y}_{t+1}$ 给定显著性水平的置信区间。当样本量 n 较大且 x_{it+1} 接近 $\bar{x}_i$ 时，则可以近似地认为：

$$y_{t+1} = \hat{y}_{t+1}$$

遵从 $N(0,\sigma)$。给定显著性水平 $\alpha=0.05$，则

$$P(\hat{y}_{t+1} - 1.96\sigma < y_{t+1} < \hat{y}_{t+1} + 1.96\sigma) = 0.95 \quad (9.1.26)$$

其中 σ 未知，用无偏估计量：

$$S_y = \sqrt{\frac{Q}{n-m-1}} \quad (9.1.27)$$

代替。因此，y_{t+1} 的 $\alpha=0.05$ 的置信区间为：

$$(\hat{y}_{t+1} - 1.96S_y, \hat{y}_{t+1} + 1.96S_y)$$

应用实例[9.1]：因变量 y 为长江中下游夏季(6～8 月)降水量，3 个自变量分别为：冬季(12 月～翌年 2 月)北太平洋涛动指数(x_1)、1 月太平洋地区极涡面积指数(x_2)、5 月西太平洋副高脊线(x_3)，数据见表 9.1。取 1953～1996 年观测样本。这里 $n=44, m=3$。建立夏季降水量的多元线性回归方程。

用最小二乘法求出回归系数 b_0, b_1, b_2, b_3，得到回归方程：

$$\hat{y}_t = 287.4350 + 2.4959x_{1t} - 1.946x_{2t} - 2.9008x_{3t} \quad (9.1.28)$$

回归平方和 $U=201\ 569.2$，残差平方和 $Q=641\ 325.3$，复相关系数 $R=0.489\ 0$。统计量值 $F=4.190\ 7$，当 $\alpha=0.05$ 时，$F_{0.05}(3,40)=2.84$，$F>F_{0.05}$，因此认为，线性回归方程(9.1.28)式在 $\alpha=0.05$ 显著性水平上具有显著性。

表 9.1　长江中下游夏季降水量试验数据

	203	250	220	212	212	172	225	220	217	238
	199	195	222	224	229	216	201	199	227	195
x_1	192	215	203	213	221	216	228	208	202	239
	204	240	209	234	247	234	222	202	211	212
	241	220	226	219						
	109	−6	−61	−272	−180	33	−64	79	83	−85
	−78	97	−69	5	47	−102	−91	86	−96	−133
x_2	1	−66	44	2	158	81	−8	−33	33	−108
	134	8	−18	138	166	120	−16	−13	−48	91
	8	16	41	29						
	105	95	90	100	149	115	108	108	99	105
	105	109	140	133	149	149	105	109	140	134
x_3	105	149	113	149	105	100	99	100	105	100
	99	110	105	145	90	105	120	115	105	90
	90	90	90	105						
	446	1000	576	496	477	349	367	398	373	570
	378	416	477	364	342	388	778	526	395	370
y	527	497	548	407	605	269	458	775	371	550
	662	512	370	481	558	478	606	433	552	466
	688	522	601	705						

根据(9.1.23)式可知，$F_3=7.1742$，$F_1=4.1154$，$F_2=0.8062$，而 $F_{0.05}(1,42)=4.07$。$F_3>F_{0.05}$，$F_1>F_{0.05}$，$F_2<F_{0.05}$，因此认为，x_1 和 x_3 对方程的作用是显著的，x_2 是不显著的。

§9.2　最优子集回归

在气候预测工作中，逐步回归算法的应用是十分广泛的，它的最大优势是计算量及内存需求小。但是，从实践和理论上可以证明，在给定的自变量条件下，并不能获得一个最优回归

方程。另外，选入和剔除自变量均基于统计检验，显著性水平 α 的选择具有任意性。很难从理论上以任何概率保证所筛选的自变量的显著性[4]。特别当引入方程中的自变量很多时，所建回归方程很容易通过回归效果的 F 检验或复相关系数检验，使检验流于形式。在计算机高速发展的今天，计算量及内存容量已不再是主要矛盾。因此，用最优子集回归替代逐步回归应成为一种趋势。最优子集回归(Optimal Subset Regression，OSR)是从自变量所有可能的子集回归中以某种准则确定出一个最优回归方程的方法。

9.2.1 方法

所有可能的回归方法是由 Garside 在 1965 年提出来的。之后，Furnival 等人对这一方法进行了完善和修改[5,6]。

假设考虑有 m 个自变量的回归，由于每个变量有在方程内或不在方程内两种状态存在，因此，m 个自变量的所有可能的变量子集就有 2^m 个。除去方程一个变量也不含的空集外，实际有 2^m-1 个变量子集。可见，计算量是随自变量个数呈指数增长的。当 m 较大时，变量个数非常之大，例如：$m=10$ 时，则有 $2^{10}-1=1023$ 个变量子集，计算量和内存量是非常之大的。对于 70 年代以前的计算条件，这种计算是无法想象的。即使在计算机高速发展的今天，当 m 很大时，计算也相当困难。因此，有学者设计了计算所有可能回归的最佳算法。

建立最优回归预测方程就是要从所有可能的回归中确定出一个效果最优的子集回归。具体做法是：按照一定的目的和要求，选定一种变量选择准则 s，每一个子集回归都能算出一个 s 值，共有 2^m-1 个 s 值(由 Furnival 设计的算法，并不需要 2^m-1 个回归)。s 越小(或越大)对应的回归方程效果就越好。在 2^m-1 个子集中，最小(或最大)值对应的回归就为最优

子集回归。

9.2.2 计算

为解决所有可能回归计算量与内存问题，统计学者相继设计出各种算法。其中 Furnival 设计出几种计算所有可能回归的方式：字典式、二进制式、自然式和家族式。表 9.2 给出自

表 9.2 计算所有可能回归的顺序($m=4$)

序号	字典式	二进制式	自然式	家族式
1	1	1	1	1
2	1 2	2	2	2
3	1 2 3	1 2	3	3
4	1 2 3 4	3	4	4
5	1 2 4	1 3	1 2	1 2
6	1 3	2 3	1 3	1 3
7	1 3 4	1 2 3	1 4	2 3
8	1 4	4	2 3	1 2 3
9	2	1 4	2 4	1 4
10	2 3	2 4	3 4	2 4
11	2 3 4	1 2 4	1 2 3	3 4
12	2 4	3 4	1 2 4	1 2 4
13	3	1 3 4	1 3 4	1 3 4
14	3 4	2 3 4	2 3 4	2 3 4
15	4	1 2 3 4	1 2 3 4	1 2 3 4

变量个数 $m=4$ 时，按 4 种方式计算所有可能回归的顺序。由表看出，这几种计算方式的共同特点是每种变量子集只出现过 1 次，就可以获得所有可能的回归。然后根据给定的准则，选择最优回归方程。但是，当 m 很大时，计算量相当可观。因此，Furnival 和 Wilson 又设计出不计算所有可能回归，而求最优子集回归的“分支定界法”。具体计算思路是：将 m 个自变量按某种原则分成若干组。设 A，B 为其中两组，若它们的残差平方和 $Q_A \leqslant Q_B$，则 B 变量组的所有可能的子集回归的残差平方和不会再比 Q_A 小，因此 B 变量组的所有可能的子集回归不需要计算。将 Q_A 视为一个界，凡是残差平方和比其

大的变量组，其子集回归不全是最优的，不必计算。用这种构思计算最优子集回归可以大大减少计算量。

9.2.3 选择最优子集回归的准则

上面提到逐步回归模型的确定是使用基于 F 检验的方法，对大型回归问题，F 临界值不好确定，F 值取得太大，方程中变量个数过少；F 值取得太小，又使得大批变量进入方程，不符合要求。因此，选择合适的最优子集回归的识别准则，是建立最优回归预测模型的一个重要问题。不同的目的可以选择不同的识别准则。这里介绍几种着眼于预测的识别准则。

9.2.3.1 平均残差平方和

设 k 为任一子集回归中的自变量个数，相应的残差平方和为：

$$Q_k = \sum_{t=1}^{n} (y_t - b_1 x_{1t} - b_2 x_{2t} - \cdots - b_k x_{kt})^2 \quad (9.2.1)$$

那么，平均残差平方和定义为：

$$MQ_k = \frac{Q_k}{n-k} \quad (9.2.2)$$

当 k 较小时，残差平方和 Q_k 随着 k 的增加而减小。一旦 k 增加到一定程度，Q_k 不会明显减小，体现了对自变量个数过多所实施的调整。依这一准则，按 MQ_k 越小越好为标准，选择回归子集。

9.2.3.2 C_p-准则

按照 C_p-准则的原意，下标 p 代表含常数项在内的子集回归中自变量个数。这里仍用 k 表示任一子集回归中自变量个数。定义 C_p 统计量为：

$$C_p = \frac{Q_k}{\hat{\sigma}^2} - (n - 2k) \quad (9.2.3)$$

其中 $\hat{\sigma}^2$ 为子集回归的方差

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (9.2.4)$$

的估计。

从 C_p 统计量的性质可以推导出，C_p 越小越好。因此，依 C_p-准则可以选择出最优子集回归。

9.2.3.3 预测残差平方和准则

顾名思义，预测平方和准则是从预测观点出发的。但是，在计算预测偏差时，它与其它预测统计量的计算方法不同。它是用独立样本即建立回归时未曾用过的样本，来计算预测偏差，期望以此准则选择出较好预测效果的子集回归。

在变量 y 及子集回归中 k 个自变量 x_i 中剔除第 j 个观测样本 y_j 和 x_{ij}，得到回归模型：

$$y' = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (9.2.5)$$

利用最小二乘法可以得到 β_i 的估计 $b_0, b_1, \cdots, b_k$，以此作出被剔除的第 j 个样本的预测值 $\hat{y}_j'$，计算预测残差：

$$d_j = y_j - \hat{y}_j' \qquad (9.2.6)$$

再剔除另一组观测样本……这样依次对 n 个样本都轮作一遍，得到 $d_1, d_2, \cdots, d_n$，其平方和为：

$$PRESS_k = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i') \qquad (9.2.7)$$

称 $PRESS_k$ 为预测残差平方和的统计量。以 $PRESS_k$ 达到最小为准则，选择最优子集回归。

在实际使用时，可以利用完整的观测样本所计算的回归结果来计算统计量，避免要计算 n 个方程的麻烦。$PRESS_k$ 简化计算公式为：

$$PRESS_k = \sum_{i=1}^{n} \left(\frac{q_i}{1 - S_{ii}}\right) \tag{9.2.8}$$

其中 q_i 为一般残差，即

$$q_i = |y_i - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki}| \tag{9.2.9}$$

S_{ii} 为最小二乘法计算过程中矩阵 $S = X(X'X)^{-1}X'$ 中的对角元素。

9.2.3.4 CSC 准则

CSC 准则是针对气候预测特点提出的一种考虑数量和趋势预测效果的双评分准则(Couple Score Criterion，CSC)。它的具体推导我们将在第十章有关节中给出。这里仅给出使用选择最优子集回归时的计算形式。

设 k 为任一子集回归中自变量个数，CSC_k 定义为：

$$CSC_k = S_1 + S_2 \tag{9.2.10}$$

其中

$$S_1 = nR^2 = n\left(1 - \frac{Q_k}{Q_y}\right) \tag{9.2.11}$$

式中 Q_k 为残差平方和；Q_y 为气候学预报。

$$Q_y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{9.2.12}$$

$$S_2 = 2I = 2\Big[\sum_{i=1}^{I}\sum_{j=1}^{I} n_{ij}\ln n_{ij} + n\ln n - \Big(\sum_{i=1}^{I} n_{i.}\ln n_{i.} + \sum_{j=1}^{I} n_{.j}\ln n_{.j}\Big)\Big] \tag{9.2.13}$$

式中 I 为预报趋势类别数，n_{ij} 为 i 类事件与 j 类估计事件的列联表中的个数，其中

$$\begin{aligned} n_{.j} &= \sum_{i=1}^{I} n_{ij} \\ n_{i.} &= \sum_{j=1}^{I} n_{ij} \end{aligned} \tag{9.2.14}$$

以 CSC_k 达到最大为准则选择最优子集回归。

应用实例[9.2]：因变量 y 为长江中下游夏季(6～8 月)降水量。选取 10 个自变量：前期春季(前一年 3～5 月)、夏季(前一年 6～8 月)、秋季(前一年 9～11 月)和冬季(前一年 12 月～当年 2 月)平均赤道东太平洋海温(x_1，x_2，x_3，x_4)，上述四季南方涛动指数(x_5，x_6，x_7，x_8)，冬季北太平洋涛动指数(x_9)和 1 月太平洋地区极涡面积指数(x_{10})。观测样本量取为 1953～1996 年。这里 $n=44$，$m=10$。利用 Furnial—wilson 设计的"分支定界法"计算所有可能的子集回归，用 CSC 准则确定最优子集回归作为预测方程，并作 1997～1998 年两年长江中下游夏季降水量预报。

表 9.3 给出所有可能的最优子集及其对应的复相关系数

表 9.3 所有可能的最优子集

k	最优子集	R	CSC
1	x_9	0.31	7.80
2	x_4 x_9	0.38	14.67
3	x_2 x_3 x_9	0.43	24.60
4	x_2 x_3 x_9 x_{10}	0.48	26.30
5	x_2 x_3 x_6 x_9 x_{10}	0.51	31.49
6	x_1 x_2 x_3 x_6 x_9 x_{10}	0.52	21.83
7	x_1 x_2 x_3 x_6 x_8 x_9 x_{10}	0.52	23.61
8	x_1 x_2 x_3 x_4 x_6 x_8 x_9 x_{10}	0.54	25.03
9	x_1 x_2 x_3 x_4 x_6 x_7 x_8 x_9 x_{10}	0.54	23.55
10	x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10}	0.54	23.26

及 CSC 值。由表看出，由 5 个自变量组成的子集回归 CSC 值达到最大，因此为最优。其回归方程为：

$$\hat{y}_t = -228.673 - 146.450x_{2t} + 232.000x_{3t} + 55.481x_{6t} + 3.404x_{9t} - 0.325x_{10t} \qquad (9.2.15)$$

将 x_{2t}，x_{3t}，x_{6t}，x_{9t}和 x_{10t}1953～1996 年观测值代入(9.2.15)，即可得到 1953～1996 年夏季降水量的回归拟合值。拟合均方

根误差 $RMSE=119.4$mm。将上述 5 个自变量 1997 和 1998 年的观测值代入(9.2.15),算出这两年夏季降水量的预测值分别为 664.22 和 802.29mm,实况为 556.00 和 779.00mm。长江中下游夏季多年平均降水量为 502mm,从数量上衡量 1997 年误差较大,1998 年误差较小,但从趋势上衡量预报均是正确的。

§9.3 主成分回归

由于多元线性回归模型通常采用最小二乘法估计其参数,因此也称为最小二乘回归。最小二乘法对于有些情况就不适用。在自变量之间存在近似线性关系,即存在复共线性时,回归的正规方程组(9.1.5)出现严重病态,导致回归方程极不稳定。如果方程中自变量相互无关,它们仅独立地对因变量有影响,这时正规方程组(9.1.5)的系数矩阵为对角矩阵,给计算带来了便利,且由自变量间的复共线性造成的问题就不存在了。在实际问题中,自变量间并非一定是相互无关的,常常需要人为地筛选或构造。对于具有多个自变量的线性回归问题,可以构造一些潜变量作为新的自变量。这些潜变量是由原有自变量进行线性变换得到的,且可以反映出原有变量所蕴涵的基本信息。主成分分析就可以达到这种目的。利用主成分分析从多元随机变量的观测样本矩阵中提取主成分,它们是原变量的线性组合且相互正交。利用某种判据选取前几项方差较大的主成分,略去方差较小的一些主成分。这样不仅保留了大部分原有信息,又消除了复共线性,克服了最小二乘回归的缺点。这种利用主成分作为新自变量进行回归的方法是 W.F. Massy 在 1965 年提出来的[7],被称为主成分回归

(Principal Component Regression，PCR)。由于这里回归系数估计值的数学期望不再等于待估系数，因此不再是无偏估计，而称为有偏估计。应当强调的是，对于有偏估计仅仅是在存在复共线性时，才优于通常的最小二乘回归无偏估计。因此，在建立回归方程时，需要先研究自变量间是否存在复共线性。若存在复共线性才使用有偏估计，否则还是利用最小二乘无偏估计，毕竟它具有坚实的理论基础和应用实践。

9.3.1 复共线性的诊断

由 $x_1, x_2, \cdots, x_m$ 标准化处理后的数据构成一个自变量矩阵，记为 X。诊断自变量阵 X 是否存在复共线性，可以采用以下两种简便的方法。

9.3.1.1 特征根法

若 $X'X$ 至少有一个特征根近似 0，则矩阵 X 至少存在一个复共线性关系。假设特征根 $\lambda_{p+1}, \lambda_{p+2}, \cdots, \lambda_m \approx 0$，则与它们对应的标准正交化特征向量为 $v_{1p+1}, v_{2p+2}, \cdots, v_{mm}$。若存在复共线性，则有：

$$v_{1i}x_1 + v_{2i}x_2 + \cdots + v_{pi}x_m \approx 0$$
$$(i = p+1, p+2, \cdots, m) \qquad (9.3.1)$$

这一方法的缺陷是没有给出一个定量的标准。

9.3.1.2 条件数

用条件数来判断是否存在复共线性及复共线性的严重程度。假设 $\lambda_p \approx 0$，则条件数定义为：

$$k = \frac{\lambda_1}{\lambda_p} \qquad (9.3.2)$$

若 $0 < k < 100$ 则认为不存在复共线性；若 $100 \leqslant k \leqslant 1000$，则认为存在较强复共线性；若 $k > 1000$，则认为存在严重的复共线性。

9.3.2 方法概述

多元线性回归模型的矩阵形式为：

$$Y = X\beta + \varepsilon \tag{9.3.3}$$

其中 X 为 $n \times m$ 维的自变量观测数据矩阵，这里假设已实施标准化处理。从矩阵 X 中提取 m 个自变量 $x_1, x_2, \cdots, x_m$ 的样本主成分矩阵。做法与第七章中介绍的经验正交函数分解类似。

设 $A = X'X$ 的 m 个特征根为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_m > 0$，其相应的特征向量为 $v_1, v_2, \cdots, v_m$，它们组成了正交矩阵 V。X 与主成分矩阵 T 的关系表示为：

$$X = TV' \text{ 或 } T = XV \tag{9.3.4}$$

主成分 T 是原自变量的线性组合：

$$t_i = v_{1i}x_1 + v_{2i}x_2 + \cdots + v_{mi}x_m \quad (i = 1, 2, \cdots, m) \tag{9.3.5}$$

将(9.3.4)式代入(9.3.3)式，回归模型为：

$$Y = TV'\beta + \varepsilon = T\alpha + \varepsilon \tag{9.3.6}$$

$$\alpha = V'\beta \text{ 或 } \beta = V\alpha \tag{9.3.7}$$

由于

$$T'T = \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix} \tag{9.3.8}$$

由(9.3.8)和(9.3.6)式，可以求出 α 的估计：

$$\hat{\alpha} = (T'T)^{-1}T'Y = \Lambda^{-1}T'Y \tag{9.3.9}$$

那么，同时可以得到 β 的估计：

$$\hat{\beta} = V\ \Lambda^{-1}T'Y \tag{9.3.10}$$

如果原自变量存在复共线性，使其正规方程组出现病态，且出现 $\lambda_{p-1}, \lambda_{p-2}, \cdots, \lambda_m$ 接近 0 时，同时 $t_{p+1}, t_{p-2}, \cdots, t_m$ 取值接近 0，就将它们略去。主成分 $t_1, t_2, \cdots, t_p$ 几乎可以反映原变量

$x_1, x_2, \cdots, x_m$ 的所有信息。将前 p 个主成分矩阵记作 T_1，回归模型则表示为：

$$Y = T_1 a_1 + \varepsilon \tag{9.3.11}$$

p 维参数 $a_1 = (a_1, a_2, \cdots, a_p)'$ 的估计为：

$$\hat{a}_1 = (T_1' T_1)^{-1} T_1' Y = \Lambda_1^{-1} T_1' Y \tag{9.3.12}$$

β 的主成分估计则为：

$$\hat{\beta}_c = V_1 \hat{a}_1 = \hat{a}_1 v_1 + \hat{a}_2 v_2 + \cdots + \hat{a}_p v_p \tag{9.3.13}$$

其中 V_1 为 $m \times p$ 维矩阵。

$\hat{\beta}_c$ 的第 i 个分量为：

$$\hat{\beta}_{ic} = \hat{a}_1 v_{i1} + \hat{a}_2 v_{i2} + \cdots + \hat{a}_p v_{ip} \tag{9.3.14}$$

那么，主成分回归方程表示为：

$$\hat{y} = \hat{a}_1 t_1 + \hat{a}_2 t_2 + \cdots + \hat{a}_p t_p \tag{9.3.15}$$

至于所保留的主成分个数 p 的确定，可以参考第七章 7.3 节中旋转经验正交函数分解中旋转经验正交函数个数的确定办法。

将(9.3.5)式代入(9.3.15)式变为：

$$\begin{aligned}\hat{y} = {} & \hat{a}_1 (v_{11} x_1 + v_{21} x_2 + \cdots + v_{m1} x_m) + \hat{a}_2 (v_{12} x_1 + \\ & v_{22} x_2 + \cdots + v_m x_m) + \cdots + \hat{a}_p (v_{1p} x_1 + v_{2p} x_2 + \\ & \cdots + v_{mp} x_m) = (\hat{a}_1 v_{11} + \hat{a}_2 v_{12} + \cdots + \hat{a}_p v_{1p}) + \\ & (\hat{a}_1 v_{21} + \hat{a}_2 v_{22} + \cdots + \hat{a}_p v_{2p}) x_2 + \cdots + (\hat{a}_1 v_{m1} + \\ & \hat{a}_2 v_{m2} + \cdots + \hat{a}_p v_{mp}) x_m \end{aligned} \tag{9.3.16}$$

将(9.3.14)式代入原回归方程(9.3.3)，就可以得到原变量的主成分回归方程：

$$\hat{y} = \hat{\beta}_{1C} x_1 + \hat{\beta}_{2C} x_2 + \cdots + \hat{\beta}_{mC} x_m \tag{9.3.17}$$

9.3.3 计算步骤

主成分回归的计算步骤可以简单地表述为：

(1)对原自变量进行标准化处理，得到矩阵 X。

(2)求协方差矩阵 $X'X$。由于数据经标准化处理，因此得到的是相关矩阵。

(3)求解相关矩阵的特征根及相应特征向量。

(4)利用(9.3.4)式求出 p 个主成分 $t_1,t_2,\cdots,t_p$。

(5)利用(9.3.12)式求出系数 $\alpha_1,\alpha_2,\cdots,\alpha_p$ 的估计值。

(6)利用(9.3.14)式求出系数 β 的主成分估计值 β_{1C}，$\beta_{2C},\cdots,\beta_{mC}$，并将它们代入(9.3.17)式，即可得到原变量的主成分回归方程。

§9.4 特征根回归

主成分回归仅从原自变量的样本数据中提取主成分，没有考虑自变量与因变量 y 的关系。作为主成分回归的推广形式，Webster 等人提出了特征根回归[8](Latent Root Regression，LRR)，将因变量也考虑进去了。同样也是从原有数据中提取相互正交的主成分，从而在消去原自变量复共线性的同时，也使所建立的回归方程能够表征自变量与因变量之间的相关关系。

9.4.1 回归系数的特征根估计

假设 X 是由因变量 y 和因变量 $x_1,x_2,\cdots,x_m$ 标准化处理后的数据矩阵。可以证明，由 $X'X$ 的特征根 λ_i 和 $v_i(i=0,1,\cdots,m)$可以表示回归系数 β_i 的最小二乘估计 b_j 可以表示为：

$$b_j=-S_{yy}\sum_{i=0}^{m}v_{ji}w_i \quad (j=1,2,\cdots,m) \qquad (9.4.1)$$

其中

$$w_i = \frac{v_{0i}}{\lambda_i \sum_{j=0}^{m} (v_{0j}^2/\lambda_j)} \quad (i = 0,1,\cdots,m) \qquad (9.4.2)$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (9.4.3)$$

现在假设 $\lambda_i \approx 0, v_{0i} \approx 0, , i = 0,1,\cdots,p-1$。那么就将它们略去，这样(9.4.1)式就由剩下 $p, p+1, \cdots, m$ 项来表示，即

$$\bar{b}_j = -S_{yy} \sum_{i=p}^{m} v_{ji} w_i \qquad (9.4.4)$$

其中，

$$w_i = \frac{v_{0i}}{\lambda_i \sum_{j=p}^{m} (v_{0j}^2/\lambda_j)} \quad (i = p, p+1,\cdots,m) \qquad (9.4.5)$$

9.4.2 计算步骤

建立特征根回归方程的计算步骤如下：

(1)对给定的因变量 y 和自变量 $x_1, x_2, \cdots, x_m$ 进行标准化处理。将处理后的因变量放在前面，自变量放在后面，构成一个数据矩阵，记为 X。

(2)计算协方差矩阵 $X'X$，得到增广相关矩阵。

(3)求出增广相关矩阵的特征根 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 及对应的特征向量 $v_{oi}, v_{1i}, \cdots, v_{mi}, i = 0,1,\cdots,m$。

(4)将同时都很接近 0 的 λ_i, v_{0i} 去掉。在实际操作时，限定 $\lambda_i \leqslant 0.05, |v_{0i}| \leqslant 0.10$ 就认为它们近似等于 0。

(5)应用(9.4.4)和(9.4.5)式计算回归系数的特征根估计。

(6)建立特征根回归方程。

应用实例[9.3]：建立长春 6～8 月平均气温的特征根回归。取 1961～1995 年 35 年观测资料。选取 7 个自变量：

x_1:前一年年平均太阳黑子相对数;

x_2:前一年全球二氧化碳含量;

x_3:前一年赤道东太平洋秋季(9～11 月)海温;

x_4:赤道东太平洋冬季(前一年 12 月～当年 2 月)海温;

x_5:当年长春 3～5 月降水总量;

x_6:前一年秋季南方涛动指数;

x_7:冬季南方涛动指数。

具体数据见表 9.4。这里 $n=35, m=7$。

首先检测自变量是否存在复共线性。将 $x_1, x_2, \cdots, x_7$ 进行标准化处理构成自变量矩阵 X, $X'X$ 的 7 个特征根分别为:

$\lambda_2=237.9285$; $\lambda_6=3.7154$; $\lambda_7=1.5646$;

$\lambda_3=1.3016$; $\lambda_5=0.2995$; $\lambda_1=0.1671$;

$\lambda_4=0.023\ 36$。

由于 $\lambda_4\approx 0$,因此条件数为:

$$k=\frac{\lambda_1}{\lambda_4}=10\ 185.2954$$

$k>1000$,所以认为 X 存在严重的复共线性。由于最小特征根 $\lambda_4\approx 0$,按(9.3.1)式以特征向量 v_{i4} 为系数的关系为:

$$-0.6267x_1-0.6263x_2-0.6199x_3-0.0199x_4-$$
$$0.6170x_5-0.6146x_6-0.6255x_7\approx 0$$

就是一个复共线性关系。可以看出,x_4 相应的系数非常小。

增广相关矩阵的特征根及对应的特征向量见表 9.5。从表 9.5 中看出,$\lambda_0=0.0284<0.05$,同时 $|v_{00}|=0.0286<0.10$,故略去。其余特征根均大于 0.05。因此,(9.4.5)和(9.4.6)式中的求和从 $p=1$ 开始计算回归系数的特征根估计。为了比较,用表 9.4 的数据计算了通常的最小二乘的多元

表 9.4　长春 6～8 月平均气温试验数据

	22.3	21.9	22.2	20.8	21.3	21.4	21.7
	21.3	20.3	21.9	21.0	20.9	21.6	21.4
y	21.9	20.7	21.5	22.1	21.2	21.4	21.3
	23.1	20.9	21.9	21.3	21.1	21.3	22.6
	21.4	21.5	21.7	21.1	21.2	23.8	21.8
	112	54	38	28	10	9	47
	94	105	106	105	67	69	38
x_1	35	16	13	28	93	155	155
	141	110	67	46	18	13	29
	100	158	142	163	105	59	36
	317.02	317.74	318.63	319.13	319.69	320.41	321.09
	321.90	322.72	324.21	325.51	326.48	327.60	329.82
x_2	330.41	331.01	332.06	333.62	335.19	336.54	338.40
	339.46	340.76	342.76	344.34	345.65	346.80	348.56
	351.30	352.71	353.99	355.45	356.28	356.98	358.78
	0.21	−0.21	−0.33	0.62	−0.85	1.04	−0.32
	−0.72	0.24	0.71	−0.53	−0.61	1.46	−0.87
x_3	−0.45	−0.94	0.88	0.22	−0.30	0.19	−0.10
	−0.25	0.99	0.34	−0.42	−0.29	0.51	1.32
	−1.10	−0.30	0.12	0.97	0.52	0.78	1.03
	0.23	−0.26	−0.33	0.62	−0.28	1.05	−0.40
	−1.01	0.47	0.80	−0.94	−0.44	1.05	−1.29
x_4	−0.38	−1.02	0.65	0.19	−0.10	0.32	−0.21
	−0.04	1.74	0.21	−0.67	−0.22	0.89	0.80
	−1.15	−0.15	0.11	1.28	0.46	0.45	1.04
	68	68	22	31	45	35	100
	106	76	57	58	72	98	79
x_5	46	136	63	80	55	127	98
	61	166	54	54	88	52	98
	36	149	65	84	42	71	120
	0.40	0.00	0.60	−1.10	0.90	−1.50	−0.30
	−0.10	−0.40	−0.90	1.30	1.30	−1.10	1.60
x_6	0.60	1.80	−0.20	−1.30	−0.30	−0.30	−0.50
	−0.10	−2.50	0.40	−0.10	−0.30	−0.50	−0.70
	1.80	0.30	−0.50	−1.40	−0.90	−0.80	−1.40
	0.40	1.00	0.50	−0.80	−0.30	−0.80	1.00
	0.20	−1.00	−0.90	1.40	0.40	−1.10	2.20
x_7	−0.10	1.80	−0.10	−0.70	−0.10	−0.20	−0.20
	0.60	−3.90	0.20	0.00	−0.20	−1.50	−0.60
	1.40	−1.10	0.00	−2.40	−1.10	−0.10	−0.80

线性回归。计算结果列于表 9.6。从表中看出，与最小二乘回归相比，特征根回归系数的数值改变了不少，而且 b_4 的符号由正改变为负。表明在两种回归方法中，自变量 x_4 对 y 的作用完全相反。这一实例可以进一步证明，如果自变量矩阵存在严重的复共线性，则可能导致最小二乘回归系数符号与实况相反。这时需要用有偏估计方法。最小二乘回归系数 b_4 为负，其含意是，前一年秋季赤道东太平洋海水温度高，则长春夏季气温亦高。而大量研究工作的结论恰好相反[9]。特征根估计将 b_4 的符号变为负号更符合实际。

表 9.5　增广相关矩阵的特征根及特征向量

i	0	1	2	3	4	5	6	7
λ_i	0.0284	0.1483	0.2203	0.6350	0.6917	1.1787	1.4208	3.6767
v_{0i}	−0.0286	−0.0071	0.0537	−0.5446	0.7365	0.0138	−0.1175	0.3784
v_{1i}	−0.1958	0.0396	0.0873	0.1075	−0.3257	−0.0440	−0.7402	0.5339
v_{2i}	−0.2491	0.1304	0.0555	0.6351	0.2602	−0.0417	0.4253	0.5140
v_{3i}	0.4193	0.3236	−0.7338	0.0793	0.0354	0.3776	−0.1051	0.1408
v_{4i}	−0.1171	0.6739	−0.1502	−0.0346	−0.0578	−0.7001	−0.2595	−0.1198
v_{5i}	0.7981	0.1767	0.5073	0.0484	−0.0698	−0.1205	0.0624	0.2209
v_{6i}	−0.2638	0.6175	0.3527	−0.2410	−0.1677	0.5690	0.0998	−0.0686
v_{7i}	0.0515	0.1001	0.2089	0.4696	0.4962	0.1588	−0.4814	−0.4690

表 9.6　回归系数

回归系数	b_1	b_2	b_3	b_4	b_5	b_6	b_7
特征根回归	0.2584	0.9826	1.9799	−0.3866	−0.5346	−1.2385	0.2805
最小二乘回归	0.0008	0.0143	0.1128	0.3769	−0.0025	−0.2532	0.5614

§9.5　岭回归

岭回归(Ridge Regression，RR)是 Hoerl 和 Kennard 在 1970 年首先提出来的[10]。它也是一种有偏估计，旨在克服自变量之间存在的复共线性。

9.5.1　方法概述

由 9.1 可知，最小二乘估计为：

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (9.5.1)$$

对于多元线性回归模型中的回归系数的岭估计为：

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'Y \qquad (9.5.2)$$

其中 k 为任一正常数；I 为单位矩阵。由于是用标准化处理过的数据，因此 $X'X$ 为相关矩阵。比较(9.5.1)和(9.5.2)式可见，岭估计是在(9.5.1)式中相关矩阵的对角线元素上加了一个常数 k，其余不变。事实上，是人为地设置了每个变量的变化范围。因此，形象地将 k 称为岭参数。假设 $X'X$ 的特征根为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_m > 0$，$(X'X + kI)$ 的特征根就为 $\lambda_1 + k, \lambda_2 + k, \cdots, \lambda_m + k$，使近似 0 的程度得到改善。从直观上想象，这种做法可以消除变量矩阵的复共线性，得到优于最小二乘估计的岭回归估计。

9.5.2　岭参数的确定

由于 k 是一任意正常数，因此 k 取值不同，得到的岭估计也不同。因此，在实际应用中，k 值的确定十分关键。近年来，陆续提出了一些确定 k 值的方法，但目前还没有一种公认的最优方法。下面给出几种常用的方法。

9.5.2.1　岭迹法

为确定 k，将岭回归估计 $b_i(k), i = 1, 2, \cdots, m$ 作为 k 的函数在平面直角坐标系上画岭迹图，k 取值范围为 $0 < k < \infty$。岭迹图反映了各自变量岭回归系数 $b(k)$ 随 k 的变化，并能直接比较系数之间的相互作用。根据岭迹图确定 k 值的选取。例如：图 9.1a 所示的岭迹曲线很不稳定，表示最小二乘估计没有很好地反映数据的实际情况。如果岭迹曲线比较平稳(图 9.1b)，表明最小二乘估计正常。如果岭迹介于二者之间，恰

当选择 k 值，用岭回归替代最小二乘回归。

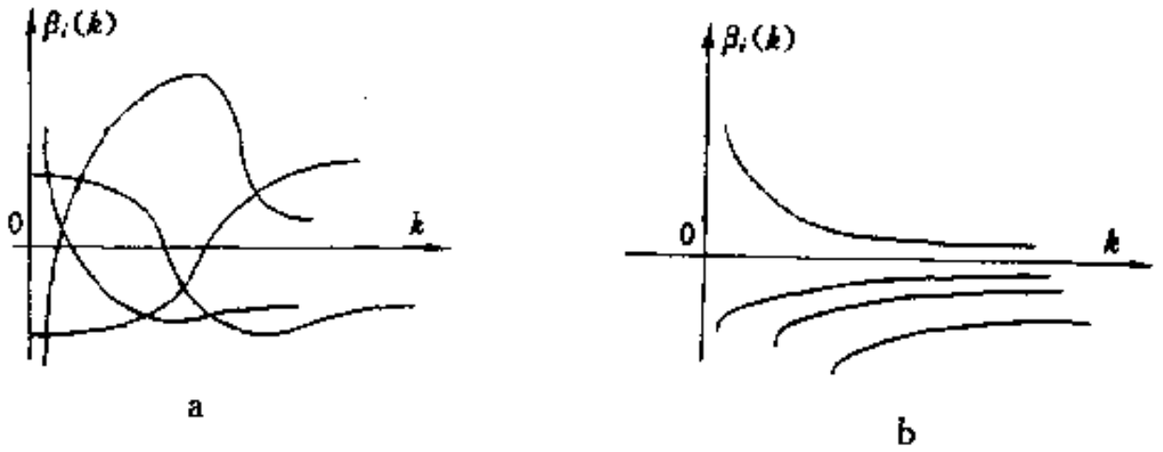


图 9.1　岭迹图

根据岭迹图选择 k 值的原则是：

(1)回归系数的岭估计基本稳定。

(2)改变最小二乘估计回归系数符号不合理现象。

(3)回归系数不出现不合理的绝对值。

(4)残差平方和增加不大。

由于资料矩阵已经过标准化，可以直接比较岭回归系数的大小。分析岭迹还可以进行变量的选取。对于岭回归系数稳定但绝对值较小的变量或岭回归系数不稳定，但随着 k 值的增大而趋于 0 的变量可以删去，对剩余变量作新的岭回归。这一方法的缺点是 k 的选择具有一定的主观随意性。

9.5.2.2　均方误差最优法

这种方法选择 k 值的原则是，使岭回归估计的均方误差达到最小。最小二乘估计的均方误差定义为：

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{E}[(\hat{\beta} - \beta')(\hat{\beta} - \beta)] \qquad (9.5.3)$$

一个好的估计应该有较小的均方误差。由于 $\hat{\beta}$ 是 β 的无偏估计，$X'X$ 是非负定实对称矩阵且有逆阵。$(X'X)^{-1}$ 的特征根为 $\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_m^{-1}$，因此推导出最小二乘估计的均方误差，可以

表示为：

$$\mathrm{MSE}(\hat{\beta}) = \sigma^2 \sum_{i=1}^{m} \frac{1}{\lambda_i} \tag{9.5.4}$$

而岭回归的均方误差 $\mathrm{MSE}[\hat{\beta}(k)]$是常数 k 的函数。若存在 $k>0$，使得：

$$\mathrm{MSE}[\hat{\beta}(k)] < \mathrm{MSE}(\beta) \tag{9.5.5}$$

在均方误差意义下，岭回归估计优于最小二乘估计。可见，最优的 k 值依赖于未知数 $\beta(k)$及 σ^2。由于 $\beta(k)$及 σ^2 未知，计算时采用迭代算法。先用最小二乘估计 $\hat{\beta}$ 和 $\hat{\sigma}^2$ 代替，求出最优的 k 值，记为 k_1，算出 $\hat{\beta}(k_1)$，再用 $\hat{\beta}(k_1)$和 $\hat{\sigma}^2$ 代替，算出最优的 k_2……如此迭代下去，直到 k 值稳定为止。

9.5.2.3　预测残差平方和法

在 9.2 节中曾介绍过用预测残差平方和准则来确定最优子集回归。预测残差平方和 PRESS 越小，回归模型性能越好。

将 *PRESS* 用于岭回归估计。对特定的 k 值，用(9.2.8)式计算出岭回归的 *PRESS*，并作出 *PRESS* 对 k 关系的曲线，将使 *PRESS* 达到最小值的 k 值作为岭参数。

应用实例[9.4]：将岭回归方法用于表 9.4 的数据。用均方误差最优法确定 k 值。表 9.7 给出不同 k 值时的岭回归系数。

表 9.7　不同 k 值时的回归系数

k	x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	−0.0017	0.0654	−0.6251	1.1376	−0.0030	−0.3891	0.9068
0.6753	−0.0017	0.0655	−0.1655	0.5439	−0.0034	−0.3075	0.7035
1.3506	−0.0018	0.0655	−0.0368	0.3619	−0.0036	−0.2670	0.6195
2.0258	−0.0019	0.0656	0.0185	0.2726	−0.0037	−0.2384	0.5656
2.7011	−0.0019	0.0656	0.0464	0.2191	−0.0038	−0.2160	0.5250

由表 9.7 看出，x_1 和 x_3 的系数比较稳定且绝对值很小，因此将它们剔除。用剩余变量再进行岭回归。

参考文献

[1]施能，曹鸿兴．基于所有可能回归的最优气候预测模型．南京气象学院学报，1992，(4)

[2]陈希孺，王松桂．近代回归分析．合肥：安徽教育出版社，1987. 217～277

[3]王学仁，温忠麟．应用回归分析．重庆：重庆大学出版社，1989. 119～140

[4]俞善贤，汪铎．试用最优子集与岭迹分析相结合的方法确定回归方程．大气科学，1988，12(4)382～388

[5]Furnival G. M. All possible with less computation. Technometrics, 1971, (13): 403～ 408

[6]Furnival G. M. and R. W. M. Wilson. Regression by leaps and bound Technometrics. 1974, (16): 499—511

[7]Massy W. F. Principle component regression in exploratory statistical research, J. Amer. Statist. Assoc., 1965, (60): 234—266

[8]Webster J. T. et al. Latent root regression analysis. Technometrics, 1974, (16): 513—522

[9]魏松林．1881～1989 年东北地区夏季低温冷害．见：章基嘉主编．长期天气预报论文集．北京：海洋出版社，1992. 138～142

[10]Hoerl A. E. andR. W. Kennard. Ridge regression: biased estimation for non-orthogonal problems, Technometrics, 1970, (12): 55—88

第十章　均生函数预测模型

在现有的时间序列预测模型中，如 AR，ARMA 和 TAR 模型等，在制作多步预测时，预测值会趋于平均值，且往往对极值的拟合效果欠佳。指数平滑模型和灰色模型等可以制作多步预测，但它们表示的是一种指数增长，对于呈起伏型变化的气候序列不适用。依据气候时间序列蕴涵不同时间尺度振荡的特征，我们拓广了数理统计中算术平均值的概念，定义了时间序列的均值生成函数（Mean Generating Function，MGF，简称均生函数），提出了视均生函数为原序列生成的、体现各种长度周期性的基函数的新构思。在基函数基础上，相继给出了几种适于不同类型序列的建模方案[1~4]。均生函数预测模型既可以作多步预测，又可以较好地预测极值，解决了上述两类模型的难题，为长期天气预报和短期气候预测开辟了一条新途径。从建模方法而言，均生函数模型是借助多元分析的手段，解决时间序列预测问题的一种尝试。

本章将给出均生函数的定义，重点叙述设计较完善、适用性较强的一种建模方案。我们将均生函数概念扩充到模糊集中，定义出模糊均生函数，给出相应的建模方案及实施步骤。

§10.1　均值生成函数

设一时间序列：

$$x(t) = \{x(1), x(2), \cdots, x(n)\} \qquad (10.1.1)$$

其中 n 为样本量。$x(t)$的均值为：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x(i) \tag{10.1.2}$$

对于(10.1.1)式定义均值生成函数

$$\bar{x}_l(i) = \frac{1}{n_l}\sum_{j=0}^{n_l-1} x(i+jl) \quad (i = 1,\cdots,l,1 \leqslant l \leqslant m) \tag{10.1.3}$$

式中 n_l=INT(n/l),m=INT($n/2$)或 INT($n/3$),INT 表示取整数。

根据(10.1.3)式,可以得到 m 个均生函数

$\bar{x}$

$\bar{x}_2(1),\bar{x}_2(2)$

$x_3(1),x_3(2),\bar{x}_3(3),$

$\vdots$

$x_m(1),x_m(2),\cdots,\bar{x}_m(m)$

由此可见,均生函数是由时间序列按一定的时间间隔计算均值而派生出来的。

将均生函数定义域延拓到整个数轴上,即作周期性延拓:

$$f_l(t) = \bar{x}_l(i) \quad t \equiv i[\mathrm{mod}(l)] \quad (t = 1,2,\cdots,n) \tag{10.1.4}$$

这里 mod 表示同余。我们称 $f_l(t)$为均生函数延拓序列是一种周期函数。由此构造出均生函数延拓矩阵:

$$F = (f_{ij})_{n\times m} \qquad f_{ij} = f_l(t) \tag{10.1.5}$$

$$F=\begin{bmatrix}\bar{x} & \bar{x}_2(1) & \bar{x}_3(1) & \cdots & \bar{x}_m(1)\\ \bar{x} & \bar{x}_2(2) & \bar{x}_3(2) & \cdots & \bar{x}_m(2)\\ \bar{x} & \bar{x}_2(1) & \bar{x}_3(3) & \cdots & \\ \bar{x} & \bar{x}_2(2) & \bar{x}_3(1) & \cdots & \vdots\\ & & & & \bar{x}_m(m)\\ \vdots & \vdots & \vdots & & \vdots\\ \bar{x} & \bar{x}_2(i_2) & \bar{x}_3(i_3) & \cdots & \bar{x}_m(i_m)\end{bmatrix} \tag{10.1.6}$$

其中 $\bar{x}_2(i_2)$表示顺序取 $\bar{x}_2(1)$,$\bar{x}_2(2)$之一,$\bar{x}_3(i_3)$表示顺序取 $\bar{x}_3(1)$,$\bar{x}_3(2)$,$\bar{x}_3(3)$之一,余类推。称 f_t 为延拓均生函数。

如同在数理统计中通常所要求的,序列样本量 n 不小于30,而对求均值的样本量不作严格限制。当然,至少要有两个数据求平均,否则失去平均的意义。

我们将均生函数延拓矩阵(10.1.6)式中第一列记为 f_1,第二列记为 f_2…第 m 列记为 f_m。

f_1 是序列 $x(t)$的均值,它是由 n 个数据相加求平均而成,故随机性最小。

f_2 是由$[\frac{n}{2}]$个数据相加求平均而成,当 n 充分大时,随机性亦小。

当 m 取为 INT($\frac{n}{2}$)时,f_m 是由两个数据相加平均而成,故随机性较大。当 m 取为 INT($\frac{n}{3}$),f_m 也只是由三个数据相加平均而得,亦有较大随机性。

由此可见,在求取 $f_1\sim f_m$ 时,由于求均值的样本量由大变小,其均值序列的随机性也由弱到强。也就是说,长周期的均生函数随机性较大,短周期的均生函数随机性小。

应用实例[10.1]:表 10.1 为合肥 1951～1998 年 7 月的

表 10.1　合肥 1951～1998 年 7 月的降水量

年份	降水量									
1951～1960	164	164	216	558	91	77	309	56	18	297
1961～1970	124	266	185	63	159	57	47	147	372	217
1971～1980	191	188	188	127	108	38	118	119	171	360
1981～1990	189	286	176	44	179	239	299	151	237	85
1991～1998	448	150	98	66	43	309	54	112		

降水量。这里取 1951～1993 年作为统计样本量，即 $n=43$。这里 m 取 $\mathrm{INT}(\frac{n}{3})=14$，计算时间间隔 l 取 1～14 的均生函数。

$$\bar{x}_1(1)=\frac{1}{43}(164+164+216+558+\cdots+98)=181.1$$

$$\bar{x}_2(1)=\frac{1}{22}(164+216+91+309+\cdots+98)=185.8$$

$$\bar{x}_2(2)=\frac{1}{21}(164+558+77+66+\cdots+150)=176.1$$

$$\bar{x}_3(1)=\frac{1}{15}(164+558+309+297-\cdots+98)=204.8$$

$$\bar{x}_3(2)=\frac{1}{14}(164+91+66+188+\cdots+448)=159.5$$

$$\bar{x}_3(3)=\frac{1}{14}(216+77+18+266+\cdots+150)=177.3$$

$$\vdots \qquad\qquad \vdots$$

$$\bar{x}_{14}(1)=\frac{1}{4}(164+159+177+98)=148.0$$

$$\bar{x}_{14}(2)=\frac{1}{3}(164+57+360)=193.7$$

$$\bar{x}_{14}(3)=\frac{1}{3}(216+47+189)=150.7$$

$$\vdots \qquad\qquad \vdots$$

$$\bar{x}_{14}(14)=\frac{1}{3}(63+119+150)=110.7$$

构造均生函数延拓矩阵：

$$F = f_{43,14} = \begin{bmatrix} 181.1 & 185.8 & 204.8 & \cdots & 148.0 \\ 181.1 & 176.1 & 159.5 & \cdots & 193.7 \\ 181.1 & 185.8 & 177.3 & \cdots & 150.7 \\ 181.1 & 176.1 & 204.8 & \cdots & 330.3 \\ & & & & \vdots \\ & & & & 110.7 \\ \vdots & \vdots & \vdots & & 148.0 \\ & & & & \vdots \\ 181.1 & 176.1 & 177.3 & \cdots & 112.7 \end{bmatrix}$$

图 10.1 为周期长度 l 取 2,4,7,14 年的均生函数延拓序列变化曲线。如图所示的上下起伏变化，可以建立直观的均生函数的概念。

§10.2 双评分准则

统计模型的选择通常是指在多元分析方程中选取自变量的个数或在时间序列中确定模型的阶数。这关系到模型预报的精度和方程的稳定性，也涉及到计算量的大小。在很长一段时间内，使用基于假设检验的方法。例如：F 检验。这种方法事先必须给定信度水平 α，带有人为性，使用也较繁琐。70 年代初，赤池弘次提出了最小信息量准则，即 AIC。之后，Schwartz 从贝叶斯定理导出了类似的 BIC 准则。AIC 和 BIC 准则是在残差平方和与变量个数间进行权衡，当两个模型的残差平方和相同时，取变量个数少的模型。这里把吝啬原理在统计模型选择中加以具体化。但是，对于气候预测而言，要求的是预测准确，尤其要求“趋势”报对。只有报对趋势，模型才有实用价值。尤其在计算机的内存、速度条件日新月异发展的今天，方程中增加几个自变量的计算量已微不足道。也就是说，对于气候预测问题，最重要的是要使诸如旱涝、冷暖等趋势报对。

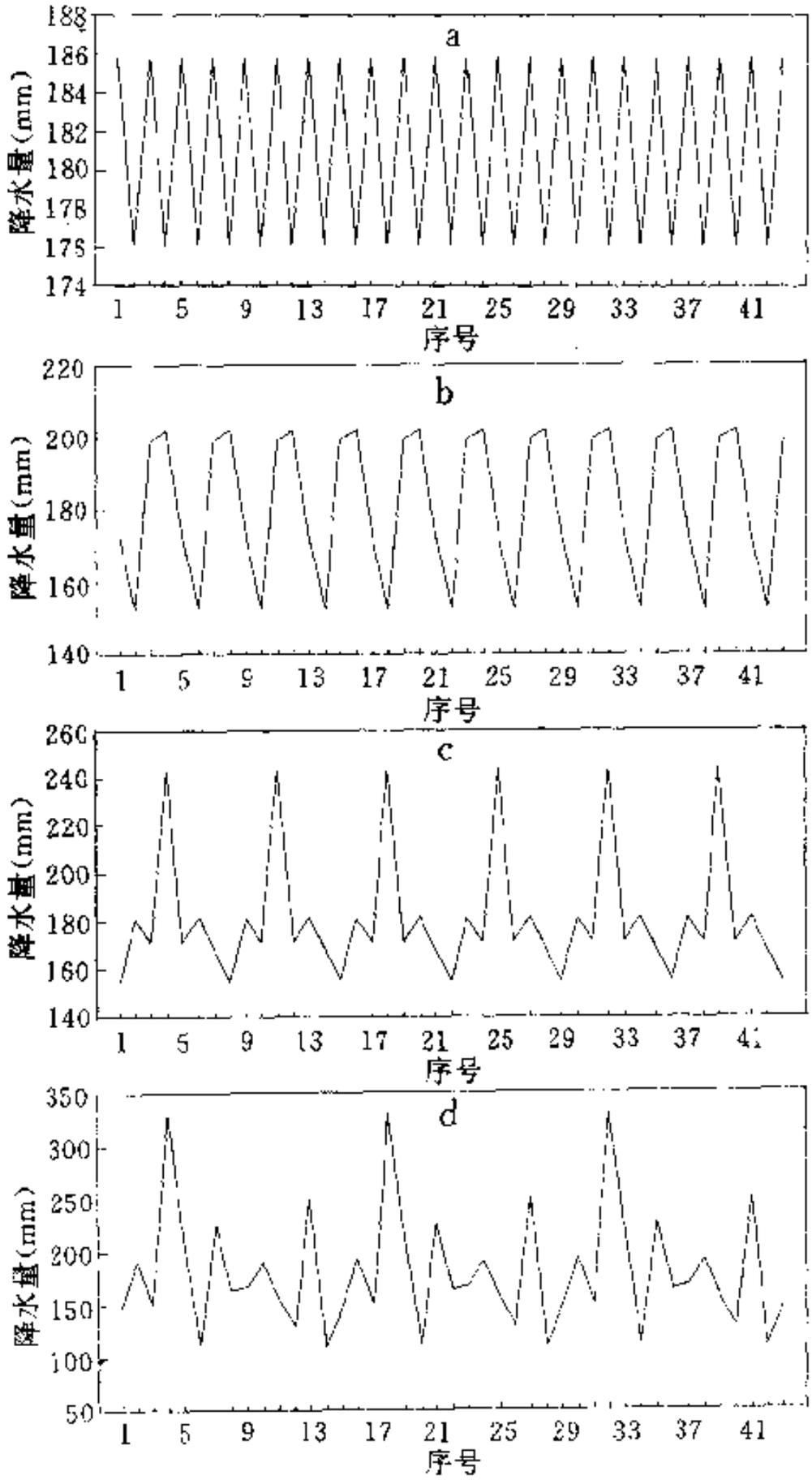


图 10.1　均生函数延拓序列 a:$l=2$; b:$l=4$; c:$l=7$; d:$l=14$

基于上述考虑，均生函数模型选择使用以数量预报的评分和趋势预报的评分来权衡变量的筛选的双评分准则(Couple Score Criterion，CSC)[5]。双评分准则旨在使数量评分和趋势评分均达到最小，以尽可能报对趋势。

用 S_1 表示数量评分，由于它是对具体量测数据和模型预测值之差的评定，故称为精评分。用 S_2 表示趋势评分，即粗评分。那么，双评分准则表示为：

$$CSC = S_1 - S_2 \tag{10.2.1}$$

预报量的均值为：

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t \tag{10.2.2}$$

令：

$$Q_x = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2 \tag{10.2.3}$$

为每次都用均值作为预报值而得到的数量评分。这种用均值所作的预报称为气候学预报。

设预报方程中引进了 k 个因子，模型残差平方和

$$Q_k = \sum_{t=1}^{n}(x_t - \hat{x})^2 \tag{10.2.4}$$

那么数量评分定义为：

$$S_1 = nR^2 = n(1 - \frac{Q_k}{Q_x}) \tag{10.2.5}$$

R^2 为复相关系数。(10.2.5)式中 Q_k/Q_x 的含意是，一个好的预报方法，其误差必须比气候预报小，即 $Q_k/Q_x<1$。当 $n\to\infty$ 时，S_1 为自由度 $\nu=k$ 的 χ_2 分布。

趋势度量取最小判别信息统计量 $2I$：

$$S_2 = 2I = 2[\sum_{i=1}^{I}\sum_{j=1}^{I} n_{ij}\ln n_{ij} + n\ln n - (\sum_{i=1}^{I} n_{i.}\ln n_{i.} + \sum_{j=1}^{I} n_{.j}\ln n_{.j})] \tag{10.2.6}$$

式中 I 为预报趋势类别数；n_{ij} 为 i 类事件与 j 类估计事件列联表的个数。其中

$$n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$$

$$n_{i\cdot} = \sum_{j=1}^{I} n_{ij}$$

至于列联表的确定，将在实例中具体说明。

在事件相互独立的假设下，$2I$ 的渐近分布为 χ^2 分布，自由度 $\nu_2=(I-1)(I-1)$。

因此，双评分准则的表达式为：

$$CSC_k = S_1 + S_2 = nR^2 + 2I \qquad (10.2.7)$$

根据 χ^2 的可加性，则

$$\chi^2_\nu = \chi^2_{\nu_1} + \chi^2_{\nu_2} \qquad (10.2.8)$$

为 χ^2_ν 分布，自由度 $\nu=k+(I-1)^2$。

显而易见，用 CSC_k 筛选回归方程的自变量的标准为：

$$\max(CSC_k) = CSC_k \qquad (10.2.9)$$

并可对 CSC_k 进行 χ^2 检验。当 $CSC_k>\chi^2_\alpha$ 时，自变量入选，显著性水平 α 视具体情况，取为 0.05，0.01 或 0.001。

趋势评分可视实际问题而定。这里仅给出一种方案。其中趋势类别数 I 根据预报量变化趋势来确定。将 n 个样本的预报量 y 分成若干类别。如降水量预报中可分为偏多、正常和偏少 3 类。计算

$$U = \frac{1}{n-1}\sum_{t=1}^{n} |\Delta x_t|$$

式中

$$\Delta x_t = x_t - x_{t-1} \qquad (t = 2,3,\cdots,n)$$

若预报量分为 3 类：

x_A：当 $\Delta x_t > U$

x_B：当 $|\Delta x_t| \leqslant U$

x_C：当 $\Delta x_t < -U$

x_A, x_B, x_C 分别表示预报量变化的升、平、降 3 种趋势。

应用实例[10.2]：取表 10.1 合肥 1951～1993 年 7 月降水量预报量，以应用实例[10.1]中计算出的单个均生函数为预报因子，分别建立一元回归方程。这时可以利用(10.2.5)和(10.2.6)式分别计算 CSC 的数量评分和趋势评分。

计算一元回归的残差平方和 Q_k 和 Q_r，代入(10.2.5)式得出 S_1。然后再求 S_2。以 CSC 值最大者长度为 9 年的均生函数来说明 S_2 的求法。表 10.2 为一元回归方程引进长度为 9 年均生函数 f_9 的列联表。

表 10.2　改进 f_9 的列联表

n_{ij}		预报			总计 $n_{i.}$
		1 类	2 类	3 类	
实况	1 类	4	3	2	9
	2 类	0	3	1	4
	3 类	4	3	22	29
总计 $n_{.j}$		8	9	25	$n=42$

表 10.2 的横栏为预报，竖栏为实况，分为 3 类，表中数字为相应类别的出现频数。例如：第一行表示实况为 1 类而预报分别为 1，2 和 3 类的频数，$n_{1.}=9$ 为实况是 1 类的频数总和。再例如：第一列表示预报是 1 类而实况分别为 1，2 和 3 类的频数，$n_{.1}=8$ 为预报是 1 类的频数总和。余类推。从列联表中的 $n_{.j}$ 和 $n_{i.}$ 可以清晰地看出预报与实况频数的差异，将列联表中的数字代入(10.2.6)式，即可算出趋势评分 $S_2=14.99$。根据一元回归的残差平方和 Q_k 和 Q_r，利用(10.2.5)式算出数量评分 $S_1=11.76$。$CSC=S_1+S_2=26.75$。这时 $\nu_1=k=1$，$\nu_2=(3$

$-1)(3-1)=4$，故 $\nu=5$。给定显著性水平 $\alpha=0.05$，那么 $\chi^2_{0.05}=11.07$，$CSC>11.07$。因此，选择均生函数 f_9 作为备选因子。这是双评分准则在选择备选因子时的应用。它在建立预测模型(用于模型识别)的应用，在以后建模步骤中还会提及。

§10.3 均生函数预测模型

构造出均生函数，就可以通过建立原时间序列与这组函数间的回归，建立预测模型。我们先后设计了几种计算方案。

10.3.1 逐步回归筛选方案

利用逐步回归技术筛选时间序列 $x(t)$ 生成的均生函数。将均生函数视为备选因子，原始序列作为预报量。按照通常的逐步回归步骤进行计算。优势周期是用均生函数与原序列的相关系数来计算方差贡献，并依次选取方差贡献的最大值来确定的。为了能选取随机性较小稳健性较大的均生函数建立方程，在方差贡献上添加“惩罚”系数，以避免随机性较大的长周期均生函数入选。

设长度为 l 的均生函数的方差贡献为 U_l，则令：

$$V_l=\alpha_l U_l \qquad \alpha_l=\frac{n}{l} \qquad (l=2,3,\cdots[\frac{n}{2}]\text{或}[\frac{n}{3}]) \tag{10.3.1}$$

当 l 较小时，α_l 较大，即对方差贡献施加较大权重。随着 l 不断增大，α_l 逐渐变小，以期筛选出隐含于序列中的周期，进行 F 检验时，再将方差贡献复原。

设作 q 步预报，将入选的均生函数作 q 步外延，则得到预报方程：

$$\hat{x}(n+q) = a_0 + \sum_{i=1}^{k} a_i f_i(n+q) \qquad (q = 1,2\cdots)$$

(10.3.2)

其中 a_0 和 a_i 为逐步回归技术估计的系数；f_i 为延拓均生函数。

10.3.2 正交筛选方案

通过 Gram-Schmidt 正交化处理，使均生函数正交化。令 f_2 作为正交化的初始向量，对 $f_3, f_4, \cdots, f_m$ 进行正交化，求得 $m-1$ 个正交化序列 $\tilde{f}_2, \tilde{f}_3, \cdots, \tilde{f}_m$。

以 $\tilde{f}_2, \tilde{f}_3, \cdots, \tilde{f}_m$ 作为自变量与 $x(t)$ 建立线性模型

$$x(t) = \sum_{i=2}^{m} \tilde{a}_i \tilde{f}_i(t) + e(t) \qquad (10.3.3)$$

向量一矩阵表达式为：

$$\underset{n\times 1}{X} = \underset{n\times(m-1)}{\tilde{F}} \quad \underset{(m-1)\times 1}{\tilde{A}} \qquad (10.3.4)$$

求最小二乘解：

$$\tilde{A} = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'X \qquad (10.3.5)$$

由于 $\tilde{f}_i$ 与 $\tilde{f}_j$ 之间是正交的，因此协方差阵为对角阵，逆矩阵 $G=(\tilde{F}'\tilde{F})^{-1}$ 亦为对角阵，其元素为：

$$\tilde{g}_{ii} = \frac{1}{\tilde{f}_{ii}} \qquad (10.3.6)$$

线性模型的系数为：

$$\tilde{a}_i = \tilde{g}_{ii} \sum_{t=1}^{n} \tilde{f}_i(t) x(t) \quad (i = 2,3,\cdots,m) \qquad (10.3.7)$$

线性模型系数的大小表示均生函数的重要程度。将 $\tilde{a}_i$ 按绝对值的大小排序，均生函数 $\tilde{f}_i$ 依 $\tilde{a}_i$ 绝对值由大到小进入，进入方程的均生函数个数由双评分准则确定。由于均生函数是正交的，因此在筛选过程中系数不必重复计算。

由于模型的系数是由正交均生函数确定的。因此，必须求出原均生函数 $f_i(t)$ 的系数 a_i，建立形如(10.3.2)式的预报方程。

10.3.3 最优子集回归建模方案

为了建立预报效果更佳的模型，我们设计了一个适用性较强的建模方案[6]。具体计算方案及步骤如下：

(1)为了拟合原序列中的高频部分，对原序列进行差分运算。这一运算实际上起着高通滤波的作用。

作一阶差分运算

$$\Delta x(t) = x(t+1) - x(t) \quad (t = 1,2,\cdots,n-1) \tag{10.3.8}$$

得到序列

$$x^{(1)}(t) = \{\Delta x(1),\Delta x(2),\cdots,\Delta x(n-1)\} \tag{10.3.9}$$

作二阶差分运算

$$\Delta^2 x(t) = \Delta x(t+1) - \Delta x(t) \quad (t = 1,2,\cdots,n-2) \tag{10.3.10}$$

得到序列

$$x^{(2)}(t) = \{\Delta^2 x(1),\Delta^2 x(2),\cdots,\Delta^2 x(n-2)\} \tag{10.3.11}$$

用(10.1.3)式分别计算原序列 $x(t)$，一阶差分序列 $x^{(1)}(t)$ 和二阶差分序列 $x^{(2)}(t)$ 的均生函数，分别记为 $\bar{x}_l^{(0)}(t)$，$\bar{x}_l^{(1)}(t)$ 和 $\bar{x}_l^{(2)}(t)$。利用(10.1.4)式即可得到它们的延拓序列 $f_l^{(0)}(t)$，$f_l^{(1)}(t)$ 和 $f_l^{(2)}(t)$。

为了拟合时间序列中向上递增和向下递减的趋势，进一步建立累加延拓序列

$$f_l^{(3)}(t) = x(1) + \sum_{i=1}^{t-1} f_l^{(1)}(i+1)$$
$$(t = 2,3,\cdots,n;\ l = 1,2,\cdots,m) \tag{10.3.12}$$

其中 $f_l^{(3)}(1)=x(1)$。累加延拓实际上是用一阶差分的均生函数代替不同时刻差分值。

这样共获得 $4m$ 个均生函数延拓序列 $f_l^{(0)}(t)$，$f_l^{(1)}(t)$，$f_l^{(2)}(t)$，$f_l^{(3)}(t)$，$l=1,2,\cdots,m$，作为自变量供筛选。

(2)建立每一个延拓序列与原序列间的一元回归，计算双评分准则 CSC 值，凡满足 $CSC>\chi_\alpha^2$ 的序列粗选为预报因子。设入选了 P 个延拓序列。

(3)用 Furnival-Wilson 设计的算法计算出所有可能的 2^P 个回归子集。从 2^P 个回归子集中根据双评分选择变量标准，选出一个最优回归子集作为预报方程。

(4)假定最优回归子集由 k 个变量构成，则均生函数模型为：

$$\hat{x}(t)=a_0+\sum_{i=1}^{k}a_i f_i(t) \qquad (10.3.13)$$

若作 q 步预报，对 k 个序列作 q 步外延代入方程(10.3.13)中即可。

10.3.4 模拟计算

为了验证均生函数预测模型的优效性，进行了下列模拟计算：

给定一时间序列，由

$$x(t)=e^{\alpha t}\sin t\cos 2t \qquad (10.3.14)$$

计算出 $t=1,2,\cdots,n$，$\alpha-0.5$。这里设样本量 $n=50$。由(10.3.14)式构造出一个样本量为 50 的时间序列 $x(t)$。

对给定序列 $x(t)$分别用最优子集回归的均生函数模型、双向灰色模型和自回归模型进行模拟。

10.3.4.1 均生函数模型

这里均生函数个数 m 取 25。按 10.3.3 叙述的计算步骤计算。当有 5 个延拓序列进入模型时，双评分准则 CSC 值达

到极大。建立预测方程：

$$x(t) = 0.001\ 747\ 8 + 0.669\ 61\ f_{13}^{(0)}(t) - 0.028\ 267\ f_{13}^{(3)}(t) + 0.663\ 62\ f_{12}^{(0)}(t) - 0.007\ 771\ 0\ f_{6}^{(0)}(t) + 0.676\ 604\ f_{35}^{(0)}(t)$$

其中 $f_{13}^{(0)}(t)$表示原序列周期长度为13的均生函数延拓序列，$f_{13}^{(3)}(t)$表示周期长度为13的累加延拓序列，其余类推。

模型拟合原序列均方根误差 $RMSE=0.40$，预测5步的 $RMSE=0.9777$。

10.3.4.2　双向灰色模型

为了充分利用包含在时间序列中的信息，我们曾对一般的灰色模型进行了改进，提出了使向前差分预报误差和向后差分预报误差之和达最小的原则的双向灰色模型。其预报方程形式为：

$$x(t+1) = \left[x(1) + \frac{b_0}{b_1}\right]e^{b_1 t} - \frac{b_0}{b_1}$$

用最小二乘法估计 b_0 和 b_1。

序列(10.3.14)式的预报方程为：

$$x(t+1) = \left(-0.35 + \frac{0.028\ 702}{0.019\ 938}\right)e^{0.019\ 938\,t} - \frac{0.028\ 702}{0.019\ 938}$$

方程拟合原序列均方根误差 $RMSE=0.64$，预测5步的 $RMSE=1.0956$。

10.3.4.3　自回归方程

对序列(10.3.14)式建立自回归模型，用 AIC 准则定阶，确定预报模型为：

$$x(t) = 0.5360\,x(t-1) - 0.9551x(t-2) - 0.5989x(t-3) - 0.4441x(t-4) + 0.000\ 01x(t-5)$$

模型拟合原序列 $RMSE=0.9162$，预测 5 步的 $RMSE=1.5003$。

从模拟计算结果可见，均生函数预测模型的拟合和预测效果均优于灰色模型和自回归模型。

应用实例[10.3]：用 10.3.3 的最优子集回归建模方案，用表 10.1 所列的合肥 1951～1993 年 7 月降水量建立均生函数预测模型并作 1994～1998 年 5 年预报。这时样本量 $n=43$，$m=14$，预报步数 $q=5$。

对 $4m$ 个均生函数延拓序列依次与原降水量序列建立一元回归，用双评分准则进行粗选，共选出 13 个延拓序列作为预报因子。计算所有可能的 2^{13} 个回归子集。再由双评分准则最后确定出由 8 个均生函数延拓序列构成的预测模型。拟合原序列的均方根误差 $RMSE=53.6$mm。图 10.2 给出 1951～1993 年合肥 7 月降水量的实际值与拟合值及 1994～1998 年实际值与预报值的比较。实线为实际值，虚线为拟合值与预报值。

由图 10.2 可以看出，模型的拟合值与实况值很接近。尤其可贵的是，极值年的拟合值与实况值吻合得相当好，这是许多模型难以做到的。由图还可以看出，模型所作的 5 年预报效果也十分好。合肥 7 月降水量多年平均为 181mm，预报 1994，1995，1997 和 1998 年低于平均值，实况均在平均值以下。预报 1996 年降水量偏多，高于平均值，实况确实如此。可见，预报模型可以把旱涝变化趋势很好地预报出来。对于夏季降水量这种变化幅度较大的变量而言，能够有这样的预报效果已经相当令人满意了。

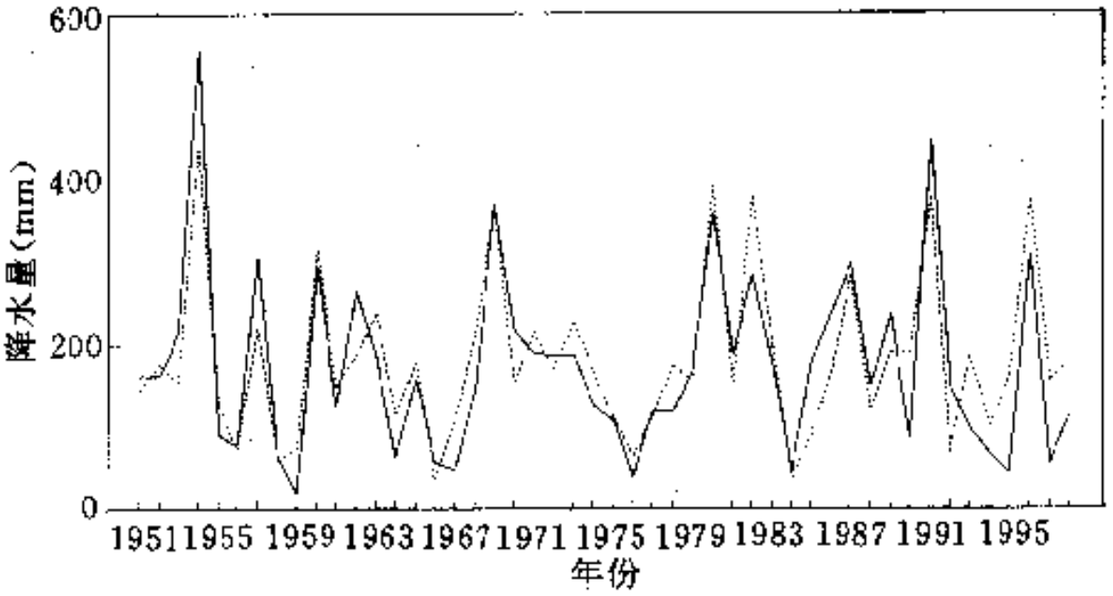


图 10.2　合肥 7 月降水量变化曲线

§10.4　模糊均生函数模型

将均生函数的概念推广到模糊集中，推导出不同类型序列的隶属度，定义出模糊均生函数，给出了相应的建模方案及实施步骤[7]。

10.4.1　模糊均生函数

将均生函数的概念推广到模糊集中。设论域

$$U = \{u_i | i = 1, 2, \cdots, n\} \tag{10.4.1}$$

其中 n 为样本量。在 U 上构造模糊子集 $\underset{\sim}{A}$

$$\underset{\sim}{A} = \mu_1/u_1 + \mu_2/u_2 + \cdots + \mu_n/u_n \tag{10.4.2}$$

我们将起报时刻记为 t_n。那么，要对未来时刻 $t_{n+1}, t_{n+2}, \cdots, t_{n+l}$ 作出预报。从预报的物理意义上考虑，愈靠近起报时刻的观测值包含对预报有用的信息愈多，对预报愈有价值。马尔柯夫过程认为，系统所处的状态与时刻 t_n 以前所处的状态无关，只有 t_n 时刻的观测值才对预报有用。也就是说，若序列具

有马尔柯夫性，则可以把隶属函数定义为

$$A_M = 0/u_1 + 0/u_2 + \cdots + 1/u_n \qquad (10.4.3)$$

A_M 表示具有马尔柯夫性序列的隶属函数。

若从统计学观点出发，则把样本 $x(1), x(2), \cdots, x(n)$ 等概率对待，即隶属函数定义为：

$$A_s = 1/u_1 + 1/u_2 + \cdots + 1/u_n \qquad (10.4.4)$$

A_s 表示统计学意义下的隶属函数。

在实际问题中，我们既不忍心丢舍过多以往的信息，又希望近期观测值对预报发挥较大作用。为此，设计了随 t_n 的远近以指数形式下降的隶属度，即

$$\underset{\sim}{\mu_A}(t_i) = \begin{cases} e^{-\beta(t_n - t_i)} & t_i < t_n \\ 1 & t_i \geqslant t_n \end{cases} \qquad (10.4.5)$$

当 $\beta=0$ 时，(10.4.5)式蜕化为(10.4.4)式。若等间隔取样，令 $\Delta t=1$，$t_i=i\Delta t$，$t_n=n\Delta t$，则(10.4.5)式可写为

$$\underset{\sim}{\mu_A}(i) = \begin{cases} e^{-\beta(n-i)} & i < n \\ 1 & i \geqslant n \end{cases} \qquad (10.4.6)$$

式中 β 按对过去观测值重视程度事先给定。显然，$0 \leqslant \underset{\sim}{\mu_A}(i) \leqslant 1$。

若序列具有周期性，则可令隶属度为：

$$\underset{\sim}{\mu_A}(i) = \begin{cases} r\sin\dfrac{2\pi}{l}(n-i) & i < n \\ 1 & i \geqslant n \end{cases} \qquad (10.4.7)$$

其中 l 为周期长度；r 为由经验或试算确定的常数。显然，当 $r \leqslant 1$ 时，有 $0 \leqslant \underset{\sim}{\mu_A}(i) \leqslant 1$。

若既考虑观测值随起报时刻远近效用逐渐下降又体现周期性，则令隶属度为：

$$\mu_{\tilde{A}}(i)=\begin{cases}re^{-\beta(n-i)}\sin\dfrac{2\pi}{l}(n-i) & i<n\\ 1 & i\geqslant n\end{cases}\quad(10.4.8)$$

基于 $\mu_{\tilde{A}}$ 构造模糊向量

$$a=(\mu_{\tilde{A}}(1),\mu_{\tilde{A}}(2),\cdots,\mu_{\tilde{A}}(n))\quad(10.4.9)$$

在论域 U 上构造另一个模糊子集 $\tilde{B}$，它的隶属函数取为：

$$\mu_{\tilde{B}}(i)=x(i)/x_{\max}\quad(10.4.10)$$

式中 $x_{\max}=\max x(i)$，显然 $0\leqslant\mu_{\tilde{B}}(i)\leqslant1$。构造模糊向量

$$b=(\mu_{\tilde{B}}(1),\mu_{\tilde{B}}(2),\cdots,\mu_{\tilde{B}}(n))\quad(10.4.11)$$

用代数加和乘定义模糊向理的内积

$$a\cdot b=\frac{1}{n}\sum_{i=1}^{n}\mu_{\tilde{A}}(i)\mu_{\tilde{B}}(i)\quad(10.4.12)$$

根据(10.4.12)式定义模糊均生函数(Fuzzy Mean Generating Function，FMGF)

$$\bar{x}_l(i)=\frac{C}{n_l}\sum_{j=0}^{n_l-1}\mu_{\tilde{A}}(i+jl)\mu_{\tilde{B}}(i+jl)\quad(10.4.13)$$

式中 C 为给定的常数，使得 FMGF 与序列 $x(t)$ 的量级相同。方便地可取 $C=x_{\max}$，(10.4.13)式就变为：

$$\bar{x}_l(i)=\frac{1}{n_l}\sum_{j=0}^{n_l-1}\mu_{\tilde{A}}(i+jl)x(i+jl)$$
$$(i=1,2,\cdots,l)\quad(10.4.14)$$

像均生函数一样作周期性延拓，得到模糊均生函数外延序列。

10.4.2 建模方案及实施步骤

设一原始序列

$$x^{(0)}(t)=\{x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n)\}$$

建立其预测模型的方案及实施步骤如下：

(1)对 $x^{(0)}(t)$ 作一和二阶差分运算，得到

$$x^{(1)}(t) = \{\Delta x^{(1)}(1), \Delta x^{(1)}(2), \cdots, \Delta x^{(1)}(n)\}$$

$$x^{(2)}(t) = \{\Delta x^{(2)}(1), \Delta x^{(2)}(2), \cdots, \Delta x^{(2)}(n)\}$$

(2)对序列 $x^{(0)}(t)$, $x^{(1)}(t)$和 $x^{(2)}(t)$分别用(10.4.14)求出模糊均生函数并作周期性延拓。这里 μ_A选用(10.4.6)式计算,β 取 0.01,这样得到三组模糊均生函数延拓序列 $f_l^{(0)}(t)$, $f_l^{(1)}(t)$, $f_l^{(2)}(t)$, $l=1,2,\cdots,m$, $m=\mathrm{INT}(n/2)$。

(3)构造一组累加延拓序列:

$$f_l^{(3)}(t) = x^{(0)}(1) + \sum_{i=1}^{t-1} f_l^{(1)}(i+1)$$

$$(t = 2,3,\cdots,n \qquad l = 1,2,\cdots,m) \quad (10.4.15)$$

其中 $f_l^{(3)}(1)=x^{(0)}(1)$。

(4)建立每一延拓序列与原序列间的一元回归,用双评分准则对 $4\times m$ 个延拓序列逐一筛选。

(5)由 4 步粗选得的因子,再进行精选。用所有可能子集回归是最好的,但计算量较大,这里采用按双评分准则 *CSC* 值由大到小,将粗选出的模糊均生函数逐一引入回归方程。当 *CSC* 出现极大值时,停止筛选。假设引入 k_0 个模糊均生函数时 *CSC* 值为极大,那么预报方程为:

$$\hat{x}(t) = a_0 + \sum_{i=1}^{k_0} a_i f_i(t) \qquad (10.4.16)$$

§10.5 全国夏季降水趋势分布预报方法

由于影响我国夏季降水的因素十分复杂,且气候噪音背景很强,致使降水的短期预测难度很大。因此,研究客观预报方法、提高预报准确率是短期气候预测的重要研究课题。依据气候系统具有不同时间尺度周期振荡的特性,构造代表不同

降水分布型式变化序列的均生函数延拓序列作为一部分预报因子。另外还考虑了其它影响夏季降水的强信号。由此设计出一套具有一定物理基础及一定统计信度支持的预报方案。1994 年以来，我们一直坚持用这一预报方案制作全国夏季降水的趋势预报，取得了比较好的预报效果[8~9]。

10.5.1 预报思路和流程

取 1951 年以来至今的中国 160 站 6～8 月降水总量作经验正交函数分解。前三项特征向量解释了总方差的 97%。第一特征向量解释总方差的 94.4%。它反映了中国夏季降水的多年平均状况，即呈东南向西北递减的降水分布型式。第二特征向量代表了江淮流域降水与其南北趋势为相反的分布型式。这是中国夏季降水最常见的分布型式。第三特征向量代表了江南与黄淮之间的降水趋势呈相反的分布型式，它也是我国夏季比较常见的降水分布型式。可见前三个特征向量概括了我国夏季降水最基本的分布形式。将我国每年的夏季降水趋势分布看作是由大范围降水多寡及不同分布型式的扰动两部分叠加而成，即

$$R = \overline{R} + R' \qquad (10.5.1)$$

式中的 $\overline{R}$ 由第一特征向量及其时间系数表示，R' 则由第二特征向量和第三特征向量及其时间系数表示。在保证全国大范围降水趋势预报正确的前提下，再报准扰动项的基本趋势，这一年的预报就会有一定的把握。具体实施时，要先建立第一、第二和第三特征向量时间系数序列的预报方程，预报出未来一年的这三个时间系数的数值，再乘以相应的特征向量，就可以得到降水场的预报。

预报依据由两部分组成：一部分是影响夏季降水的强信号。其中包括赤道东太平洋地区的海温、南方涛动指数、冬季

北太平洋涛动指数和太平洋地区极涡面积指数。另一部分则由降水量本身不同时间尺度的变化构成。将长期振荡及各种短期振荡从代表降水年际趋势变化的特征向量时间系数序列中提取出来，即按(10.1.3)式计算特征向量时间系数序列的均生函数。将这部分因子加到预报方程中，期望气候振荡在预报中发挥作用。

设计出一个制作夏季降水趋势分布预报的流程，如图10.3所示。

计算步骤如下：

(1)用中国160站1951年以来的6～8月降水总量作经验正交函数分解。例如：制作1994年预报，则用1951～1993年6～8月降水总量分解，提取前3个特征向量的时间系数作为预报量。

(2)分别计算前3个特征向量的时间系数与上述预报因子之间的相关系数，把相关系数达到给定信度的因子选出来作为备选因子。

(3)假定粗选出 k 个预报因子，用 Furnival wilson 设计的算法计算出所有可能的 2^k 个回归子集。从 2^k 个回归子集中根据双评分选择变量标准，选出一个最优回归子集作为预报时间系数的方程。

(4)用预报出的该年前3个特征向量的时间系数乘以相应的特征向量，得到该年全国160站6～8月降水总量的预报。

(5)将降水量预报值对1961～1990年平均求偏差，即得到该年降水距平百分率的预报。

10.5.2 预报效果检验

分别对1951～1993，1951～1994，1951～1995，1951～1996和1951～1997年6～8月降水量场进行经验正交函数

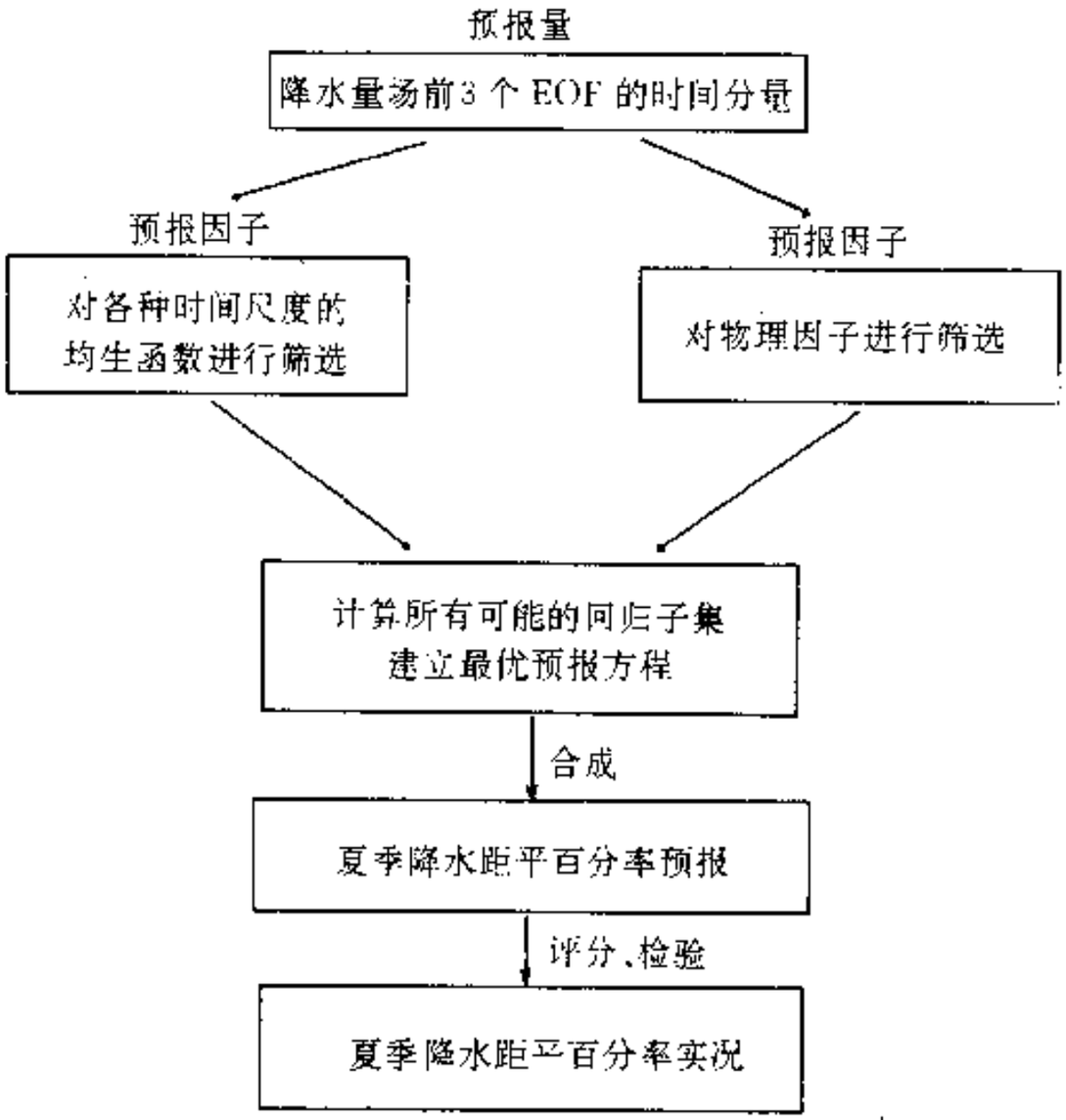


图 10.3　中国夏季降水预报流程图

分解，按照预报流程，算得 1994～1998 年夏季降水距平百分率的预报。这 5 年的预报均提供给全国汛期预报会商大会。表 10.3 为 1994～1998 年预报的效果检验。表中 R 表示预报场

表 10.3　1994～1998 年预报的效果检验

年份	1994	1995	1996	1997	1998	平均
R	0.36	0.14	0.18	0.31	0.31	0.26
S	80	75	77	82	83	79.4

与实况场之间的距平相关系数。距平相关系数是国际上通用

的、比较客观的评定办法。S 表示采用国家气候中心气候预测室评分标准的得分[8]。评分公式为：

$$S = \frac{N + N_0 \div N' + N''}{M + N' + N''} \times 100\% \quad (10.5.2)$$

其中 M 为检验的站数，与气候预测室做法一致取中国东部 100 个站；N 为同号的站数；N_0 为距平百分率绝对值<20%异号的站数；N' 为距平百分率在 20%～40%的同号站数；N'' 为距平百分率≥50%的同号站数。S 评分公式虽然存在一定缺陷，但列出来可以与业务预报效果加以比较。

从表 10.3 中看出，预报具有较高的技巧。从距平相关系数衡量，5 年的预报技巧平均为 0.26。从预报评分来看，5 年平均为 79.4，两项平均评分指标在参加全国汛期降水预报会商单位和个人的预报中是比较高的。其中 1994，1997 和 1998 年的预报取得了比较好的效果。

1994 年预报与实况之间的距平相关系数为 0.35，该年我国夏季出现了南北两条明显的多雨带，一条位于华南大部和江南南部，另一条位于北方，多雨区主要位于东北南部、华北北部和西北东部(图 10.4a)。从图 10.4b 中看出，南北两条雨带都预报出来了，位置也基本正确。

1997 年我国夏季长江以北大范围地区出现了异常干旱，只有江南及河套地区降水偏多(图 10.5a)。对于这种降水异常趋势分布，作出了较准确的预报(图 10.5b)，只是多雨的位置有些偏差。

1998 年夏季，我国长江流域降水超常偏多，雨区范围广，降水强度大，持续时间长，造成了百年来罕见的特大洪涝灾害。与此同时，内蒙古东部、东北西部也出现了特大洪涝。而华北中部、华南东部及陕西、甘肃东部降水偏少(图 10.6a)。

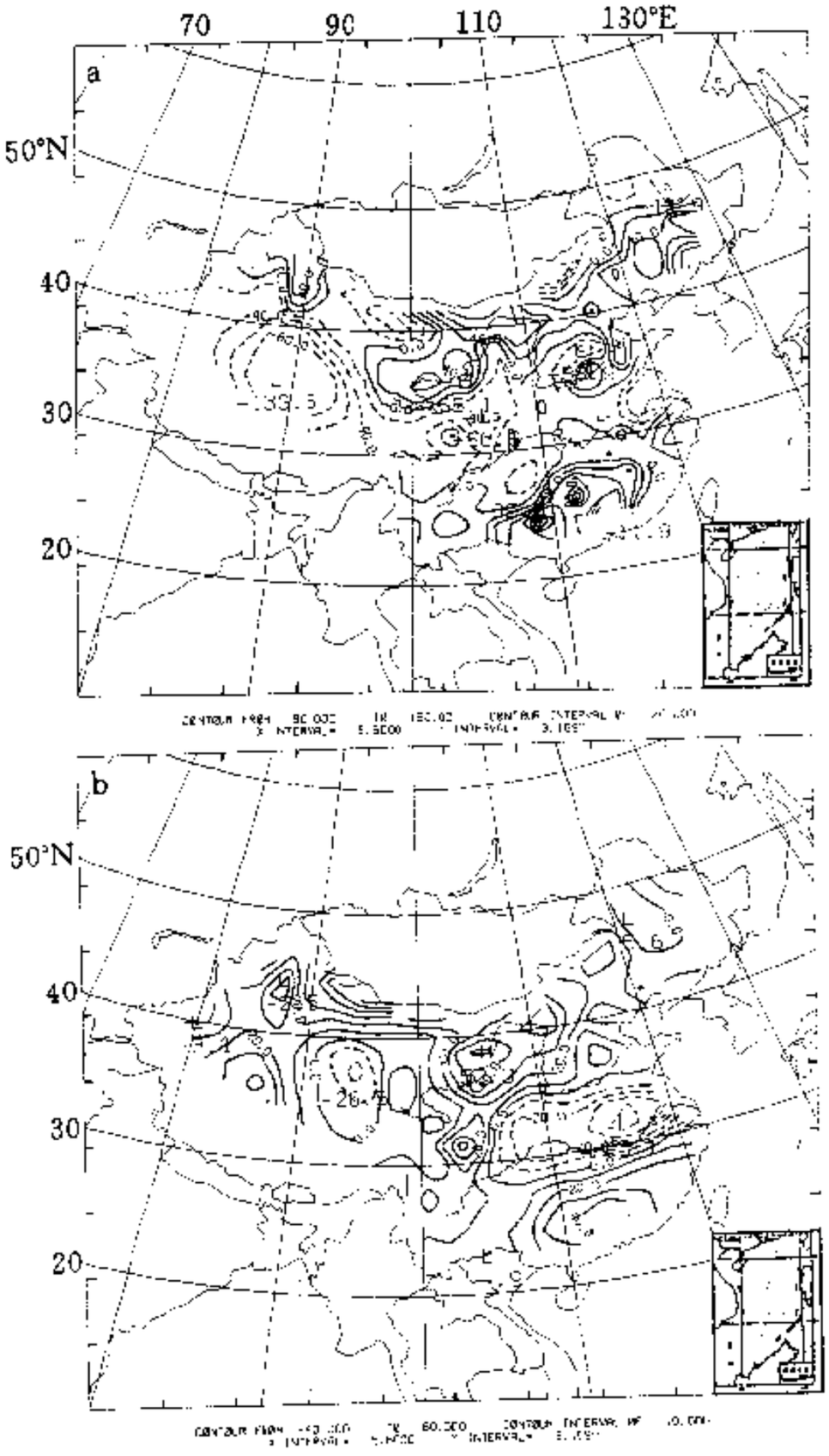


图 10.4　1994 年夏季降水距平百分率(a 为实况;b 为预报)

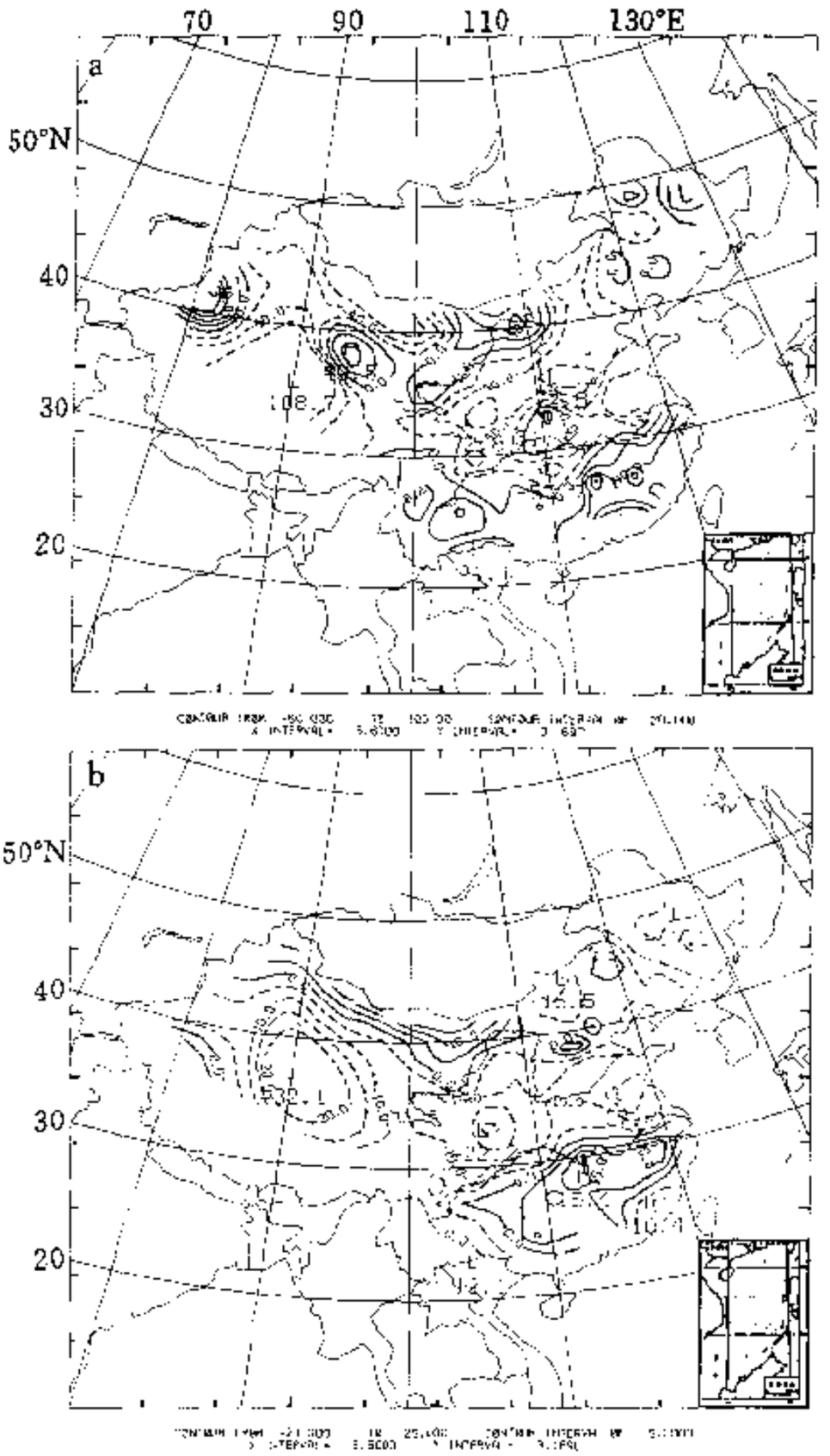


图 10.5　1997 年夏季降水距平百分率(a 为实况;b 为预报)

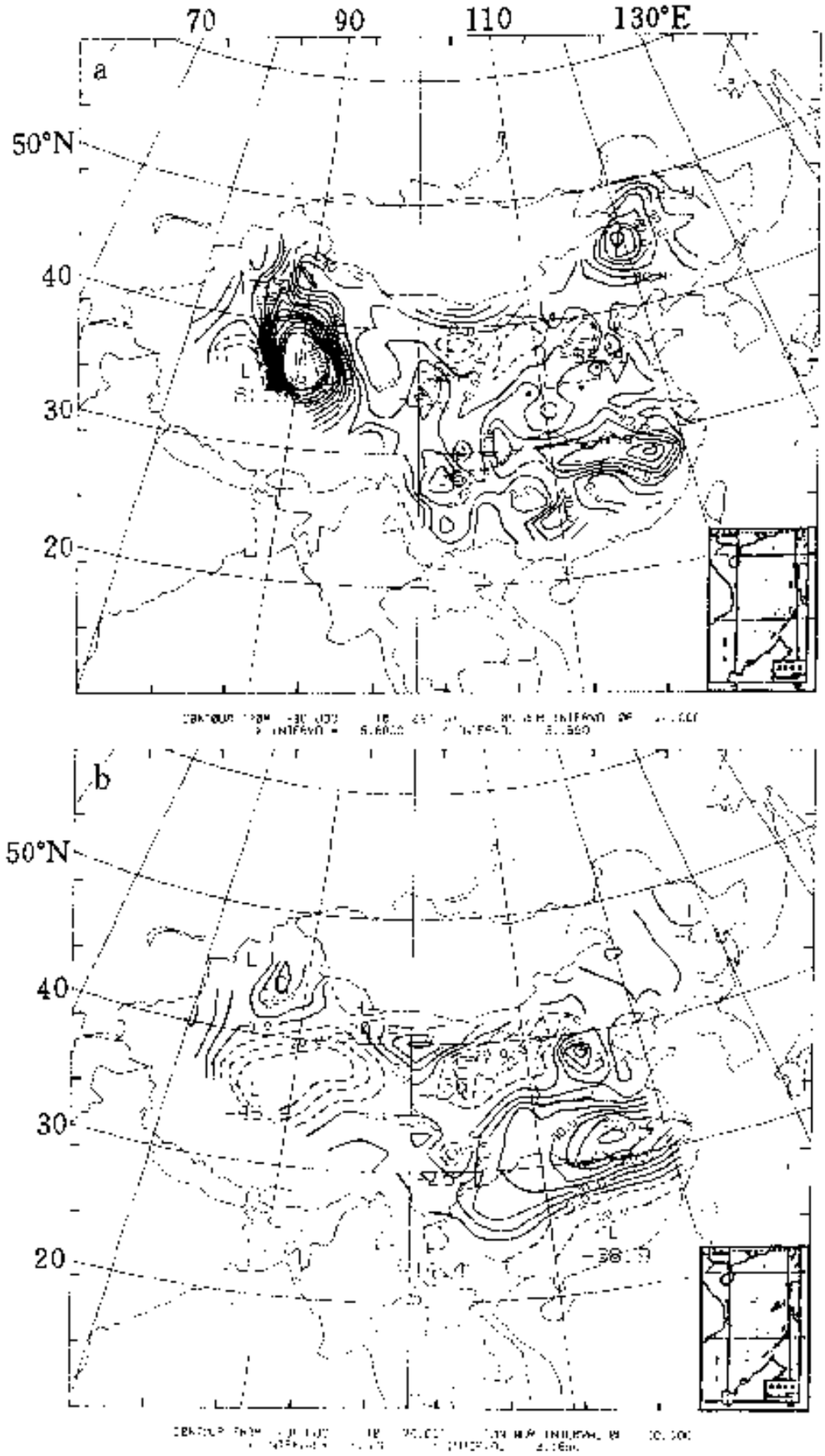


图 10.6　1998 年夏季降水距平百分率(a 为实况;b 为预报)

从图 10.6b 中看出，对于长江流域降水超常偏多趋势，作出了相当不错的预报。降水距平百分率为 30%的等值线覆盖了长江流域，50%等值线的范围也不小，预报的强度之大是不多见的，说明预报模型对于超常偏多趋势有明显反映。不足的是，最大降水距平百分率的中心位置略偏北。另外，内蒙古、东北的多雨，陕西、甘肃的偏少趋势均与实况相符。当然，强度不够。对于降水出现如此异常的趋势分布，预报达到这种程度已经十分令人满意了。

§10.6 最优气候均态模型

前面叙述的是以时间序列各种时间长度均值为基函数的预测模型。这里介绍另一种形式的均值模型。从近几年美国气候预测中心(Climate Prediction Center)发布的气候预测公报中看到，典型相关和最优气候均态(Optimal Climate Normal，OCN)是美国制作短期气候预测的两种常用统计方法[10]。其中 OCN 主要用于温度的预测。其实，OCN 的基本思想并无新意，但在计算上有其独特之处。它是相对于持续性预测概念而言的一种预测。持续性气候预测的概念是，用现时值作为下一时刻的预测值。而最优气候均态预测则是用前 k 个时刻的平均值作为下一时刻的预测值。

气候系统并不是静态不变的。因此，计算平均值所取的平均数 k 过于大，未必能够得到最小误差的预测。按照世界气象组织 WMO 的建议，气候平均值基于一个特定的 30 年，如 1951～1980，1961～1990 年。这样使得世界各地均在一个统一标准下。距平的正负可以明确表示异常冷暖。事实上，许多研究表明，用最近 k 年($k<30$)平均值作为预测，其预测技巧

要比用 30 年平均值好。把 WMO 推荐的均值作下一年预测作为预测评分的一个标准。

OCN 方法的最大特点是计算简便，而预测效果并不比复杂模型差。因此，这里对它的基本做法作一简单介绍。

10.6.1 方法

假设一气候变量序列 $x_i, i=1,2,\cdots,n$。构造序列

$$\bar{x}_{i,k}=\frac{1}{k}\sum_{j=1}^{k}x_{i-j} \qquad (10.6.1)$$

$$(k=1,2,\cdots,n \quad i=n_1+1,n_1+2,\cdots,n_1+L)$$

其中 n_1 为统计基本样本量，通常取为 30 年；k 代表所计算的气候平均的年数；L 为试验样本量；$n=n_1+L$。

(10.6.1)式表示分别求出 $1,2,\cdots,n_1$ 年的平均值，以这些平均值依次作出 $n_1+1,n_1+2,\cdots,n_1+L$ 时刻的预测。再以预测值与实况值最接近为标准，得出试验预测的每个时刻"最优"平均数。以某种准则确定出作下一时刻预测的平均数。

10.6.2 确定最优平均数准则

可以选用以下几种方法确定最优平均数。

10.6.2.1 距平相关系数

$$R(k)=\frac{\sum_{i=k+1}^{n}(\bar{x}_{i,k}-C_{\mathrm{WMO}})(x_i^{obs}-C_{\mathrm{WMO}})}{\sqrt{\sum_{i=k-1}^{n}(\bar{x}_{i,k}-C_{\mathrm{WMO}})^2\sum_{i=k+1}^{n}(x_i^{obs}-C_{\mathrm{WMO}})^2}}$$

$$(k=1,2,\cdots,n_1) \qquad (10.6.2)$$

其中 x_i^{obs} 表示预测年份的观测值。C_{WMO} 为 WMO 推荐的 30 年平均值。(10.6.2)式是对统计样本而言，对独立样本的距平相关系数为：

$$R_{\text{indep}} = \frac{n[R(k)]}{n-1} - \frac{1}{(n-1)R(k)} \qquad (10.6.3)$$

以距平相关系数达到最大为标准，确定最优平均数。

10.6.2.2 绝对误差

$$ABS(k) = \sum_{i=k+1}^{n} |\bar{x}_{i,k} - x_i^{obs}|/(n-k)$$
$$(k = 1,2,\cdots,n_1) \qquad (10.6.4)$$

以 $ABS(k)$ 最小为标准确定最优平均数。

10.6.2.3 均方误差

$$RMS(k) = \sqrt{\sum_{i=k+1}^{n} (\bar{x}_{i,k} - x_i^{obs})/(n-k)} \qquad (10.6.5)$$

以 $RMS(k)$ 达到最小为标准，确定最优平均数。

10.6.2.4 频率指数

在 OCN 的应用中，经反复试验，设计出一种以最优平均数出现的频率，来确定作下一时刻预测的平均数的准则。定义一个指数

$$I(k) = \frac{m(k)}{L} \qquad (10.6.6)$$

其中 $m(k)$ 为相同 k 出现的次数；L 为试验预测次数。以 $I(k)$ 达到最大为标准，确定最优平均数。

10.6.3 计算步骤

这里给出以频率指数为准则的 OCN 的计算步骤：

(1)利用(10.6.1)式计算出 $1,2,\cdots,n_1$ 年的平均值。用这些平均值依次向前作出 $n_1+1, n_1+2, \cdots, n_1+L$ 年的试验预测。具体实施时，以 1 年平均值作为 $n_1+1, n_1+2, \cdots, n_1+L$ 年的预测值，以 2 年平均值作为 $n_1+1, n_1+2, \cdots, n_1+L$ 年的预测值……以 n_1 年的平均值作为 $n_1+1, n_1+2, \cdots, n_1+L$ 年

的预测值。

(2)逐一计算以1年平均值,2年平均值,…,n_1年平均值作为$n_1+1, n_1+2, \cdots, n_1+L$年预测值与观测值之间的绝对值误差。从每一年的预报中挑选出绝对值误差最小的平均值。这样得到L次试验预测的最优平均数。

(3)统计L次试验预测的最优平均数发生的频次$m(k)$,代入(10.6.6)式,以$I(k)$达到最大为准则,确定出预测$n+1$年的最优平均数k。

(4)以最近k年样本的平均值作为$n+1$年的预测。

应用实例[10.4]:用最优气候均态法预测1996年北京1月平均气温。以1961～1990年30年资料作为统计样本量,以1991～1995年的资料为试验预测样本量。

5次试验预测的最优平均数分别为3,3,9,3和3。以频率指数准则确定出预测下一年的最优平均数为3年。根据这一结果,将1993,1994和1995年1月平均气温的平均值−2.0℃作为1996年北京1月平均气温的预测值。这一年的实况值为−2.2℃。

参考文献

[1]魏凤英等.逐步回归周期分析.气象,1983,9(1)

[2]魏凤英等.带有周期分量的多元逐步回归.气象科学研究院院刊,1986,1(1)

[3]魏凤英,曹鸿兴.建立长期预测模型的新方案及其应用.科学通报,1990,35(10):777～780

[4]魏凤英,曹鸿兴.长期预测的数学模型及其应用.北京:气象出版社,1990.49～90

[5]曹鸿兴等.统计模型的双评分准则及其在气象、水文预报中的应用.数理统计与应用概率,1989(1)

[6]曹鸿兴等．多步预测的降水时序模型．应用气象学报，1993，4(2)
[7]魏凤英，曹鸿兴．模糊均生函数模型及其应用．气象，1993，19(2)
[8]魏凤英，张先恭．一种夏季大范围降水趋势分布的预报方法．气象，1995，21(12)
[9]魏凤英，张先恭．中国夏季降水趋势分布的一个客观预报方法．气候与环境研究，1998，3(3)
[10]Huang Jin et al. Long-Lead seasonal temperature prediction using optimal climate normals. J. Climate，1996(9)：809～817

附录 1：

现代气候统计诊断与预测程序使用说明

这里给出本书中所介绍的现代气候统计诊断和预测方法的 FORTRAN 程序的使用说明，提供的程序共计 32 个。这些程序可在 PC486 和 586 等微机上运行，也可以移植到工作站或其它大型计算机上使用。大多数程序在各种高低 FORTRAN 版本中均可以执行，少数程序有高版本的要求。程序按书中内容出现的顺序排列。

1. MV. FOR

程序功能：

用递推算法计算一时间序列的均值和方差。方法描述见本书第二章 2.1 和 2.2 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；X(N)：实型数组，存放原始数据。

(2)输出量。XM(N)：前 N－1 个值为中间均值，第 N 个值为序列的均值；S(N)：前 N－1 个值为中间方差，第 N 个值为序列的方差。

2. IP. FOR

程序功能：

计算序列的偏度系数和峰度系数，并进行检验。方法描述见本书第二章 2.3 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；X(N)：实型数组，存

放原始数据。

(2)输出量。G1:偏度系数;TEST1:在 0.05 显著性水平下偏度系数检验值;G2:峰度系数;TEST2:在 0.05 显著性水平下峰度系数检验值。

3. C01. FOR

程序功能:

计算两变量间的相关系数。方法描述见本书第二章 2.4 节。

参数和数组:

(1)输入量。N:整型数,序列样本量;XY(N,2):实型数组,变量 X 放在前,变量 Y 放在后。

(2)输出量。RHO:两变量间相关系数。

4. C02. FOR

程序功能:

计算多个变量滞后长度为 j 的自互相关系数。方法描述见本书第二章 2.4 节。

参数和数组:

(1)输入量。M:整型数,变量个数;N:整型数,样本量;NE:滞后长度;X(M,N):实型数组,存放原始数据。

(2)输出量。XR(M,M):存放任一滞后长度的自互相关系数。共输出 J 个这样的数组,J=0,1,…,NE。

5. LR. FOR

程序功能:

计算变量 X 与其对应的时间 t 之间的一元线性回归。方法描述见本书第四章 4.1 节。

参数和数组:

(1)输入量。N:整型数序列样本量;X(N):实型数组,存

放序列 X 原始数据；IT(N)：整型数组，存放 X 对应的时间 t。

(2)输出量。A：回归常数；B：回归系数；r：相关系数；LX(N)：存放回归计算值。

6. MM. FOR

程序功能：

对序列作给定滑动长度 K 的滑动求和平均计算。方法描述见本书第四章 4.2 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；IH：整型数，滑动长度；NYEAR：整型数，序列初始年或初始序号；X(N)：实型数组，存放原始数据。

(2)输出量。NNY(N－IH＋1)：年序号或序号；X1(N－IH＋1)：存放滑后序列。

7. CA. FOR

程序功能：

计算序列的累积距平。方法描述见本书第四章 4.3 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；NYEAR：整型数，序列初始年或初始序号；X(N)：实型数组，存放原始数据。

(2)输出量。NYEAR：年份或序号；X1(N)：存放累积距平值。

8. SS. FOR

程序功能：

对序列进行五、七、九点二次平滑。方法描述见本书第四章 4.4 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；NL：整型数，控制参

数，NL＝5 执行五点平滑，NL＝7 执行七点平滑，NL＝9 执行九点平滑；NW：整型数，控制参数，NW＝1 执行平滑，否则不执行平滑；X(N)：实型数组，存放原始数据。

(2)输出量。Y(N)：存放平滑后序列。

9. TS. FOR

程序功能：

对序列进行五点三次平滑。方法描述见本书第四章 4.5 节。

参数和数组：

(1)输入量。N：整型数，序列样本量，N＞5；NYEAR：整型数，序列初始年或初始序号；X(N)：实型数组，存放原始数据。

(2)输出量。XX(N)：存放平滑后序列。

10. SPLINES. FOR

程序功能：

用三次样条函数对给定的序列进行分段曲线拟合。方法描述见本书第四章 4.6 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；M：整型数，所分的段数；N1：整型数，N1＝7×M－3；Y(N)：实型数组，存放原始序列数据；Z(M)：实型数组，存放插入 M 个分点的值；IZ(M)：整型数，存放每段所含的节点个数。

(2)输出量。Y1(＊)：存放拟合序列。

11. MTT. FOR

程序功能：

用滑动 t-检验来检测序列是否发生突变。采用滑动的办法连续设置基准年，依次计算基准年前后两段子序列的平均值，进而得到统计量 t 序列 。方法描述见本书第五章 5.1 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；IH：整型数，子序列长度；NYEAR：整型数，序列初始年或初始序列；Y(N)：实型数组，存放原始数据。

(2)输出量。NNY1：年份或序号；T(N－2＊IH－1)：存放 t 统计量值；A，B：t_α 的临界值。

12. CRAMER. FOR

程序功能：

用比较一个子序列与总体序列平均值的显著差异来检测突变。采用滑动的办法连续设置基准年，依次计算子序列与总序列的平均值，最终得到统计量 t 序列。方法描述见本书第五章 5.2 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；IH：整型数，子序列长度；NYEAR：整型数，序列初始年或初始序号；Y(N)：实型数组，存放原始数据。

(2)输出量。NNY1：序列的年份或序号；T(N－IH＋1)：存放 t 统计量值；A，B：t_α 的临界值。

13. YAMA. FOR

程序功能：

用信噪比来检验两子序列均值的差异是否显著。采用滑动的办法连续设置基准点，计算出信噪比序列。方法描述见本书第五章 5.3 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；IH：整型数，子序列长度；NYEAR：整型数，序列初始年或初始序号；Y(N)：实型数组，存放原序列数据。

(2)输出量。NNY1:序列的年份或序号;SN(N－2＊IH＋1):存放信噪比序列;A_1,A_2:临界值。

14. MK. ROR

程序功能:

按序列的顺序 $x_1, x_2, \cdots, x_n$ 计算一次秩统计量 U,记为 UF。同时,按序列的逆序 $x_n, x_{n-1}, \cdots, x_1$ 计算一次秩统计量 U,记为 UB。UF 和 UB 交点大于临界值,则认为交点处为突变点。方法描述见本书第五章 5.4 节。

参数和数组:

(1)输入量。N:整型数,序列样本量;NYEAR:整型数,序列初始年或初始序号;Y(N):实型数组,存放原序列数据。

(2)输出量。NNY1:序列的年份或序号;UF(N):存放 UF 序列;UB(N):存放 UB 序列;A,B:临界值。

15. LAPAGE. FOR

程序功能:

采用连续设置基准点的办法,以滑动的方式计算基准点前后时段范围内的秩序列,进而得到秩统计量序列。方法描述见本书第五章 5.5 节。

参数和数组:

(1)输入量。N:整型数,序列样本量;IH1:整型数,滑动长度;IH:整型数,子序列长度;NYEAR:整型数,序列初始年或初始序号;Y0(N):实型数组,存放原始序列数据。

(2)输出量。NNY1:序列的序号;HK(＊):统计量序号;T95,T99:95%,99%置信水平值。

16. SA. FOR

程序功能:

计算给定时间序列的连续功率谱,并计算红噪音过程和

白噪音过程进行功率谱显著性检验。方法描述见本书第六章6.1节。

参数和数组：

(1)输入量。N：整型数，序列样本量；IH：整型数，滑动平均长度，若不作滑动 IH=1；M：整型数，最大落后长度，取$\frac{N}{3}$—1；X(N)：实型数组，先作为输入变量，存放序列原始数据。

(2)输出量。C：白噪音值；X(*)：后存放红噪音序列；R1(M+1)：先放自相关系数，后放周期长度；R(M+1)：平滑功率谱密度。

17. MESA. FOR

程序功能：

对给定的时间序列进行最大熵谱分析。方法描述见本书第六章 6.2 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；M：整型数，最大落后时间长度，一般取$\frac{N}{2}$；X(N)：实型数组，存放序列原始数据；NW2：整型数，输出控制水平，NW2=1 时，在屏幕上显示计算过程，NW2≠1 时，无显示。

(2)输出量。K0：自回归阶数；FI(M)：自回归系数；L：落后长度；T(L)：各落后长度的周期长度；SE(L,K0)：对应各落后长度的最大熵谱值。

18. CSA. FOR

程序功能：

对给定的两个时间序列进行交叉谱分析。计算出正交谱、协谱、凝聚谱和位相谱。方法描述见本书第六章 6.3 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；M：整型数，最大落后时间长度；XY(N,2)：实型数组，存放两个时间序列数据。

(2)输出量。P11(M)，P12(M)：序列 X(N)和 Y(N)的功率谱；Q12(M)：正交谱；P12(M)：协谱；R12(M)：凝聚谱；SITA(M)：位相谱；CL(M)：落后时间长度谱。

19. SSA. FOR

程序功能：

对给定的一时间序列按一定的嵌套空间维数生成一资料矩阵。在此基础上进行奇异谱分析。方法描述见本书第六章 6.5 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；M：整型数，嵌套空间维数，X(N)：实型数组，存放原始数据。

(2)输出量。A(M,M)：生成资料矩阵 F(N－M＋1,M)的协方差阵；D(M)：A 阵的特征值；V1(M,M)：A 阵的特征向量；T(M,N)：特征向量的时间系数；H1(M)：累积方差；DLM(M)：特征值误差；DLM1：为 DLM(1)，DLM(2)中最小值；DLMD1：为 ABS|DLM(2)－DLM(1) ；IX(M)：滞后步长；R(M)：滞后相关系数；IX(M) * 4：周期长度。

20. WA. FOR

程序功能：

计算时间序列墨西哥帽状的小波变换。方法描述见本书第六章 6.6 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；M：整型数，时间尺度个数；N5：整型数，开始尺度；KS：整型数，尺度的间隔；

X(N)：实型数组，存放原序列数据；NYEAR：整型数，初始序号(年、月或其它)。

(2)输出量。F(M,N)：小波变换系数；IT(M)：延拓尺度。

21. EOF. FOR

程序功能：

用雅可比方法计算变量场的特征值和特征向量，并求出特征向量对应的时间系数、方差贡献和累积方差贡献。方法描述见本书第七章 7.1 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；M：整型数，站点数或变量个数；JOB：整型数，控制参数，进行经验正交函数分解时所用数据的标记，JOB＝0，用原始数据，JOB＝1，计算距平后展开，JOB＝2，数据标准化后展开。F(N,M)：实型数组，存放原始数据。

(2)输出量。V(M,M)：存放协方差矩阵；IX(M)：特征值序号；D(M)：存放特征值；V1(M,M)：存放特征向量；T(N,M)：存放时间系数；H1(M)：存放累积方差。

22. EEOF. FOR

程序功能：

在建立滞后两个时次的新资料基础上进行扩展经验正交函数分解，从而得到滞后 0 时次、滞后 1 时次和滞后 2 时次的特征值和特征向量。方法描述见本书第七章 7.2 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；M：整型数，站点数或格点数；JT：整型数，落后时间长度；NN：N－2＊JT；M3：3＊M；JOB：整型数，控制参数，分解时所用资料形式，JOB＝0，用原始数据；JOB＝1，用距平数据；JOB＝2，用标准化数据。X

(N,M):实型数组,存放变量场原始数据,F(N,M):实型数组,存放处理后数据。

(2)输出量。D(M3):特征值;A(M3,M3):特征向量;T(N,5):存放前5个特征向量的时间系数;H1:累积方差贡献。

23. REOF. FOR

程序功能:

对给定变量场进行极大方差准则的旋转经验正交函数展开。方法描述见本书第七章7.3节。

参数和数组:

(1)输入量。N:整型数,序列样本量;M:整型数,空间点数;KS:整型数,旋转特征向量个数;JOB:整型数,控制参数,JOB=0,用原始数据展开,JOB=1,用距平形式展开,JOB=2,用标准化形式展开;F(N,M):实型数组,存放原资料场。

(2)输出量。V1(M,KS):旋转特征向量;T(N,KS):旋转特征向量的时间系数;DL(KS):特征值;DC(KS):累积方差。

24. CEOF. FOR

程序功能:

对变量场进行复经验正交函数展开。建立Hermite复矩阵,求此阵的特征值和特征向量。然后计算出有关移动性质的四个量,即空间位相函数,空间振幅函数,时间位相函数,时间振幅函数。方法描述见本书第七章7.4节。

参数和数组:

(1)输入量。IT1:整型数,样本量;M:整型数,测站数或格点数;M1:整型数,取特征向量个数;L:整型数,滤波截断长度;IT4:IT1-L-L;UR(IT1,M1):实型数组,存放原始数据。

(2)输出量。H(*):存放Hibert变换序列;ZR(M,M):存放特征向量实部;ZI(M,M):存放特征向量虚部;TR(IT4,

M)：存放时间系数实部，TI(IT4，M)：存放时间系数虚部；SIT(M，M)：存放空间位相函数；SN(IT4，M)：存放空间振幅函数；FI(IT4，M)：存放时间位相函数；RN(，IT4，M)：存放时间振幅函数。

25. CCA. FOR

程序功能：

在计算第一组变量与第二组变量的相关矩阵基础上计算典型变量和典型相关系数，进而计算出典型变量与原两组变量的相关系数。方法描述见本书第八章 8.2 节。

参数和数组：

(1)输入量。N：整型数，样本量；M1：整型数，第一组变量个数或第一变量场空间点数；M2：整型数，第二组变量个数或第二变量场空间点数；Z1(N，M1)：实型数组，存放第一变量场数据；Z2(N，M2)：实型数组，存放第二变量场数据。

(2)输出量。SR：典型相关系数；XU(M)：存放 U 的组合系数；XT(M)：存放 V 的组合系数；X(N)：逐年典型变量 U；Y(N)：逐年典型变量 V；GU(M1)：典型变量与原第一变量场的相关；GV(M2)：典型变量与原第二变量场的相关。

26. SVD. FOR

程序功能：

在计算左场和右场交叉协方差阵基础上，求解实非对称矩阵的奇异值和奇异向量。根据左奇异向量和右奇异向量分别求出时间系数 A 和 B。进一步计算时间系数 A 和 B 之间相关系数、异性相关系数场和同性相关系数场。方法描述见本书第八章 8.4 节。

参数和数组：

(1)输入量。N：整型数，样本量；L1：整型数，右场空间点

数;L2:整型数,左场空间点数;X1(N,L1):实型数组,存放右场资料数据;X2(N,L2):实型数组,存放左场资料数据。

(2)输出量。B1(L2,L2):存放奇异值,后又放每对奇异向量的解释方差及累积方差;B2(L2,L2):存放左奇异向量;A2(L1,L1):存放右奇异向量;T1(N,10):存放右奇异向量的时间系数;T2(N,10):存放左奇异向量的时间系数;A5(L1,10):存放右场同性相关系数;A6(L2,10):存放左场同性相关系数;FFZ1(10):存放右奇异向量解释原场方差比;FFZ4(10):存放左奇异向量解释原场方差比。

27. SWR.FOR

程序功能:

根据各预报因子方差贡献的显著性检验结果,筛选出配合较好,且方差贡献大的自变量建立回归方程,并计算出回归值及预报值,即实施逐步回归算法。当引入和剔除因子的FF1,FF2 检验值取为 0 时,即为一般的多元回归过程。方法描述见本书第九章 9.1 节。

参数和数组:

(1)输入量。M:整型数,预报因子个数;N:整型数,样本量;NN:整型数,预报长度;L1,L2:整型数,引入回归方程中因子个数的下、上限。FF1,FF2:实型数,引入和剔除预报因子F 检验值。XY(N+NN,M+1):实型数组,前 M 存放预报因子数据,M+1 存放预报量数据。

(2)输出量。F1,F2:引入和剔除因子的 F 检验值;SXY:总离差平方和;STEP:引入和剔除因子步数;K12:引入或剔除因子序号:"--"表示被剔除;F12:该因子的 F 检验值;L:方程目前引入变量的个数;RY12M:复相关系数;F:方程 F 检验值;QQ:残差平方和;VV:回归平方和;ERM:平均绝对误

差；RV：方差缩减；TL(N+NN)：存放回归值和预报值。E(N)：存放残差。

28. OSR. FOR

程序功能：

用 Furnial-Wilson 算法计算预报因子(限制在 30 以内)所有可能的子集回归。用双评分准则确定出最优子集回归，并计算最优回归方程的拟合值和预报值。方法描述见本书第九章 9.2 节。

参数和数组：

(1)输入量。N：整型数，样本量；NN：整型数，预报步数；N1：整型数，$N_1=N+NN$，样本量加待报的因子样本量；IP：整型数，预报因子个数；IIP：整型数，IIP=IP+1，预报因子个数加预报量；XY(N,IIP)：实型数组，存放预报因子和预报量资料，预报因子在前，预报量在后。

(2)输出量。XYM(IIP)：预报因子和预报量的平均值；RMSE：观测值与拟合值间的均方根误差；X1(N1)：存放回归方程拟合值和预报值；X(N)：存放预报量原序列。

29. ERR. FOR

程序功能：

用于建立自变量之间存在复共线性的特征根回归方程。方法描述见本书第九章 9.4 节。

参数和数组：

(1)输入量。NW：整型数，输出控制参数。NW=1，不输出任何结果，NW=2，输出特征根和特征向量及剩余值和回归系数，NW=3，输出剩余值及回归系数，NW=4，输出回归系数。N：整型数，样本量；IP1：整型数，自变量个数和预报量个数之和。YX(N,IP1)，实型数组，存放预报量和自变量资料，

预报量在前，自变量在后。

(2)输出量。SEU：剩余平方和；SS：剩余标准差；RNAM(IP1)：存放特征根；GAM(IP1,IP1)：存放特征向量；B(IP1)：存放回归系数，包括 B_0；F(N)：存放剩余值。

30. RR. FOR

程序功能：

给出偾岭估计的风险小于最小二乘估计的风险条件，并根据已知条件建立岭回归方程。根据岭回归方程计算出回归计算值及预报值。方法描述见本书第九章 9.5 节。

参数和数组：

(1)输入量。N：整型数，样本量；M：整型数，自变量个数；NN：整型数，预报步数；X(N,M+1)：实型数组，前 M 列放自变量数据，M+1 列放预报量数据。XA(NN,M)：实型数组，存放待预报自变量数据。

(2)输出量。B(M)：存放回归系数；P：回归计算值及预报值；D：剩余值；Q：剩余平方和。

31. MGF. FOR

程序功能：

对给定的一时间序列，计算其一阶差分和二阶差分序列，生成原序列及差分序列的均生函数，视其为预报因子。计算粗选出的因子的所有可能的子集回归，利用双评分准则确定出最优子集回归作为预报方程。方法描述见本书第十章 10.1～10.3 节。

参数和数组：

(1)输入量。N：整型数，序列样本量；NN：整型数，预报步数；IN：整型数，预报方程中含均生函数(即预报因子)的最大个数；X(N)：实型数组，存放原序列数据。

(2)输出量:①各阶最优子集回归结果。源程序中有英文提示,主要包括入选的预报因子、回归系数、各阶子集方程及拟合值和预报值;②最终确定的最优子集回归的结果。X1(N+NN):存放方程的拟合值和预报值。

32. OCN. FOR

程序功能:

对有 N 个统计样本量的时间序列分别求出 1,2,…,N 年的平均值。用这些平均值分别作 N+1,N+2,…,N+L 年的预报。以预报值与实况值最接近为标准,得出试验预报的每年"最优"平均年数。再以最优平均年数出现的频率和数值误差为准则,确定出制作下一年预报的平均年数。方法描述见本书第十章 10.6 节。

参数和数组:

(1)输入量。N:整型数,统计样本量;L:整型数,试验预报样本量;NN:整型数,预报长度;NYEAR:整型数,试验预报初始年-1;X(N+L):实型数组,存放统计和试验样本数据。

(2)输出量。XF(L):存放试验预报值;IK(L):存放最优平均年数;I0:预报下一步的最优平均年数。

附录 2：

附表 1a　正态分布表

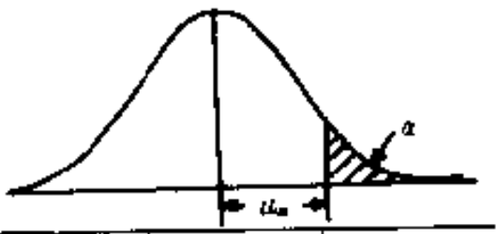
$$\int_{u_\alpha}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=\alpha$$

u_α	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.00990	.00964	.00939	.00914	.00889	.00866	.00842
2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
u_α	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
3	.00135	0^3968	0^3687	0^3483	0^3337	0^3233	0^4159	0^3108	0^4723	0^4481
4	0^4317	0^5207	0^4133	0^5854	0^5541	0^5310	0^5211	0^5130	0^6793	0^6479
5	0^6287	0^6170	0^7996	0^7579	0^7333	0^7190	0^7107	0^8599	0^8332	0^8182
6	0^9987	0^9530	0^9282	0^9149	$0^{10}777$	$0^{10}402$	$0^{10}206$	$0^{10}104$	$0^{11}523$	$0^{11}260$

注：本表除 u_α 一栏外，小数点前省略 0。

附表 1b　u_α值表

$$\frac{1}{\sqrt{2\pi}}\int_{u_\alpha}^{\infty} e^{-\frac{x^2}{2}}dx=\alpha$$



α	0	1	2	3	4
0.00	∞	3.09023	2.87816	2.74778	2.65207
0.0	∞	2.32635	2.05375	1.88079	1.75069
0.1	1.28155	1.22653	1.17499	1.12639	1.08032
0.2	0.84162	0.80642	0.77219	0.73885	0.70630
0.3	0.52440	0.49585	0.46770	0.43991	0.41246
0.4	0.25335	0.22754	0.20189	0.17637	0.15097

α	5	6	7	8	9
0.00	2.57583	2.51214	2.45726	2.40892	2.36562
0.0	1.64485	1.55477	1.47579	1.40507	1.34076
0.1	1.03643	0.99446	0.95417	0.91537	0.87790
0.2	0.67449	0.64335	0.61281	0.58284	0.55338
0.3	0.38532	0.35846	0.33185	0.30548	0.27932
0.4	0.12566	0.10043	0.07527	0.05015	0.02507

注：当 $\alpha>0.5$，$u_\alpha=-u_{1-\alpha}$，当 $\alpha=0.5$，$u_{0.5}=0$。

附表 2　t 分布表

ν	0.10	0.05	0.02	0.01	0.001
1	6.31	12.71	31.82	63.66	636.62
2	2.92	4.30	6.97	9.93	31.60
3	2.35	3.18	4.54	5.84	12.94
4	2.13	2.78	3.75	4.60	8.61
5	2.02	2.57	3.37	4.03	6.86
6	1.94	2.45	3.14	3.71	5.96
7	1.90	2.37	3.00	3.50	5.41
8	1.86	2.31	2.90	3.36	5.04
9	1.83	2.26	2.82	3.25	4.78
10	1.81	2.23	2.76	3.17	4.59
11	1.80	2.20	2.72	3.11	4.44
12	1.78	2.18	2.68	3.06	4.32
13	1.77	2.16	2.65	3.01	4.22
14	1.76	2.15	2.62	2.98	4.14
15	1.75	2.13	2.60	2.95	4.07
16	1.75	2.12	2.58	2.92	4.02
17	1.74	2.11	2.57	2.90	3.97
18	1.73	2.10	2.55	2.88	3.92
19	1.73	2.09	2.54	2.86	3.88
20	1.73	2.09	2.53	2.85	3.85
21	1.72	2.08	2.52	2.83	3.82
22	1.72	2.07	2.51	2.82	3.79
23	1.71	2.07	2.50	2.81	3.77
24	1.71	2.06	2.49	2.80	3.75
25	1.71	2.06	2.48	2.79	3.73
26	1.71	2.06	2.48	2.78	3.71
27	1.70	2.05	2.47	2.77	3.69
28	1.70	2.05	2.47	2.76	3.67
29	1.70	2.04	2.46	2.76	3.66
30	1.70	2.04	2.46	2.75	3.65
40	1.68	2.02	2.42	2.70	3.55
60	1.67	2.00	2.39	2.66	3.46
120	1.66	1.98	2.36	2.62	3.37
∞	1.65	1.96	2.33	2.58	3.29

附表 3a　F 分布表($\alpha=0.25$)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	60	∞
1	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.41	9.49	9.58	9.76	9.85
2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.41	3.43	3.46	3.48
3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.46	2.46	2.47	2.47
4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.87	1.87
6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.76	1.76	1.74	1.74
7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69	1.68	1.68	1.67	1.65	1.65
8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.64	1.63	1.62	1.62	1.61	1.59	1.58
9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.54	1.53
10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.53	1.52	1.50	1.48
11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51	1.50	1.49	1.47	1.45
12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.48	1.47	1.44	1.42
13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.46	1.45	1.42	1.40
14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.40	1.38
15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.43	1.41	1.38	1.36
16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.41	1.40	1.36	1.34
17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.41	1.40	1.39	1.35	1.33
18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40	1.39	1.38	1.34	1.32
19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.38	1.37	1.33	1.30
20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.37	1.36	1.32	1.29
21	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.31	1.28
22	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37	1.36	1.34	1.30	1.28
23	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.30	1.27
24	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36	1.35	1.33	1.29	1.26
25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.28	1.25
30	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34	1.32	1.30	1.26	1.23
40	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.31	1.30	1.28	1.22	1.19
60	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.27	1.25	1.19	1.15
120	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.16	1.10
∞	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.12	1.00

附表 3b　F 分布表($\alpha=0.05$)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	60	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	252.2	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.48	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.57	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.69	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.43	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.74	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.30	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.01	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.79	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.62	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.49	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.38	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.30	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.22	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.16	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.11	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.06	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.02	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	1.98	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	1.95	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	1.92	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	1.89	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	1.86	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.84	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.82	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.74	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.64	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.53	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.43	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.32	1.00

附表 3c　F 分布表($\alpha=0.01$)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	60	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6313	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.48	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.32	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.65	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.15	10.05	9.89	9.72	9.55	9.20	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.06	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.82	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.03	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.48	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.08	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.78	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.54	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.34	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.18	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.05	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	2.93	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	2.83	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.75	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.67	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.61	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.55	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.50	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.45	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.40	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.36	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.21	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.02	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	1.84	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.66	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.47	1.00

附表 4　χ^2 分布表

ν	0.99	0.98	0.95	0.90	0.50	0.10	0.05	0.02	0.01	0.001
1	0.000	0.001	0.004	0.016	0.455	2.71	3.84	5.41	6.64	10.83
2	0.020	0.040	0.103	0.211	1.386	4.61	5.99	7.82	9.21	13.82
3	0.115	0.185	0.352	0.584	2.366	6.25	7.82	9.84	11.34	16.27
4	0.297	0.429	0.711	1.064	3.357	7.78	9.49	11.67	13.28	18.47
5	0.554	0.752	1.145	1.610	4.351	9.24	11.07	13.39	15.09	20.52
6	0.872	1.134	1.635	2.204	5.35	10.65	12.59	15.03	16.81	22.46
7	1.239	1.564	2.167	2.833	6.35	12.02	14.07	16.62	18.48	24.32
8	1.646	2.032	2.733	3.490	7.34	13.36	15.51	18.17	20.09	26.13
9	2.088	2.532	3.325	4.168	8.34	14.68	16.92	19.68	21.67	27.88
10	2.558	3.059	3.940	4.865	9.34	15.99	18.31	21.16	23.21	29.59
11	3.05	3.61	4.57	5.58	10.34	17.28	19.68	22.62	24.73	31.26
12	3.57	4.18	5.23	6.30	11.34	18.55	21.03	24.05	26.22	32.91
13	4.11	4.76	5.89	7.04	12.34	19.81	22.36	25.47	27.69	34.53
14	4.66	5.37	6.57	7.79	13.34	21.06	23.69	26.87	29.14	36.12
15	5.23	5.99	7.26	8.55	14.34	22.31	25.00	28.26	30.58	37.70
16	5.81	6.61	7.96	9.31	15.34	23.54	26.30	29.63	32.00	39.25
17	6.41	7.26	8.67	10.09	16.34	24.77	27.59	31.00	33.41	40.79
18	7.02	7.91	9.39	10.87	17.34	25.99	28.87	32.35	34.81	42.31
19	7.63	8.57	10.12	11.65	18.34	27.20	30.14	33.69	36.19	43.82
20	8.26	9.24	10.85	12.44	19.34	28.41	31.41	35.02	37.57	45.32
21	8.90	9.91	11.59	13.24	20.34	29.61	32.67	36.34	38.93	46.80
22	9.54	10.60	12.34	14.04	21.34	30.81	33.92	37.66	40.29	48.27
23	10.20	11.29	13.09	14.85	22.34	32.01	35.17	38.97	41.64	49.73
24	10.86	11.99	13.85	15.66	23.34	33.20	36.42	40.27	42.98	51.18
25	11.52	12.70	14.61	16.47	24.34	34.38	37.65	41.57	44.31	52.62
26	12.20	13.41	15.38	17.29	25.34	35.56	38.89	42.86	45.64	54.05
27	12.88	14.12	16.15	18.11	26.34	36.74	40.11	44.14	46.96	55.48
28	13.56	14.85	16.93	18.94	27.34	37.92	41.34	45.42	48.28	56.89
29	14.26	15.57	17.71	19.77	28.34	39.09	42.56	46.69	49.59	58.30
30	14.95	16.31	18.49	20.60	29.34	40.26	43.77	47.96	50.89	59.70

附表 5　检验相关系数 $\rho=0$ 的临界值表

$P(|r|>r_\alpha)=\alpha$

n \ α	0.10	0.05	0.02	0.01	0.001
1	.98769	.99692	.999507	.999877	.9999988
2	.9000	.95000	.98000	.99000	.99900
3	.8054	.8783	.93433	.95873	.99116
4	.7293	.8114	.8822	.91720	.97406
5	.6694	.7545	.8329	.8745	.95074
6	.6215	.7067	.7887	.8343	.92493
7	.5822	.6664	.7498	.7977	.8982
8	.5494	.6319	.7155	.7646	.8721
9	.5214	.6021	.6851	.7348	.8471
10	.4973	.5760	.6581	.7079	.8233
11	.4762	.5529	.6339	.6835	.8010
12	.4575	.5324	.6120	.6614	.7800
13	.4409	.5139	.5923	.6411	.7603
14	.4259	.4973	.5742	.6226	.7420
15	.4124	.4821	.5577	.6055	.7246
16	.4000	.4683	.5425	.5897	.7084
17	.3887	.4555	.5285	.5751	.6932
18	.3783	.4438	.5155	.5614	.6787
19	.3687	.4329	.5034	.5487	.6652
20	.3598	.4227	.4921	.5368	.6524
25	.3233	.3809	.4451	.4869	.5974
30	.2960	.3494	.4093	.4487	.5541
35	.2746	.3246	.3810	.4182	.5189
40	.2573	.3044	.3578	.3932	.4896
45	.2428	.2875	.3384	.3721	.4648
50	.2306	.2732	.3218	.3541	.4433
60	.2108	.2500	.2948	.3248	.4078
70	.1954	.2319	.2737	.3017	.3799
80	.1829	.2172	.2565	.2830	.3568
90	.1726	.2050	.2422	.2673	.3375
100	.1638	.1946	.2301	.2540	.3211

注：本表小数点前省略 0。

附表 6　标准正态分布曲线下的面积

Z	00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.499767									
3.6	.499841									
3.7	.499892									
3.8	.499928									
3.9	.499952									
4.0	.499968									
4.1	.499979									
4.2	.499987									
4.3	.499991									
4.4	.499995									
4.5	.499997									
4.6	.499998									
4.7	.499999									
4.8	.499999									
4.9	.500000									

附表 7　Lillifors 检验临界值表

n	E			
	0.10	0.05	0.02	0.10
3	0.367	0.376	0.381	0.383
4	0.345	0.375	0.400	0.413
5	0.318	0.343	0.375	0.397
6	0.297	0.323	0.352	0.370
7	0.280	0.304	0.332	0.350
8	0.265	0.288	0.315	0.333
9	0.252	0.274	0.299	0.317
10	0.241	0.261	0.286	0.303
11	0.231	0.251	0.275	0.291
12	0.222	0.242	0.265	0.281
13	0.214	0.233	0.255	0.271
14	0.207	0.226	0.247	0.262
15	0.201	0.219	0.240	0.254
16	0.195	0.213	0.233	0.247
17	0.190	0.207	0.227	0.241
18	0.185	0.201	0.221	0.234
19	0.180	0.196	0.215	0.229
20	0.176	0.192	0.210	0.223
21	0.172	0.187	0.206	0.218
22	0.168	0.183	0.201	0.214
23	0.165	0.180	0.197	0.209
24	0.162	0.176	0.193	0.205
25	0.158	0.173	0.189	0.201
26	0.156	0.169	0.186	0.198
27	0.153	0.167	0.183	0.194
28	0.150	0.164	0.180	0.191
29	0.148	0.161	0.177	0.188
30	0.146	0.158	0.174	0.185

[General Information]
书名=现代气候统计诊断与预测技术
作者=
页数=269
SS号=10103896
出版日期=

目录
正文