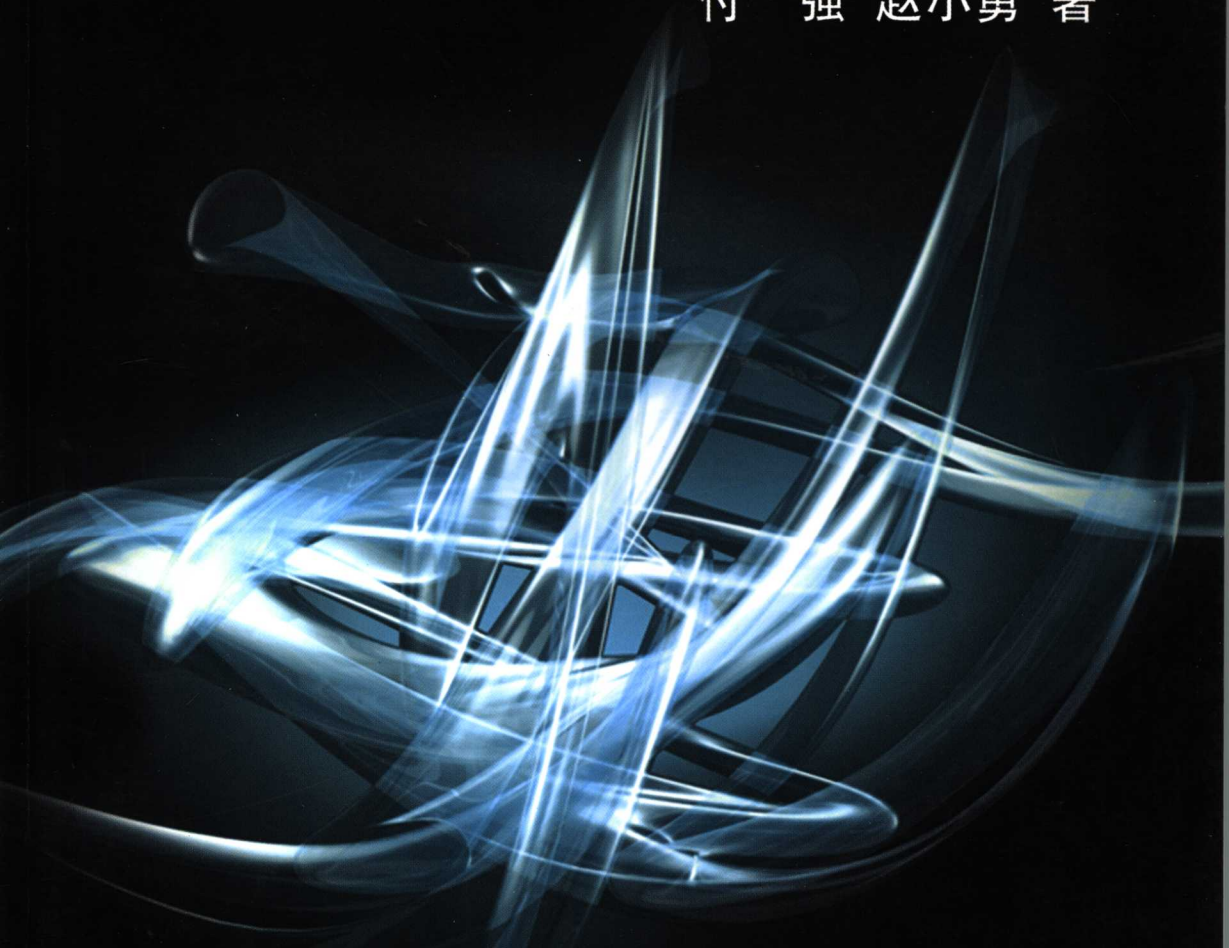
# 投影寻踪模型原理及其应用

付　强　赵小勇　著



科学出版社
www.sciencep.com

# 投影寻踪模型原理及其应用

付 强 赵小勇 著

科学出版社
北京

## 内 容 简 介

本书介绍了非线性复杂系统中数据处理的投影寻踪降维技术，给出了投影寻踪在综合评价和预测等方面的统计新模型，这些模型能充分提取数据信息，描述复杂系统的规律．书中深入浅出地介绍了各种投影寻踪模型方法的思想、原理和程序步骤，通过实例分析论证了投影寻踪模型稳健性好和准确度高等优点．

本书可供从事农业水土工程、环境工程、农业系统工程、水文学及水资源、农林经济管理等专业的科研、管理和工程技术人员阅读，也可作为相关专业的研究生参考教材．

**图书在版编目(CIP)数据**

投影寻踪模型原理及其应用/付强，赵小勇著．—北京：科学出版社，2006

ISBN 7-03-017094-6

Ⅰ.投… Ⅱ.①付… ②赵… Ⅲ.投影寻踪模型-研究 Ⅳ.O212

中国版本图书馆 CIP 数据核字(2006)第 028790 号

责任编辑：鄢德平 赵彦超／责任校对：包志虹

责任印制：安春生／封面设计：陈 敬

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

深海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2006 年 6 月第 一 版 开本：B5(720×1000)

2006 年 6 月第一次印刷 印张：12 1/2

印数：1—2 500 字数：243 000

**定价：32.00 元**

(如有印装质量问题，我社负责调换〈路通〉)

# 前　言

投影寻踪是由美国科学家 Kruskal 提出的一种用来分析和处理高维观测数据，尤其是非线性、非正态高维数据的新兴统计方法. 是统计学、应用数学和计算机技术的交叉学科，属当今前沿领域. 它通过把高维数据投影到低维子空间，寻找能反映原高维数据结构或特征的投影，达到研究分析高维数据的目的. 它具有稳健性好、抗干扰性强和准确度高等优点，可以在许多领域，诸如预测、模式识别、遥感分类、优化控制、导航、模拟雷达、图像处理、分类识别等领域广泛应用.

投影寻踪模型理论和方法处于发展阶段. 本书是近五年来作者将这一理论应用于农业系统及相关领域开展研究工作的一个总结，包括以下 8 章：第 1 章简要介绍了投影寻踪的由来、特点，分析了投影寻踪在各个研究领域中的应用现状；第 2 章介绍了基于遗传算法的投影寻踪模型；第 3 章介绍了投影寻踪数据特征分析，详细介绍了投影寻踪指标和投影寻踪的小波估计；第 4 章介绍了投影寻踪聚类模型及其应用；第 5 章介绍了投影寻踪主成分分析及其应用；第 6 章介绍了解不确定型决策问题的投影寻踪模型及其应用；第 7 章介绍了投影寻踪回归模型及其应用；第 8 章介绍了投影寻踪自回归模型及其应用.

本书的特点是内容新颖，理论联系实际，深入浅出，便于理解和实际分析计算. 本书可为高校从事数据处理的高年级本科生、研究生和教师提供帮助，同时适合于有关科技工作者参考.

本书参考和引用了国内外许多学者的相关论著，吸收了同行们的辛勤劳动成果，作者从中得到了很大的教益与启发，在此谨向各位学者表示衷心的感谢！

本书得到了国家自然科学基金(No.30400275)、国家“863”计划项目(No.2002AA2Z4251-09)、中国博士后科学基金(No.2004035167)、黑龙江省青年基金(No.QC04C28)、黑龙江省教育厅科研基金(No.10541033)、黑龙江省教育厅人文社科基金(No.1054xy006)、黑龙江省博士后科学基金(No.LSZH-04081)、北大荒集团公司博士后科研工作站博士后科研基金(No.LRB04-069)、东北农业大学博士后科学基金(No.240009)的联合资助.

付　强

2006 年 2 月 10 日

# 目　录

# 第1章 绪 论

人类、自然和社会体现了大自然的和谐统一. 在漫漫的历史长河中, 人类学会了认识和利用自然; 人类将认识事物的手段作了细致明确的分工, 形成了众多的学科, 建立了相应的理论体系和研究方法. 经过长期的研究与实践, 人们发现自然界的变化有着惊人的规律和秩序, 有着高度的组织性和系统性, 它像一个有机生命体, 内部的器官间有着丰富、有序的信息传递, 同时它与外部还有着信息交换和对外部信息的反映. 所有这些为人类认识世界和考察事物提供了可能的信息来源. 在如此宏大和复杂的信息集合中辨识事物现象与其本质、现象与现象之间的关系是一项十分有意义的研究. 物质、能源、信息是构成现代社会大厦的三大支柱. 物质是构筑社会的基础, 能源是构筑社会的动力, 而信息是构筑社会的神经系统, 信息的重要性已经被人们所认识, 信息理论也已经被广泛地应用到军事、医学、社会学、经济学、工业和农业等各个领域. 信息科学的最新发展表明, 建立在概率论基础上的Shannon 信息论, 只着重表达了信息的传递, 难以表达数据信息本身的含义. 而信息科学不仅要研究数据信息 "量" 的问题, 更重要的还在于数据的信息特征及信息的定性问题. 这就涉及数据信息的提取、描述、推理、判断和决策等富有挑战性的处理工作.

农业系统是国民经济大系统的重要组成部分, 农业系统内部结构错综复杂, 同时, 农业系统随时间演变的过程受到众多因素影响, 这些因素之间存在着复杂的关系, 随着人类活动对农业系统影响的加深, 使得因素之间的关系更加复杂, 因此对研究方法也提出了更高的要求, 要求数理统计方法能够充分描述系统中各个因素之间的相互作用关系, 比较全面地揭示农业系统演化规律.

在农业系统研究中, 不确定分析方法占有重要位置. 根据农业系统的特点以及试验数据资料, 运用不确定分析方法, 例如随机、模糊、灰色、人工神经网络、混沌等, 并结合农业系统的特点, 就可以建立农业系统的不确定分析模型和研究系统的变化规律. 建立模型时, 每个独立的因素就是一个独立参数, 因此有几个独立影响因素, 其参数空间就是几维, 研究对象的参数空间视运动复杂程度而定, 可以是高维的. 研究对象的参数空间常常超过形体或空间界限, 描述的是信息架起来的数理模型复杂度, 如果系统受到 $n$ 个独立因素的影响和制约, 就有理由认为此系统处于 $n$ 维空间. 当独立影响因子增加时, 所张开的空间维数随之增加, 要建立效果良好的农业系统模型, 就要求有足够的资料来估计模型参数. 由于农业系统资料十分有限, 因此资料长度与估计精度之间的矛盾更加突出, 在统计学中称之为高维问题, 高维

问题降低了参数估计的稳健性. 为此, 在建立农业系统多因子模型时, 需要引进新的、可靠的方法解决上述问题.

在近代统计学中, 出现了一种解决高维问题的统计方法 —— 投影寻踪, 其出发点是将高维问题引入低维空间后再进行研究. 在农业生产系统研究中, 用这种方法可以建立多因子预测、评价模型, 解决资料长度与预测精度之间的矛盾. 本书是在前人研究的基础上, 解决投影寻踪方法在应用时出现的关键问题, 解决农业系统领域研究中关于预测、多维评价以及多元复杂性问题, 使投影寻踪高维降维理论和技术得以发展和完善, 并与农业系统相结合, 解决农业系统中先前诸多悬而未解的实际问题, 使该方法在理论与实际应用问题上迈上新台阶, 为解决农业系统复杂性问题开辟新的研究途径与模式.

## 1.1　投影寻踪简介

### 一、投影寻踪的产生背景

随着人们对事物复杂性认识不断深入, 加之计算机技术日新月异的发展, 使得高维数据的统计分析越来越重要. 在许多实际问题中数据的维数相当高, 因为事物在其演变过程中必然会受到众多因素的影响和制约, 为了避免忽略掉任何可能的相关信息, 往往在搜集资料时要全面考虑各个因素, 从而使多元分析方法的应用不但非常普遍而且很重要. 传统的多元分析方法是建立在总体服从正态分布的基础上的, 而实际中有许多数据不满足正态假定, 需要用稳健的、实用的方法来解决. 遗憾的是当数据的维数较高时, 这些方法将面临一些困难, 主要困难有三: 一是随维数的增加, 计算量迅速增大, 而且不可能将其画出可视的分布图或其他图形. 二是当维数较高时, 即使数据的样本点很多, 散在高维空间中仍显得非常稀疏. 例如, 设有一个容量很大的高维点云均匀分布于 10 维单位球内, 则含有点云 5%点的小球体半径约占原单位球体半径的 74%; 如果该小球体的半径只占原单位球体半径的 5%, 则该小球体只含有 $(0.05)^{10} \approx 0$ 个资料点, 几乎是个空球. 1961 年, Bellman[1] 将这种现象称为 "维数祸根". 高维点云的稀疏性使许多在一维情况下比较成功的传统方法, 如关于密度函数估计的核估计法、邻域法等不能适用. 因而在研究高维数据时, 希望找到降维的方法, 如聚类分析、因子分析、典型相关分析等等, 但这些方法仅着眼于变量间的距离, 而忽略了不相干变量的存在, 使人无法确定结果的正确性. 三是在低维时稳健性很好的统计方法到了高维空间, 其稳健性就变差了. 以上情况表明, 传统的数据分析方法对于高维非正态、非线性数据分析很难收到很好的效果. 其原因在于它过于形式化、数学化, 难以适应千变万化的客观世界, 无法找到数据的内在规律和特征, 远不能满足高维非正态分布数据分析的需要. 投影追踪方法就是在这种形势下应运而生的.

投影寻踪产生的另一背景是由于一种非常直观的、无太多深刻道理的思维方式的存在. 对于一维和二维的数据结构, 常常采用直方图来了解数据的特征, 并通过观测这些图形的变化趋势来判断已有或未知数据的结构. 例如, 非正态的密度图、计算设计洪峰的皮尔逊Ⅲ型分布图就可以在平面上直接绘出. 虽然这种观察方式非常粗糙, 但也能为进一步研究提供启示. 当数据维数大于 4 时, 无法用眼直接观察数据结构, 需要将原始数据投影到可以观察到的空间维上, 即 1~3 维, 通过在低维空间的观测来看数据在高维空间的结构. 在科学研究中, 常常也有类似做法, 比如在研究多个变量与一个变量的关系时, 可以先挑选其中几个变量与一个变量来研究, 再挑选另外的变量逐一研究, 只是投影的思想更直接一些.

投影寻踪是用来处理和分析高维数据, 尤其是来自非正态总体的一类统计方法, 既可作探索性分析, 又可作确定性分析, 其基本思想是把高维数据投影到低维子空间上, 寻找出能反映高维数据结构或特征的投影, 以达到研究分析高维数据的目的. 投影寻踪方法的特点, 主要可以归纳为以下几点:

(1) 自然科学中有许多数据不符合正态分布或人们对数据没有多少先验信息, 需要从数据本身找出其结构或特征. PP 方法能成功地克服高维数据的 "维数祸根" 所带来的严重困难, 这是因为它对数据的分析是在低维子空间上进行的, 对 1~3 维的投影空间来说数据点就足够密了, 足以发现数据在投影空间中的结构或特征;

(2) 投影寻踪方法可以排除与数据结构和特征无关的, 或关系很小的变量的干扰;

(3) 投影寻踪方法为使用一维统计方法解决高维问题开辟了途径. 因为投影寻踪方法可以将高维数据投影到一维子空间上, 再对投影后的一维数据进行分析, 比较不同一维投影的分析结果, 找出好的投影;

(4) 投影寻踪方法与其他非参数方法一样可以用来解决某种非线性问题. PP 问题虽然是以数据的线性投影为基础, 但它找的是线性投影中的非线性结构, 因此它可以用来解决一定程度的非线性问题, 如多元非线性回归.

投影寻踪方法的关键在于找到观察数据结构的角度, 即数学意义上的线、平面维或整体维空间, 将所有数据向这个空间维投影, 得到完全由原始数据构成的低维特征量, 反映原始数据的结构特征.

## 二、 投影寻踪研究的主要内容

投影寻踪方法最早出现在 20 世纪 60 年代末 70 年代初. 为了发现数据的聚类结构, 1969 年 [2] 和 1972 年 [3], Kruscal 首先使用投影寻踪方法, 把高维数据投影到低维空间, 通过数值计算, 极大化一个反映数据聚类程度的指标, 从而找到反映数据结构特征的最优投影. 1970 年, Switzer 等人 [4] 也通过高维数据的投影和数值计算解决了化石分类问题. 1974 年, Friedman 和 Tukey[5] 用数据的一维散布和局部密度

的积构造了一类新投影指标, 用来进行一维或二维情形下的聚类和分类, 并利用这个新指标成功分析了计算机模拟的均匀分布随机数的散布结构、单纯形顶点上的高斯分布以及有名的鸢尾花聚类问题, 并将此方法命名为投影寻踪. 他们还领导编制了一个用来寻找数据聚类、散布的超曲面结构的计算机图像系统 PRIM-9[6].

之后, 关于投影寻踪方法的一系列研究成果在理论与应用研究领域引起很大重视. 在 1979 年美国数理统计学会年会上, 数据分析专题组织者 P.J. Huber 邀请 Friedman 作了关于投影寻踪的报告, 成为投影寻踪理论研究的引子, 随后相继派生出投影寻踪回归 [7,8]、投影寻踪聚类 [9]、投影寻踪密度估计 [10] 等等方法. 1981 年, Donoho 提出了用 Shannon 熵来定义一个投影指标 [11]. 1981 年, 李国英和陈忠链 [12] 等人用投影寻踪方法给出了散布阵和主成分的一类稳健估计, 并讨论了其统计特性, 另外, 许多统计学工作者也讨论了关于投影寻踪的问题 [13~20].

1985 年, 应 *The Annals of Statistics* 杂志的邀稿, Huber[21] 发表了关于投影寻踪的综合性学术论文, 并附有从事这一研究的理论工作者的讨论文章. 至此, 初步建立了投影寻踪在统计学中的独立体系, 大大推动了此方法的深入研究和实际应用.

从投影寻踪的理论与应用研究来看, 主要涉及三方面内容, 包括投影寻踪聚类分析、投影寻踪回归以及投影寻踪学习网络.

1. 投影寻踪聚类分析

1936 年, Fisher[22] 在研究鸢尾花数据的判别问题时, 开创了线性判别分析思路, 其实质是一种投影寻踪算法. 1970 年, Switzer[4] 对牙买加化石数据进行分类时, 引入了 Fisher 的上述思想, 提出投影寻踪聚类设想. 1974 年, Friedman 和 Tukey[5] 明确提出了投影寻踪思想: 将数据集投影到低维子空间上, 对投影得到的低维构形, 通过定义好的投影指标, 用计算机寻求使投影指标达到极大的一个 (或几个) 投影方向 (或平面), 给出直线 (或平面) 上的数据投影, 由计算机图像系统显示出来, 然后用眼直接判断数据结构. 以上一系列有代表性的研究为拓宽投影寻踪在实践中应用提供了基本思路.

之后, 投影寻踪聚类方法被广泛应用于模式识别领域, 其基本思路是利用投影寻踪压缩和提取系统的高维特征量后, 再对系统模式进行识别.

文献 [23] 的研究证明, 利用投影寻踪技术压缩高维特征的空间维数, 更有利于识别高维系统模式. 文中还构造了一个便于实现的投影指标, 同时给出了寻找投影方向的新途径.

文献 [24] 将投影寻踪技术用于遥感领域, 给出了识别卫星云图的新的投影指标.

文献 [25] 采用投影寻踪的思想构造稳健协差阵, 建立了一种新的能抗异常值干扰的稳健判别方法, 新方法的计算结果不易受异常值干扰.

文献 [26] 提出了基于核的投影寻踪方法, 并将其应用到滚动轴承的质量分类中, 取得了较为理想的效果.

以上研究表明, 投影寻踪聚类方法为多元数据分析方法的实践提供一种新思路, 取得了优于传统方法的良好效果.

2. 投影寻踪回归

Friedman 等很早便意识到投影寻踪方法所显示出来的处理高维数据的优势, 因此将投影寻踪方法引入多元回归分析, 建立了一种广义多元回归分析方法, 在一定程度上克服了维数祸根的问题, 取得了相当满意的预测效果.

杨力行等 [27,28] 在前人研究工作的基础上, 根据投影寻踪回归思想研制了投影寻踪回归分析软件包, 在预测 [29]、优化 [30] 等领域取得了丰富成果.

史久恩 [31] 将投影寻踪方法用于气象研究, 指出这是一条新的、有用的途径.

李祚泳等 [32~36] 将投影寻踪回归方法成功用于环境预测以及环境影响因子的污染作用分析等方面.

虽然投影寻踪方法应用还不广泛, 但从目前的应用结果表明, 投影寻踪方法起点较高, 思路新颖, 较之常规多元分析方法的确表现出一定优势, 可以解决参数估计时的高维问题.

3. 投影寻踪学习网络

从国内情况来看, 对投影寻踪方法的应用研究是较薄弱的. 在国外, 自投影寻踪方法出现以来, 引起了许多领域学者的重视, 包括应用统计和神经网络研究方面的学者. 在 Barron[37] 倡导的统计学习网络思想影响下, 许多研究神经网络的学者将投影寻踪回归思想引入网络学习中, 改变了前馈型神经网络中常用的 BP 算法以及神经元函数形式, 提出了基于投影寻踪回归学习策略的投影寻踪学习网络 (Projection Pursuit Learning Network, 即 PPLN), 其实质是一种更广泛意义上的网络回归模型.

Maechler 和 Mertin[38] 对比研究了人工神经网络 (ANN) 和非参数 PPLN 的学习策略和网络结构, 分别用这两种模型模拟了五种不同类型的二维函数, 模拟结果表明, 在同一精度下, PPLN 的训练速度比 ANN 快几十倍; 在训练的精度方面, 就平均水平而言, ANN 稍优于 PPLN, 主要原因是建立模型的样本个数有利于 ANN 的参数估计, 而不能满足 PPLN 的非参数估计. 通过对比研究, 作者明确指出了 PPLN 的学习策略优于 ANN.

由于非参数估计方法尚不完全成熟, 且应用时有诸多不便, 虽然其使用面广, 但在解决一些很复杂问题时具有一定局限性, 因此以参数神经元函数为主的 PPLN 模型依然是主要发展方向.

颜光宇 [39] 针对传统因子分析方法易受异常值干扰的缺陷, 采用稳健 M 估计和投影寻踪方法求解稳健相关阵, 提出了一种新的可抗异常值干扰的稳健因子分析

方法, 应用表明, 当数据中含有少量异常值时, 此方法可抗异常值干扰, 优于传统因子分析方法. 国外学者还提出投影寻踪与模糊神经网络耦合的模型 [40,41], 对投影寻踪方法及其应用的未来发展趋势进行了讨论 [42~46].

投影寻踪方法的研究进展表明了此方法的应用价值, 能适应形式灵活的网络发展要求, 对于不同研究对象采用各种形式的模型进行研究, 是探索复杂系统规律的有效方法之一.

## 1.2　投影寻踪模型研究进展

投影寻踪是一种新兴的、有价值的高新技术, 是统计学、应用数学和计算机技术的交叉学科, 属当今前沿领域. 它是用来分析和处理高维观测数据, 尤其是非线性、非正态高维数据的一种新兴统计方法. 它通过把高维数据投影到低维子空间, 寻找能反映原高维数据结构或特征的投影, 达到研究分析高维数据的目的. 它具有稳健性好、抗干扰性强和准确度高等优点, 可以在许多领域, 诸如预测、模式识别、遥感分类、过程优化控制、导航、模拟雷达、图像处理、分类识别等领域广泛应用. 目前, 投影寻踪模型已在工业、农业、水利、医学及遥感等领域得到广泛应用并相继取得了一批可喜成果.

1990 年, 文献 [47] 利用投影寻踪技术帮助海军沿着一条有利的路线到达目标点. 即使由于位置测量存在误差, 投影寻踪方法仍能排除干扰, 给出稳定的方向解.

1991 年, 文献 [48] 给出了一种参数 PPLN 形式, 成为参数 PPLN 模型中的代表. 作者研究和对比了两种解决回归问题的模型, 即人工神经网络中含一个隐层的 BP 网络和以统计学为基础的投影寻踪学习网络. 从比较的结果可以看出, 反向传播学习策略与投影寻踪学习策略存在明显差异. 用一个隐层的 BP 网络和投影寻踪学习网络分别对五种类型函数的逼近效果来看, 投影寻踪学习策略更优, 取得的逼近效果更好. 作者从模型精度、吝啬程度 (指使用神经元个数少、隐层数少) 和学习速度三方面进行了细致比较, 发现基于非参数 (超级平滑 (super smooth)) 的投影寻踪学习网络的学习精度优于 BP 学习网络, 而参数 (基于 Hermite 多项式) 投影寻踪学习网络优于非参数投影寻踪回归模型; 相同精度下, 参数投影寻踪学习网络要求的神经元个数少于 BP 网络和非参数投影寻踪回归模型; 在所有模拟试验中, BP 网络与投影寻踪学习网络都可以达到在 100 次循环后收敛, 两种模型具有相当的收敛速度. 总的来看参数投影寻踪学习网络优于 BP 学习网络, 并在多个方面较非参数投影寻踪回归显示出优势.

文献 [49] 用投影寻踪学习模型学习机器人手臂的反向动力变化规律, 证明了投影寻踪回归的分组学习策略在应用时的有效性, 认为参数投影寻踪回归较非参数投影寻踪回归具有较高的精度和收敛速度, 而且参数投影寻踪用较少的参数可以取得

较一个隐层的 S 型神经网络模型更高的精度, 并给出了含一个隐层的神经网络模型的参数个数计算式：$N \approx pd$ ($p$ 为隐含节点数, $d$ 为输入空间的维数). 投影寻踪模型的参数经过分组后, 其神经元个数的计算式为：$N' \approx \frac{p}{s}d$($s$ 为在每个投影方向上平行的超平面个数). 可以看出, 投影寻踪学习要求的参数数目实际上少于一个隐层神经网络的参数个数.

1994 年, 文献 [50] 将投影寻踪技术用于大气颗粒源解析分析. 由于观测的资料是一些高维数据序列, 用投影寻踪方法投影后, 选出其中极有效的几维, 去捕捉数据的主要特征, 并借助于风向资料判定大气颗粒的来源. 文献 [51] 用投影寻踪技术识别模拟雷达信号, 并解决了时间相依的分类问题.

1997 年, 文献 [52] 用投影寻踪技术压缩可观测到的图像信息, 进而识别其余未能观测到的系统灰信息. 文献 [53] 将投影寻踪回归分析方法用于导弹目标追踪问题的研究, 由于高维特征量压缩与提取是声纳目标信号分类首先要解决的关键问题, 文中基于投影寻踪理论提出了采用投影寻踪压缩与提取, 进而分类的理论和方法. 将此方法用于实测数据, 结果表明其是降低特征空间维数, 正确进行分类的行之有效方法.

1998 年, 文献 [54] 用投影寻踪学习为高维小样本序列设计了一个神经网络, 将投影寻踪的思想与 Slicing Inverse Regression(SIR) 统计思想联合, 建立了快速投影寻踪学习模型. 将其用于短期负荷的电力预测, 取得了满意成果, 证明投影寻踪学习对解决小样本问题有许多优势.

1999 年, 文献 [55] 将 PP 技术用于对遥感影像信息的判读, 在森林类型等方面, 大幅度提高了利用卫星遥感影像分类的精度, 提高了遥感信息的利用率, 为遥感应用创立了具有突破技术瓶颈意义的崭新技术手段和开发利用遥感信息资源的捷径. 文献 [56] 应用投影寻踪回归技术, 建立了流域年均含沙量的预测模型, 用降雨量和年平均径流等 4 个因子建立的某流域平均含沙量的 PPR 预测结果的拟合合格率达 100%, 预留检验样本报准率为 75%, 表明 PPR 用于泥沙输移规律的预测研究是可行的.

2000 年, 文献 [57] 针对现有紫坪埔洪水预报模型或不能充分挖掘样本信息或不便于实际应用的问题, 用投影寻踪回归方法建立了紫坪埔洪水预报模型, 分别对洪峰和洪水过程进行了预报, 并与其他方法进行对比, 取得了满意的效果, 可以作为紫坪埔洪水预报的新方案. 文献 [58] 针对我国金融数据分布的非正态性和高维性特点, 提出了一种新型模型 —— 投影寻踪判别分析模型, 研究我国商业银行的信用风险评估问题. 实证结果表明, 与传统的判别分析方法和近邻法相比, 投影寻踪判别分析模型在处理具有非正态、高维性的信用风险评估问题时, 精度更好. 文献 [59] 建立了应用于大型船舶运动的极短期预报的多维投影寻踪学习网络结构及算法, 并将该算法所取得的预报结果与自回归预报法和周期图预报法的结果进行比较, 预报

结果说明了该算法的可行性.

2001 年, 文献 [60] 将投影寻踪回归建模技术用于悬板过流区自由水面的模拟和仿真, 并与边界元法的计算结果进行了对比分析, 得到了比较满意的结果. 该技术具有计算快速简便, 无需求解复杂的微分方程等特点, 是用统计方法解决复杂工程水力学问题的有益尝试. 文献 [61] 研究非线性自回归模型投影寻踪学习网络逼近的收敛性, 证明了在 $L^k(k$ 为正整数) 空间上, 投影寻踪学习网络可以以任意精度逼近非线性自回归模型, 给出基于投影寻踪学习网络的非线性时间序列模型建模与预报的计算方法和应用实例, 对太阳黑子数据、山猫数据及西安数据进行了拟合和预报, 将其结果与改进的 BP 网络和门限自回归模型相应的结果进行比较, 结果表明基于投影寻踪学习网络的非线性时间序列的建模和预报方法是一类行之有效的方法. 文献 [62] 针对水稻节水效益评价问题, 采用高维降维技术 —— 投影寻踪分类模型, 利用基于实数编码的加速遗传算法优化其投影方向, 将多维数据指标 (样本评价指标) 转换到低维子空间, 根据投影函数值的大小评价出样本的优势, 从而作出决策, 最大限度避免了模糊综合评判等方法中权重矩阵取值的人为干扰, 取得了满意效果, 为节水效益评价及其他评判决策问题提供一条新的方法与思路.

2002 年, 文献 [63] 用投影寻踪的方法搜寻理想的投影方向, 以便使高维数据降维而发现数据中化合物的分类信息, 并利用这样的分类信息对样本进行分类建模, 取得了理想的结果. 文献 [64] 为预测年径流这类高维复杂动力系统, 提出了投影寻踪门限回归模型 (PPTR), 构造了新的投影指标函数, 用门限回归模型描述投影值与预测对象间的非线性关系, 并用实码加速遗传算法优化投影指标函数和门限回归模型参数, 实例的计算结果表明, 用 PPTR 模型预测年径流是可行而有效的.

2003 年, 文献 [65] 应用投影寻踪回归技术, 对非正态、非线性悬栅消能率实验数据, 用 1/5 数据建模拟合, 4/5 数据留作预留检验, 拟合合格率 92%, 预留检验合格率 92%, 并与激光测速得出的消能率及原型观测的消能率完全吻合. 文献 [66] 分析了基于信息散度指标投影的寻踪方法在高光谱图像处理中的应用, 给出了它与主成分分析处理结果的对比, 并提出 PP 与高光谱研究将来的发展方向. 文献 [67] 针对现有模糊图像的复原方法, 提出了一类新型人工神经网络 —— 投影寻踪子波学习网络, 并将其用来处理图像的去模糊问题. 这类新型网络具有投影寻踪学习网络优点, 在先验条件知道甚少的情况下, 不用求点扩展函数, 直接通过网络的学习, 提取参数, 以达到自适应剔除图像的模糊信息, 恢复原图像, 且具有小波函数的时域局部性, 可以对多种噪声源的模糊图像进行恢复. 模拟结果表明, 该方法对于图像的无监督恢复明显优于现有的图像恢复方法.

2004 年, 文献 [68] 针对动态多指标决策中指标和时段的权重确定问题, 提出了基于投影寻踪的理想点法新模型 (动态多指标决策问题的投影寻踪模型). 该模型利用决策矩阵样本的内部信息, 把方案的三维决策矩阵综合成一维投影值, 投影值越

大表示该方案越优, 根据投影值的大小就可对各方案进行综合排序决策. 文献 [69] 研究采用大样本数据, 利用投影寻踪、遗传算法、插值型曲线和水质评价标准, 为水质综合评价建立了一种新的数学模型 —— 遗传投影寻踪插值模型, 实例研究表明, 遗传投影寻踪插值模型建模方法直观、可靠、精度高, 既具有较强的分类功能, 又具有较好的排序功能, 可广泛应用于各种环境质量的综合评价.

2005 年, 文献 [70] 提出房地产投资多目标决策模型, 结合指标及数据分布特点将投影寻踪方法应用到房地产风险评价中, 采用基于实数编码的加速遗传算法来简化 PP 模型建模过程, 该方法直接面向数据建模, 将多种指标进行线性投影, 为决策者提供了一个综合全部指标信息的决策依据, 且具有简便、通用、准确等优点.

## 1.3 本书的主要内容

投影寻踪模型理论和方法处于发展阶段. 从已有的自然科学各领域应用看, 投影寻踪模型具有很大的发展潜力. 本书是作者近五年来应用于农业系统及相关领域中所开展研究工作的一个总结, 包括以下 8 章: 第 1 章简要介绍了投影寻踪的由来、特点, 分析了投影寻踪在各个研究领域中的应用现状; 第 2 章介绍了基于遗传算法的投影寻踪模型, 详细介绍了遗传算法的原理、特点和研究动态, 并给出了两种改进的遗传算法, 两种改进的遗传算法分别为基于实数编码的加速遗传算法和基于实数编码的加速免疫遗传算法, 详细介绍了投影寻踪的基本概念和遗传算法优化投影寻踪模型的投影方向; 第 3 章介绍了投影寻踪数据特征分析, 详细介绍了投影寻踪指标和投影寻踪的小波估计, 投影寻踪指标分为密度型投影指标和非密度函数型投影指标, 并指出不管是采用哪种投影寻踪指标, 其实质都是度量一个分布与其同方差的正态分布间的距离, 因此偏离正态分布的程度是投影寻踪指标的重要特性, 并给出了偏离正态分布的程度的计算公式; 第 4 章介绍了投影寻踪聚类模型及其应用, 详细介绍了投影寻踪聚类模型降维思路、研究内容和应用归类, 投影寻踪聚类模型用于多因素影响问题的综合评价, 但根据具体的分析问题的特点, 目前可将其主要应用归纳为两个大的方面: 投影寻踪分类模型和投影寻踪等级评价模型, 详细介绍了投影寻踪分类模型原理及其在水稻节水效益评价、水稻灌溉制度优化、节水灌溉项目投资决策、农业生产力综合评价、工程评标、农机选型及优序关系研究、农村能源区划、水资源工程方案优选、生态农业综合评价、小流域效益分类评价、水资源承载能力评价和物流规划分类评价中的应用, 投影寻踪等级评价模型原理及其在土壤质量变化评价、水质评价、土壤养分等级评价、区域水资源可持续利用评价、湖泊水污染综合评价等中的应用, 基于逻辑斯谛曲线的投影寻踪等级评价模型原理及其在耕地资源可持续利用综合评价中的应用, 基于倒 S 型曲线的投影寻踪等级评价模型原理及其在土壤质量变化综合评价、农业水资源供需状况评价中

的应用, 投影寻踪插值模型原理及其在创业农场水资源可持续利用评价中的应用; 第 5 章介绍了投影寻踪主成分分析及其应用, 详细介绍了投影寻踪主成分分析在水利项目评价、灌溉模式优选中的应用; 第 6 章介绍了解不确定型决策问题的投影寻踪模型及其应用, 详细介绍了解不确定型决策问题的投影寻踪模型原理及其在几种产品生产分析中的应用; 第 7 章介绍了投影寻踪回归模型及其应用, 详细介绍了投影寻踪回归模型原理及其在酸雨 pH 预测、酒埠江水库流量预测、地下水埋深模拟中的应用, 投影寻踪门限回归模型原理及其在降水量预测中的应用, 基于神经网络的投影寻踪耦合模型原理及其在年径流预测、水文相关分析、描述作物 — 水模型中的应用, 基于偏最小二乘回归的投影寻踪耦合模型原理及其在水稻腾发量预测中的应用, 基于偏最小二乘回归的神经网络投影寻踪耦合模型原理及其在城市水资源承载力预测中的应用; 第 8 章介绍了投影寻踪自回归模型及其应用, 详细介绍了投影寻踪自回归模型原理及其在降水量预测中的应用, 投影寻踪门限自回归模型原理及其在海洋冰情等级预测中的应用, 基于神经网络的投影寻踪自回归耦合模型原理及其在水稻单产预测中的应用.

# 第 2 章　基于遗传算法的投影寻踪模型

## 2.1　概　　述

利用投影寻踪方法解决实际问题的关键是构造能够找到最佳投影方向的有效算法.

1969 年, Kruscal[2] 提出借助计算机扩展眼功能的投影寻踪思想, 这种方法是将散布于高维空间的点云投影到低维子空间 (人眼可以观测的空间), 优化某一投影指标, 找到若干个投影方向, 使得低维空间点的散布结构最能反映高维点云的散布特征, 通过研究高维数据在低维空间的散布结构, 从而找到高维数据的特征. 寻找最佳投影方向的手段是通过人眼对连续方向的观察, 并没有给出能用计算机直接确定最佳投影方向的有效算法.

1974 年, Friedman[5] 根据 Kruscal 的思想给出了多元数据分析的投影寻踪算法, 此算法的主要目的是寻找一两个揭示多元数据特征的线性投影. Friedman 运用固定角旋转 (Solid Angle Transport, 即 SAT) 技术, 在初始方向的附近区域搜索最优的投影方向, 从任意一个初始点开始, 变动一个微小的固定角, 当投影质量改善时, 就沿这个方向继续搜索, 否则就取相反的方向, 对于投影方向对应的每一维向量都必须进行相应 SAT 运算. 当空间维数增加后, 数据结构变得复杂, 可以从若干个不同的方向, 多次进行 SAT 运算, 搜索最优的投影方向. 算法的实际应用表明, 它解决了投影寻踪的两个基本问题: 一是在低维空间中寻找更能揭示高维数据结构的投影; 二是由多元数据在低维空间的散布以及局部密度两个测度的乘积构造的投影指标, 作为优化投影方向时的目标函数.

以后的投影寻踪算法寻优, 主要是基于上述旋转变换的思想. 可以用各种方式的优化计算方法来实现, 例如梯度下降法 [45]、高斯 — 牛顿法 [7] 等等.

当研究对象复杂时, 多元数据具有复杂的拓扑结构, 以上算法存在的问题是: 如何从成千上万个区域内选取若干个采样点作为初始方向. 初始方向选取不妥, 收敛到最优解的时间就长, 有时甚至很难找到最优解. 即使找到某些方向, 那么旋转角度的大小直接影响了算法的寻优效率, 角度愈小, 计算耗时愈大, 角度过大, 可能会失去某些最优解. 针对传统的优化方法处理多变量同时寻优时往往易陷入局部最优、早熟或提前收敛, 寻求不到真正的最优解的问题, 本书引入一种全局优化算法 —— 遗传算法 (Genetic Algorithm, 简称 GA), 结合由目标函数反映的高维数据结构特性, 在优化区域内直接寻找最优解, 给出一种确定投影方向的新途径.