现代数学方法在序列数据处理与解释中的应用

Xiandai Shuxue Fangfa Zai Xulie Shuju Chuli Yu Jieshizhong De Yingyong

刘 诚 著

西南财经大学出版社
Southwestern University of Finance & Economics Press

现代数学方法在序列数据处理与解释中的应用

Xiandai Shuxue Fangfa Zai Xulie Shuju Chuli Yu Jieshizhong De Yingyong

刘　诚 著

西南财经大学出版社
Southwestern University of Finance & Economics Press

图书在版编目(CIP)数据

现代数学方法在序列数据处理与解释中的应用/刘诚著. —成都:西南财经大学出版社,2015.5
ISBN 978-7-5504-1898-1

Ⅰ.①现… Ⅱ.①刘… Ⅲ.①数学—应用—数据处理
Ⅳ.①N37

中国版本图书馆 CIP 数据核字(2015)第 096969 号

现代数学方法在序列数据处理与解释中的应用

刘 诚 著

责任编辑:张 岚
助理编辑:高 玲
责任校对:魏上兰
封面设计:张姗姗
责任印制:封俊川

出版发行	西南财经大学出版社(四川省成都市光华村街 55 号)
网 址	http://www.bookcj.com
电子邮件	bookcj@foxmail.com
邮政编码	610074
电 话	028-87353785 87352368
照 排	四川胜翔数码印务设计有限公司
印 刷	四川森林印务有限责任公司
成品尺寸	148mm×210mm
印 张	8.625
字 数	215 千字
版 次	2015 年 5 月第 1 版
印 次	2015 年 5 月第 1 次印刷
书 号	ISBN 978-7-5504-1898-1
定 价	48.00 元

序

数学是科学研究的重要工具。随着数学方法研究的深入化和应用领域的广泛化，科学研究从定性分析向定量分析的转变已成必然趋势，数学的用量已逐渐成为衡量研究价值的指标之一。本书是作者根据自己多年的学习和研究，分别从地学、生物学、经济学三个方面出发，系统地介绍了几个常用的重要的现代数学方法以及它们在序列数据处理与解释中的应用。

本书从工程应用角度出发，将现代数学方法和传统数据处理方法相结合，简明扼要地介绍了现代数学方法在序列数据处理与解释中的应用。全书共五章，主要内容有：现代数学方法与研究背景介绍，现代数学方法在处理地学、生物学、经济学中序列数据的原理和方法，包括储层物性参数预测、人工地震多次波分离、小麦条锈病预测、胎儿体重预测、生物医学信号降噪、经济时序数据降噪和经济预测等。

作为人工智能算法的典型代表，神经网络经过60年的发展，现已有超过40种的网络算法，其中包括BP网络、自组织映射、Hopfield网络、波尔兹曼机、适应谐振理论等非常典型和常用的算法。这些方法被广泛应用于自动控制、最优化、模式识别、图像处理、医疗等领域。独立分量分析是由盲源分离技术发展来的一种新的多维数据处理方法。它是从序列数据的高阶统计特性出发，提取其中的独立成分，从而达到对信号分解

的目的。它作为新兴算法，虽然发展时间短，但其取得的成绩却是不容忽视的。新的算法不断被提出，模型也开始向非线性发展，应用领域也在不断扩大。我国在这方面起步虽然较晚，但在应用方面却取得了不错的成果。发展近20年，支持向量机因在解决小样本、非线性及高维模式识别等问题中表现出许多特有的优势，能够有效避免经典学习方法中出现的过学习、欠学习、“维数灾难”以及陷入局部极小点等诸多问题，被广泛应用于模式识别、回归估计和概率密度函数估计等领域。灰色系统由邓聚龙于1982年提出，到现在已30余年。它在处理贫信息建模和预测方面展示了独特优势，尤其为国民经济的发展做出了很大贡献。聚类分析作为传统的数据处理方法，其应用仍然经久不衰，一直在数据处理领域体现着应用价值。

地质条件的高度非线性，勘探手段的高度复杂化，勘探领域的深度化和广度化，使得勘探数据中大量有效信息难以被发现和提取。储层评价仍是油气勘探的一个重要方面。地震勘探中对多次波的研究不仅没有消退，反而更加深入。因此对现代数学方法进行研究，并将其引入到油气勘探中具有非常重要的现实意义。

作物病虫害严重影响到我国粮食的安全和品质，因此对小麦条锈病发病率的精确预测具有重要意义。它不仅可以有效预防和控制小麦条锈病的发生，还可以提高农业生产中的管理水平，发展精准农业，减少病害损失，提高农产品的产量和品质。

生物医学信号中关键信息的提取，是临床医学中重要的研究内容。胎儿体重的精确预测，对产科的产前护理、分娩方式的选择、减少产科并发症，具有十分重要的意义。提高心电信号的分辨率，对于特征信号的提取，病情的分析和诊断，有着重要的实际意义。

经济分析和经济预测的定量化分析，已成为经济研究的重

要内容。如何对经济数据进行降噪，如何进行高精度的经济预测，对于经济决策至关重要。

现代数学方法是一门旨在应用的科学。本书略去了繁琐的数学推导和背景情况介绍，直接利用实例来阐述相关数学方法的基本概念及应用方法和分析技术，对要解决的关键性理论和实际问题分析透彻。本书在每一章节，都分别针对某一问题，利用相关数学方法，解决其中的关键问题。一些研究成果具有开创性和先进性，这些成果均是著者长期研究的积累。本书内容充实，观点鲜明，论述简明扼要，具有广泛的参考价值，可作为相关专业工程技术人员的参考用书。应当指出，由于著者水平有限，本书缺点错误在所难免，不妥之处望广大读者批评指正。

本书获国家自然科学基金项目“基于机场场面非视距信道建模的 MLAT 定位算法研究”（项目编号：U1433129）支持。

刘　诚

2014 年 12 月

目　录

1 绪论

数学是科学研究的重要工具，序列数据处理作为科学研究中数据处理最为常见的模式，更离不开数学。最近半个多世纪，涌现出了大量的现代数学方法，这些方法被广泛应用到各个科学领域。本书是作者多年来学习和研究的成果，分别从地学、生物（医）学、经济学三个领域出发，着重阐述了几个常用的重要的现代数学方法在序列数据处理与解释中的应用，目的在于总结和抛砖引玉。本书附录中给出了书中部分主要算法的Matlab源程序。

本章的主要内容有：

（1）对几个常用现代数学方法进行综述，包括人工神经网络、独立分量分析、支持向量机、灰色系统、聚类分析等，介绍这些方法的原理，简述研究进展，总结方法的特点和改进的方向。

（2）介绍本书的研究背景，主要包括：①测井空间序列数据的预测与反演，人工地震时间序列数据的处理与解释；②农业病虫害时间序列数据处理与预测，生物医学信号的降噪处理与分辨率提高；③经济时间序列数据的降噪处理，股票序列数据的处理与分类分析。

（3）本书的主要研究内容与结构安排。

1.1 现代数学方法研究综述

1.1.1 人工神经网络

人工神经网络模型（ANN，the Artificial Neural Network）于20世纪50年代由心理学家W. S. 麦克洛克（W. S. McCulloch）和数理逻辑学家W. 皮特（W. Pitts）建立。作为一种人工智能算法，经过半个多世纪的发展，人工神经网络以其自学习、联想存储和高速寻优的特点，取得了很大的发展和应用，在智能算法领域已具有举足轻重的作用。它目前已有超过40种的网络算法，其中包括BP网络、自组织映射、Hopfield网络、波尔兹曼机、适应谐振理论等非常典型和常用的算法。其应用也涉及各个领域，包括自动控制、最优化、模式识别、图像处理、机器控制、医疗等。

BP网络是众多算法中使用率较高的一种，它属于多层前馈型网络。其基本思想是，学习过程由信号的正向传播和误差的反向传播两个过程组成。正向传播时，输入样本从输入层传入，经各隐层逐层处理后，传向输出层。若输出层的输出与期望输出（导师信号）不符，则转入误差的反向传播阶段。误差反传是将输出误差以某种形式通过隐层向输入层逐层反传，并将误差分摊给各层的所有单元，从而获得各层的误差信号，此误差信号即作为修正各单元权值的依据。此过程一直进行到网络输出的误差减少到可以接受的程度，或进行到预先设定的学习次数为止。BP网络的拓扑结构如图1.1所示。

理论上可以证明，将BP算法用于具有非线性转移函数的三层前馈网，可以以任意精度逼近任何非线性函数。然而标准的

BP 算法在应用中暴露出不少的缺陷：

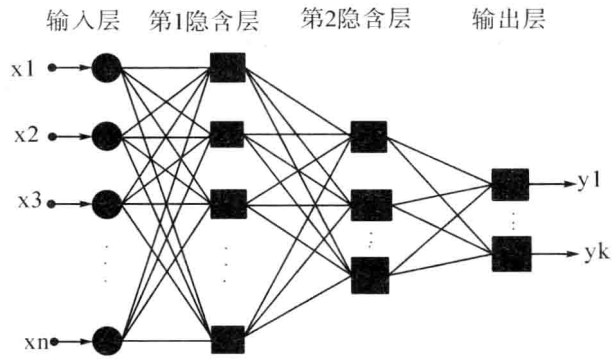

图 1.1　多层前馈神经网络

（1）易形成局部最小而得不到全局最优。

（2）训练次数多使得学习效率低，收敛速度慢。

（3）隐节点的选取缺乏理论指导。

（4）训练时学习新样本有遗忘旧样本的趋势。

针对以上问题，国内外已经提出不少有效的改进算法，以下几种是常用的改进算法：

1. 增加动量项[3]

一些学者提出，标准 BP 算法在调节权值时，只按 t 时刻误差的梯度下降方向调整，没有考虑 t 时刻以前的梯度方向，从而使得训练过程发生振荡，收敛减慢。为了提高网络的收敛速度，可以在权值调整公式中增加一动量项，即：

$$\Delta W(t) = \eta\delta X + \alpha\Delta W(t-1) \tag{1.1}$$

可以看出，增加动量项即从前一次权值调整量中取出一部分加到本次权值调整量中，α 称为动量系数，一般有 $\alpha \in (0, 1)$，动量项反映了以前积累的调整经验，对于 t 时刻的调整起阻尼作用。当误差曲面出现骤然起伏时，可以减少振荡趋

势，提高训练速度。

2. 自适应调节学习率[3]

学习率 η 也称为步长，在标准 BP 算法中定为常数，然而在实际应用中，很难确定一个从始至终都合适的最佳学习率。为了加快收敛速度，一个较好的思路就是自适应改变学习率，使其该大时增大，该小时减小。

通常，设一初始学习率，若经过一批次权值调整后使总误差 $E_{总}$ 增加，则本次调整无效，且 $\eta=\beta\eta(\beta<0)$；若经过一批次权值调整后使总误差 $E_{总}$ 减少，则本次调整有效，且 $\eta=\theta\eta(\theta>0)$。

3. 隐节点调整的基本思想

首先，我们根据学习样本的容量选取较少数目的隐节点，组成网络进行学习训练，根据参数预测的具体情况，选择了每间隔 150 次迭代，来考察误差 E 的变化是否满足：

$$\Delta E=\left|\frac{E_{150\cdot i}-E_{150\cdot(i-1)}}{E_{150\cdot i}}\right|<\varepsilon \tag{1.2}$$

式中：ε 为预定的误差，其值为 0.1.

若满足则网络继续训练，否则网络增加若干隐节点，然后继续训练网络，直到满足网络精度要求或者隐节点数超过某一上限为止，实验统计显示上限值为样本数的两倍。我们采用随机赋值的方法对新增的隐节点赋以初始权值。网络训练完成后，将多余节点删除。

4. 调整误差优化方法

传统的 BP 网络，其误差函数是通过牛顿法来进行优化的，实际上我们可以选用其他更好的优化算法，来提高整个网络的寻优速度，同时提高求解全局最优解的能力。

1.1.2 独立分量分析

独立分量分析[25]（ICA，Independent Component Analysis）是近年来由盲源分离技术（BSS，Blind Source Separation）发展而来的一种新的多维信号处理方法，其基本思路是将多维观察信号按照统计独立的原则建立目标函数，通过优化算法将观测信号分解为若干独立成分，从而帮助实现信号的增强和分析。ICA 从多维观测数据的高阶统计特性出发，提取其中的独立成分，从而使得分解结果更具实际意义。与传统的二阶空间去相关技术[26][27][28]相比，ICA 不仅可以去除各分量之间的一阶、二阶相关性，同时还具有发掘并去除数据间的高阶相关信息的能力，使得输出分量之间相互独立。因此 ICA 可以被看作二阶空间去相关技术的一种扩展。

ICA 的发展经过了一个曲折的过程。1986 年，在还没有出现 ICA 这一名字之前，西班牙学者珍妮·埃罗（Jeanny Herault）和克里斯汀·朱德（Christian Juten）在美国犹他州举行的一次关于神经网络计算的会议上发表了名为《神经网络模型的空时自适应信号处理》[29]的论文。他们在论文中建立了一种基于神经网络和 Hebb 学习规则的新的计算方法，使用这种方法可以实现独立信号混合的盲分离。这是最早的独立分量分析技术的雏形。而在随后的很长一段时间内，ICA 的研究基本上只限于法国。直到 1994 年，法国学者 P. 科蒙（P. Comon）才比较系统地阐述了 ICA 的概念并构造出了一种基于高阶统计量的目标函数[30]。1995 年，A. J. 贝尔（A. J. Bell）和 T. J. 索诺斯基（T. J. Sejnowski）从信息论的角度说明了盲源分离问题，并且证明了神经网络输出信息熵的最大化就意味着输入和输出之间的互信息最小化；同时，他们还使用随机梯度下降学习算法，构造了熵的最大化实现，这就是通常所称的信息最大化 ICA 算法

(Infomax)[31]。虽然这一方法只对处理超高斯信号有效，但它在当时还是产生了很大的影响，从此 ICA 的发展逐渐加快。1997 年，S. I. 阿马里（S. I. Amari）进一步证实，使用自然梯度的 Infomax 算法可以使算法的计算量减小，并说明了它和最大似然法间的联系[32]。1998 年，李泰源（Te-Won Lee）通过和马克·基洛拉米（Mark Girolami）等人的合作对 Infomax ICA 方法做了扩展，使它可以用来处理一般的非高斯信号[33]，包括超高斯信号和欠高斯信号。随后 A. 于瓦里宁（A. Hyvarinen）和 E. 奥亚（E. Oja）提出了一种名为快速 ICA 的固定点算法[34]。这种算法计算简单且有很好的收敛性质，它极大地促进了 ICA 在各种领域的实际应用研究。

ICA 的应用范围非常广泛，并有进一步扩大的势头。ICA 的应用首先是从对生物医学信号的处理开始的。1996 年，马克格（Makeig）等人使用 Infomax 算法对 EEG 和 ERP 数据进行了处理[35]，实验显示这种算法有一定效果。随后 ICA 的应用又扩展到图像处理、语音信号处理方面。近年来，随着人们研究的不断深入，ICA 在数据压缩、图像处理、模式识别、通信以及经济等领域的研究成果[36][37][38]也越来越多。

ICA 发展的时间虽然很短，但其取得的成绩却是不容忽视的。在理论方面，新的算法不断被提出，ICA 模型也从开始的线性模型向非线性模型发展[39]。在实际应用方面，其范围也在不断扩大，并且随着一些新算法的出现，其应用研究也逐渐从理想条件下的研究向更加实用的方面发展。从目前国际上的发展情况来看，美国、法国、芬兰、日本在 ICA 方面的研究处于领先地位。我国在 ICA 方面的研究起步比较晚，且其研究主要集中在应用上，特别是在生物医学方面的应用研究。最近几年，不少学者专家将 ICA 应用到地学上，也取得不小的成绩。其中主要体现在对地震资料进行分解识别[40]、对地震灾害系统中声

波/振动信号进行分离[41][42]、地震数据去噪[43][44][45][46]，另外ICA在地震信号多次波压制应用中也取得了初步成绩，但是仍然有许多需要改进和进一步完善的地方[47][48][49]。

1.1.3 支持向量机

支持向量机（SVM，Support Vector Machine）是科尔特斯（Cortes）和万普尼克（Vapnik）在1995年最先提出的。它建立在统计学习理论和结构风险最小化原理的基础上，通过寻求结构化风险最小化来提高学习机泛化能力，以实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。

支持向量机通过控制超平面的间隔度量来抑制过拟合；通过采用核函数巧妙地解决维数问题以降低运算量。因此SVM在解决小样本、非线性及高维模式识别等问题中表现出许多特有的优势，能够有效避免经典学习方法中出现的过学习、欠学习、“维数灾难”以及陷入局部极小点等诸多问题，被广泛应用于模式识别、回归估计和概率密度函数估计等领域。

支持向量机有如下几个特点：

（1）非线性映射是SVM方法的理论基础，SVM利用内积核函数代替向高维空间的非线性映射。

（2）对特征空间划分的最优超平面是SVM的目标，最大化分类边际的思想是SVM方法的核心。

（3）支持向量是SVM的训练结果，是在SVM分类决策中起决定作用的支持向量。

（4）SVM是一种有坚实理论基础的新颖的小样本学习方法。它基本上不涉及概率测试及大数定律等，因此不同于现有的统计方法。从本质上看，它避开了从归纳到演绎的传统过程，实现了高效的从训练样本到预报样本的“转导推理”，大大简化

了通常的分类和回归等问题。

（5）SVM 的最终决策函数只由少数的支持向量所确定，计算的复杂性取决于支持向量的数目，而不是样本空间的维数，从而避免了“维数灾难”。

（6）少数支持向量决定了最终结果，可以帮助我们抓住关键样本、“剔除”大量冗余样本，这注定了该方法不但算法简单，而且具有较好的“鲁棒”性。SVM 的“鲁棒”性主要体现在以下几个方面：

①增、删非支持向量样本对模型没有影响。

②支持向量样本集具有一定的鲁棒性。

虽然支持向量机有以上诸多优点，但这种算法也存在着一定的缺陷，如针对大规模训练样本难以实施，并且在求解上花费的训练时间较长等。最小二乘支持向量机是支持向量机学习算法的重要扩展，用于解决传统 SVM 算法在应用于大规模训练样本和求解困难等方面的缺点。与传统 SVM 相比，LSSVM 主要是通过引入最小二乘线性系统到传统的 SVM 中，从而将原来的不等式约束变成等式约束，并且将解二次规划变为解一组等式方程，从而提高模型的求解速度。同时就传统 SVM 常用的 ε - 不敏感损失函数而言，LSSVM 则不再需要指定逼近精度 ε，这也使得 LSSVM 更容易理解和操作。

1.1.4 灰色系统分析

灰色系统（GS，Gray System）是邓聚龙在 1981 年提出的，是以信息不完全的系统为研究对象，运用特定的方法描述信息不完全的系统并进行预测、决策、控制的一种系统理论。它通过对“部分”已知信息的生成、开发，提取有价值的信息，实现对系统运行行为、演化规律的正确描述和有效监控。灰色系统是以“灰色朦胧集”为基础的理论体系，以灰色关联空间为