改进的高维非线性偏最小二乘回归模型及应用

Improved High-Dimension and Nonlinear Partial Least-Squares Regression Model and Applications

郭建校 著

中国物资出版社

天津外国语大学国际商学院精品专著系列
天津外国语大学首届“青蓝之星”基金资助项目
天津外国语大学《统计学》精品课程系列著作

改进的高维非线性偏最小二乘回归模型及应用

郭建校 著

中国物资出版社

图书在版编目（CIP）数据

改进的高维非线性偏最小二乘回归模型及应用/郭建校著．—北京：中国物资出版社，2010.9

（天津外国语大学国际商学院精品专著系列）

ISBN 978-7-5047-3556-0

Ⅰ.①改… Ⅱ.①郭… Ⅲ.①非线性回归 Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字（2010）第 181766 号

策划编辑 寇俊玲

责任编辑 左卫霞

责任印制 何崇杭

责任校对 孙会香 杨小静

中国物资出版社出版发行

网址：http://www.clph.cn

社址：北京市西城区月坛北街 25 号

电话：（010）68589540 邮政编码：100834

全国新华书店经销

中国农业出版社印刷厂印刷

开本：710mm×1000mm 1/16 印张：9.75 字数：160 千字

2010 年 9 月第 1 版 2010 年 9 月第 1 次印刷

书号：ISBN 978-7-5047-3556-0/O·0041

定价：26.00 元

（图书出现印装质量问题，本社负责调换）

前　言

偏最小二乘（Partial Least-Squares，PLS）回归是一种基于高维投影思想的新的非参数回归方法，可有效地将多元回归、主成分分析以及典型相关分析等功能有机地结合起来，因此，它已被誉为第二代多元统计分析方法。识别特异点和对变量集实施降维是回归建模前两个重要的数据分析预处理过程。本书基于偏最小二乘回归模型，结合非线性核主成分分析、二叉树知识等多种方法，提出了改进的非线性偏最小二乘回归模型、二叉树降维方法和降维二叉树评价方法，并扩展了特异点识别方法。主要论述内容如下。

提出了一种改进的非线性偏最小二乘回归模型。传统的线性及非线性PLS回归模型计算因变量集与提取的主成分之间的线性回归，没有考虑因变量集和主成分之间可能是非线性关系。本书把因变量集对各个主成分的线性回归改进为可根据具体情况选择线性回归或非线性回归，每个主成分依旧表示成原始自变量集的线性回归方程。本书还具体分析并建立了汽车油耗与其他十个设计及性能方面的指标之间的非线性回归模型。

提出了高维空间的二叉树降维方法及降维二叉树评价方法。本书提出了将传统的整体降维，改进为从局部降维再延伸到全局降维的一种逐步降维的新方法。如果样本变量数 n 过大，可对相关性最强的两个变量实施主成分分析或核主成分分析：提取第一个成分变量代替原来的两个变量，样本变量数则降维为 $n-1$，循环执行此降维过程，直到满足精度为止。整个降维过程表现为一棵二叉树或残缺二叉树。根据降维二叉树评价方法，采用天津市2008年各区县经济发展指标，具体对天津市18个区县的经济发展水平进行了科学的评价。

分析并扩展了高维空间的特异点识别方法。在基于PLS回归识别特异点的分析技术基础上，将识别特异点的二维平面 T^2 椭圆图方法扩展到三维空间 T^2 椭球和高维空间 T^2 超椭球，同时基于谱系聚类法，提出了基于高维空间主成分谱系图的特异点识别方法，并对我国主要省份、城市的汽柴油价格进

行了分析。

不能付诸实践的理论是空洞的理论。本书内容编写力求实用，所有模型和方法皆有相应的应用案例分析，语言通俗易懂。在保证内容完整性的前提下，尽量简化理论内容以及一些数学公式的推导和演算，侧重方法介绍及案例分析。

本书所有应用案例均采用R语言进行了实证，R程序或命令均以R 2.9.1版本为基础编写并运行，而且全部程序均运行通过。读者如果需要作者自编的R程序或命令，可以通过电子邮件向作者索取。

学无止境，研亦无止境，书中介绍的内容还可继续扩展，例如，可结合非线性核函数，消除偏最小二乘回归计算方法中所有的线性化计算步骤，以实现完全彻底的非线性偏最小二乘回归模型；把二叉树降维方法扩充为大局部降维方法，直至实现可调局部降维方法；继续丰富基于主成分谱系图的特异点识别方法，实现定性与定量方法相结合的综合识别方法等。

本书介绍的模型、理论和方法等可用于评价、预测、数据预处理、数据挖掘等领域，可为从事数据分析的科研人员提供计算模型和方法，也可以作为相关领域科研人员的参考书。

由于作者水平有限，且受时间和资料的限制，书中不妥之处在所难免，欢迎读者不吝指正，作者的电子邮箱地址：cpropro@163.com。

郭建校

2010年8月

目　录

1 绪 论

在实际应用问题中，尤其是在经济管理与工程技术的预测和控制研究中，经常需要分析研究两组多重相关变量之间的相互依赖关系，即回归关系。统计回归的研究具有悠久的历史，在理论和实践方面都取得了丰硕的成果，在社会科学和自然科学等许多领域都得到了广泛应用。在众多的统计回归方法中，偏最小二乘回归综合了多元统计分析中典型相关分析、主成分分析和多元线性回归等多种方法的优点，是一种新型多元数据分析方法。近些年，有关偏最小二乘回归模型的理论和方法的研究逐渐成熟，该回归方法在很多领域得到了广泛应用与推广。

在建立回归模型前，需要对数据进行各种预处理工作，如识别并剔除特异点，对高维数据进行降维处理，以简化计算过程和结果。改进的非线性偏最小二乘回归模型、二叉树降维方法和特异点识别方法是本书论述的三个主要内容。其中，改进的非线性偏最小二乘回归模型是本书的核心内容，二叉树降维方法及降维二叉树评价方法是本书首次提出的局部降维与评价方法，也是本书的主要创新内容。本章首先简述了三种方法的基本内容，然后介绍了三种方法的国内外研究现状与进展、本书研究内容采用的基本思路与研究方法，最后介绍了本书的主要内容和创新之处。

1.1 引 言

1.1.1 数据预处理与回归建模

如果研究对象的内在特征和各因素间的关系比较复杂，无法分析并确定因素关系进而建立数学模型时，常用的办法是借助于搜集到的大量统计数据，基于对数据的统计分析建立回归模型。回归分析是处理变量之间关

系的一种常用统计方法，研究用一组变量（即自变量集或预测变量集）去预测另一组变量（即因变量集或响应变量集）时，最小二乘回归、一元线性回归方程（成对变量之间）等是很常见的方法。现已出现的基于最小二乘准则下的多元线性回归分析（MLR）、基于自变量集主成分的主成分回归分析（PCR）等方法，在数据分析、回归建模、评价预测等方面得到了广泛的应用。

在满足一定的条件时，单一的回归方法可取得较好的预测效果。然而，正是因为受限于使用条件，使得回归方程往往有一定的应用局限性。近年来，很多回归模型借鉴了其他一些常用统计方法的优点，或者把其他统计方法和回归模型相融合，嵌入到回归模型中，改进回归模型的计算方法，借以降低或取消自身限制条件，极大地扩展和提高了回归模型的适应性。偏最小二乘回归就是集成了经典的多元线性回归和一些常见统计方法的综合回归方法，使得模型适应性极强，应用范围非常广泛。

所有的回归模型在计算创建之前，都要对数据样本进行预处理，识别特异点和对多变量集样本实施降维是常用的两种预处理方法。大规模数据的降维技术一直是统计研究中一个非常重要的问题，传统的降维方法往往注重于从所有变量中抽取少数共性的成分变量，在满足一定信息提取精度的前提下，代替原始变量集。如果在变量集中局部变量保持分块共性，即一部分变量的共性不同于另一部分变量，这种情形在多变量数据中是比较常见的，此时很难一次性从全部变量中抽取共性成分变量，或者抽取的共性变量很难解释其物理意义。如果充分考虑到多变量集的局部变量可能存在相同特性，进而对部分变量实施局部降维，然后逐渐扩展降维范围，也可以实现多变量集的降维。

特异点的识别方法非常多，由于样本的各变量之间复杂的相关关系，如多重共线性等，使得特异点的很多识别方法有一定的局限性，很多的限制条件使得识别方法的应用范围变小。在很多应用领域，基于主成分的一些特异点识别方法往往比直接采用原始样本变量集的特异点识别方法有着更加广泛的应用，因为主成分之间互不相关，相互独立，这使得特异点计算方法中不用考虑各成分变量之间的多重共线性等干扰因素。

本书是以论述改进的非线性偏最小二乘回归模型的理论与方法及应用

为主的专著，也论述了在回归建模之前的数据预处理方法。主要论述了一种改进的非线性偏最小二乘回归模型、高维空间的特异点识别方法、二叉树降维方法和降维二叉树评价方法等。所述内容可作为评价、预测和控制研究而应用于各领域，具有广泛的应用前景。

在一些应用领域中，大规模数据样本越来越常见，回归分析是常见的数据分析方法，为了精练回归模型和提高回归模型的精度，在创建回归模型之前经常需要对原始数据预处理。首先识别并剔除特异点，消除噪声，其次对大规模变量进行降维，消除变量集多重共线性等不良因素的干扰和影响，简化计算步骤，最后建立回归模型。

回归建模的常见步骤如图 1－1 所示。

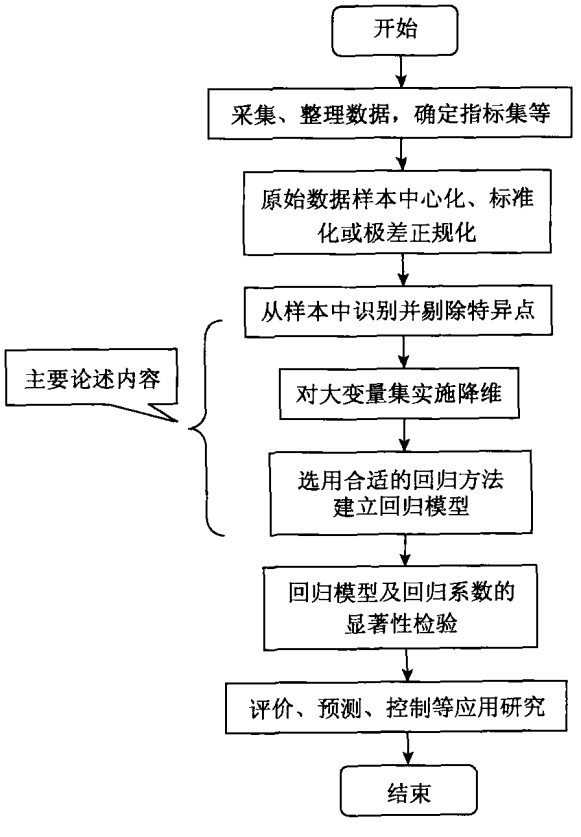

图 1－1 回归建模的常见步骤

综上所述，降维方法、特异点识别方法、非线性偏最小二乘回归模型方法与技术等研究是管理科学与工程的前沿课题，具有重要的理论价值和宽广的应用背景。本书详述了回归建模过程中回归模型、降维、特异点识别三类方法中改进的非线性偏最小二乘回归、二叉树降维及评价方法和基于主成分的特异点识别方法，并对其进行了实证和应用分析。

当今社会处于知识爆炸的时代，信息过量是人人必须面对的问题，海量数据就是信息过量的集中表现。在信息技术飞速发展、经济管理与工程技术问题日益复杂、信息数据快速增长的今天，需要快速有效的数据分析，从海量的、不完全或有缺失数据的、有少量噪声的、模糊的、随机的数据中，去掉噪声，简化、提纯、压缩数据，提取隐含在其中的潜在有用的信息，发现隐藏在数据之间的潜在规律。

识别并剔除特异点可以去除海量数据中的噪声，对海量数据实施降维可以实现数据的简化和提纯，并可提取隐藏于海量数据中的潜在信息，回归建模可以发掘存在于数据中的潜在规律。本书论述研究的出发点和背景正是基于上述三个知识点而展开。

1.1.2 偏最小二乘回归

近年来，随着经济管理与工程实际问题日趋复杂，对数据分析、回归建模的要求也越来越高，在一些应用领域，基于最小二乘准则的一些回归方法已经不能满足要求。在这种背景下，一种新的回归分析方法——偏最小二乘（Partial Least-Squares，PLS）回归应运而生。

偏最小二乘回归是一种基于高维投影思想的新的非参数回归方法，是对一般最小二乘的改进。它比较有效地克服了最小二乘的部分缺点。PLS回归的基本思想在20世纪60年代就出现了，但直到20世纪70年代提出的NIPALS（Nonlinear Iterative Partial Least Squares）算法，才解决了偏最小二乘算法的实现问题；20世纪90年代，在法国专门召开了关于偏最小二乘回归模型研究的国际研讨会，真正有力地促进了PLS方法、理论和应用的快速发展。近年来，国内外关于偏最小二乘回归模型的研究逐渐深入并有所扩展，在理论、方法和应用方面都得到了迅速的发展。

偏最小二乘回归提供了一种多对多线性回归建模的方法，特别是当自

变量和因变量两组变量的个数均很多，且还存在多重相关性，而观测数据的数量（样本量）又较少时，用偏最小二乘回归建立的模型具有传统的回归分析等方法所不具有的优点。

偏最小二乘回归分析在建模过程中集中了主成分分析、典型相关分析和线性回归分析等方法的优点，因此在分析结果中，除了可以提供一个更为合理的回归模型外，还可以同时完成一些类似于主成分分析和典型相关分析的研究内容，提供更丰富、更深入的一些信息，如基于两个主成分变量的特异点识别方法。偏最小二乘分析方法可以有效地将回归建模、主成分分析以及典型相关分析的基本功能有机地结合起来，以致很多文献认为“偏最小二乘＝典型相关分析＋主成分分析＋多元回归”。目前，国外的很多专家学者，如美国顾客满意度指数模型的创立者、密歇根大学的福内尔（Fornell）教授等，都把偏最小二乘回归誉为第二代多元统计分析方法。近年来，偏最小二乘实际应用不断扩展，涉及化学、经济学、社会学、工业、生物、地质、医学以及药物学等领域。

与传统多元线性回归模型相比，偏最小二乘回归能够在各自变量之间存在严重多重相关性的条件下继续进行回归建模，也可以在样本点个数少于变量个数的条件下进行回归建模。偏最小二乘回归在最终模型中包含原有的所有自变量，因此，更易于辨识应用系统的信息与噪声（甚至一些非随机性的噪声），每一个自变量的回归系数也更容易解释其意义。

偏最小二乘回归是一种非常优秀的基于小样本数据集，同时很好地解决了多重共线性问题的一种多元回归方法，但是在建模计算过程中也存在一定的问题，因此，本书提出了改进的非线性偏最小二乘回归方法，修补了原模型计算方法中存在的缺陷。改进的非线性偏最小二乘回归方法具有十分重要的意义，同时也有非常广阔的应用前景。

1.1.3 降维方法

复杂数据在不损失或基本不损失信息的条件下，简单化是数据分析和研究所追求的目标之一，简化的数据也会简化一系列的数据分析计算过程和结果，这对于很多研究领域，如数据挖掘、模式识别、机器学习和评价预测等，都是非常重要的，这就需要研究复杂数据的降维技术。

变量降维是实现复杂数据简单化的一种常见方法。降维方法和技术可以应用于数据挖掘、机器学习、模式识别、多元统计分析等各种领域中，下面以数据挖掘为例简述降维方法。近年来，数据挖掘研究引起了信息产业界的极大关注，其主要原因是存在的大量数据，可以被广泛使用，并且迫切需要把这些庞大数据转换成有用的信息和知识。从中获取的信息和知识有着广泛的实际应用，包括商务管理、经济管理、市场分析预测、生产控制过程、工程设计流程和科学研究等。数据挖掘就是研究隐藏在大规模数据中的有用信息的过程和方法，计算结果可以帮助决策者有效地分析历史和当前的数据，并挖掘出隐藏于其中的相互关系和模式，借以预测未来行为。

一般来说，大规模数据会体现出两个特点，一个是多指标（变量），另一个是大样本。在数据分析过程中，大样本是普遍存在的，获得的样本容量越大，就越有可能抽取出准确的信息。所以，大容量数据样本在数据分析过程中是普遍受到欢迎的。数据分析比较难处理的是多指标问题，首先，大指标集有时并不有助于分析和解决问题，如果原始大指标集中的一些变量并不包含一些有用的解释信息，那么这个大指标集反而有可能会造成最终模型急剧膨胀，数学模型将会变得非常复杂庞大，甚至有可能干扰真实信息。但是，如果丢弃这些信息，一些有用的信息就有可能损失掉。其次，存在于指标集之间的相关性使得多指标之间关系异常复杂，以致很难获得最终的数学模型。最后，如果把存在于指标集之间的相关性和其对模型的线性及非线性影响全部考虑并计算进数学模型，那么，最终数学模型中就会存在大量原始指标集衍生的组合变量，纷繁复杂的数据模型对数据分析和研究将会造成极大的困难，也不利于问题的研究和解决。这就需要变量降维技术，即用少量的变量或成分变量代替原始数据中的相关性较强的变量子集。

传统的降维计算方法侧重于研究数据样本整体降维，本书则从局部数据视角出发，研究从局部降维延伸到全局降维的一种新的降维方法。

1.1.4 特异点识别方法

在样本中，远离样本一般水平的极端大值和极端小值称为特异样本

点，简称为特异点，也称孤立点、歧异值、野值、新颖点、偏离点、离群点、离群值、异常值、极端值、影响点、异常点、例外点、噪声、异常物等。

从大量随机数据样本中识别并剔除特异点，是统计分析、数据挖掘、机器学习、模式识别等很多领域在数据分析研究之前就需要完成的工作。特异点在很多大量的观测样本中存在，对最终数学模型有一定的破坏作用，它会影响、干扰甚至破坏模型的精确度，因此，需要在建模之前从样本中剔除。

很多特异点识别方法有一定的局限性，本书基于一些基本的特异点识别方法，如基于偏最小二乘主成分的 T^2 椭圆图，扩展这些方法以消除其局限性，新的特异点识别方法具有更加广泛的应用范围。

1.2 研究综述

1.2.1 国内研究现状

1. 偏最小二乘回归

偏最小二乘及相关方法国际会议于 1999 年发起，是偏最小二乘回归领域最高水平的国际学术会议，至今已经在欧洲成功举办了 5 届，对于推动偏最小二乘回归的研究与应用起到了重要的作用。2009 年 9 月 4～7 日，由北京航空航天大学经济管理学院主办，法国 ESSEC 高等经济商学院、巴黎 HEC 商学院协办的第 6 届偏最小二乘及相关方法国际会议（The 6th International Conference on Partial Least Squares and Related Methods）在北京航空航天大学召开。此次偏最小二乘及相关方法会议是首次在亚洲国家举行，来自 20 余个国家的偏最小二乘及相关领域的近百位专家学者共同交流了偏最小二乘及相关方法的新成果，探讨了其发展趋势及在实际应用中面临的问题与挑战。

在国内，北京航空航天大学经济管理学院院长王惠文教授于 1994 年开始研究并应用偏最小二乘回归模型，是我国最早引入这一方法的学者之一，并撰写了目前国内经典的两部论述偏最小二乘回归的专著，即《偏最

小二乘回归方法及其应用》和《偏最小二乘回归的线性和非线性方法》。我国著名管理科学家成思危教授、中国人民大学吴喜之教授等，都为偏最小二乘回归理论、技术与应用的研究与推广作出了极大贡献。

近几年，国内外关于偏最小二乘回归的理论和方法研究主要集中在通径分析方法和递阶模型、非线性方法、一些辅助分析技术以及偏最小二乘回归理论的进一步研究与探讨上，这些研究使得偏最小二乘回归研究成果非常丰富，并构成偏最小二乘回归理论与方法体系。鉴于偏最小二乘回归在评价、预测与控制等很多领域的强大功能，其应用研究较为广泛，几乎涉及社会科学和自然科学的各个领域。

非线性偏最小二乘（Nonlinear Partial Least-Squares，NLPLS）回归是偏最小二乘回归理论体系中非常重要的内容，国内外关于 NLPLS 的研究主要还是借助于非线性模型的线性化思想，如基于变量代换、样条函数、核函数变换的 NLPLS 回归方法和偏最小二乘 Logistic 回归等。这些方法采用不同的技术，从不同侧面，在一定程度上实现了非线性偏最小二乘回归方法，但各种 NLPLS 回归模型的计算方法中依然保留了线性化计算过程，并未从根本上彻底实现 NLPLS 回归模型的非线性化。

目前，国内关于线性与非线性偏最小二乘回归的研究成果最多的还是应用领域，即把偏最小二乘回归模型应用于化学、经济学、社会学、生物、工业、军事、地质、农学、医学以及药物学等领域，有些研究还针对每种应用领域，提出了一些适合该领域应用的改进 PLS 模型与方法。

西南交通大学的白裔峰博士把偏最小二乘算法应用于风险最小化的机器学习领域中，提出了高维空间中的核偏最小二乘算法（KPLS），并基于简化的偏最小二乘算法，提出简化核偏最小二乘算法（S-KPLS），实现偏最小二乘算法、支持向量机和模糊系统建模的有机结合。有的研究者把偏最小二乘回归算法应用于建立基于加权块式递推偏最小二乘算法的大坝安全监控统计模型，并提出基于正交信号修正的偏最小二乘回归统计模型。这些研究成果对偏最小二乘回归方法进行了改进，并应用于适合该模型或方法的领域。

2. 降维方法

国内关于降维方法的研究成果很多，很多研究者提出了一系列的降维方

法，根据这些降维方法采用的基本思想，可对它们进行分类，如图 1－2 所示。

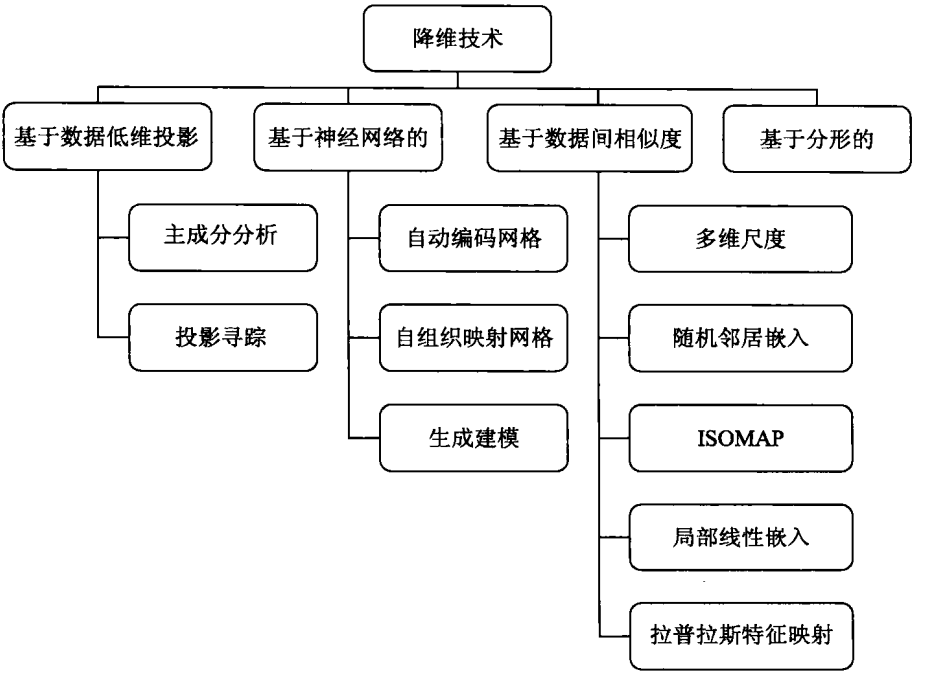

图 1－2 降维技术分类

如图 1－2 所示的降维方法有各自适合的应用领域，且各有优缺点。表 1－1 详细总结了这些降维技术的优缺点。根据表中所总结的常用降维算法的优缺点，在实际应用中可根据具体应用的实践背景和相关领域知识，选择合适的算法，充分发挥各自降维算法的优势。

表 1－1 一些常用降维方法及其优缺点分析

降维方法	优 点	缺 点
主成分分析（PCA）	计算方便，概念简单，具有最优线性重构误差	没有准确的方法可以确定提取主成分的数量；不能处理非线性数据
核主成分分析（KPCA）	能处理非线性数据；是 PCA 在非线性领域的扩展	对核函数依赖性比较强

续 表

降维方法	优 点	缺 点
主曲线	PCA 在非线性领域的扩展	不能证明算法的收敛性
投影寻踪	能有效地剔除噪声	计算量较大；高度非线性数据不适用
自组织映射	实现高维数据可视化	缺乏收敛性定义；不能定义可优化代价函数
贝叶斯神经网格	可以求得参数的后验分布	先验分布是任意选择的
生成拓扑映射	可以应用于数据可视化	不适用于硬降维，即不适用于将数据从很高维空间向较低维空间进行投影处理
多维尺度	可保持数据间的差异性	缺乏统一准则评价嵌入维的质量
局部线性嵌入	计算比较简单直观；属于探索性的数据分析方法，能揭示高维空间中的低维结构	目前主要用于可视化，其他方面应用尚待扩展
基于分形的降维	可以得到数据的分数维估计；体现数据集的本性特征	为得到 D 维数据准确的本征维估计，要求集中数据的个数 $N \geqslant 10^{D/2}$

3. 特异点识别方法

特异点识别是数据挖掘研究的一个重要方面，特异点的识别方法有很多，可以分为五类：基于分布的、基于聚类的、基于距离的、基于深度的和基于密度的特异点识别方法。在国内，基于聚类的和基于距离的特异点识别研究成果较多一些。基于聚类的特异点识别算法先将数据样本分成若干类，不属于任何类的观测点就是特异点。如长沙理工大学的曾颖等人提出了“基于 K-均值聚类和凝聚聚类的离群点查找方法”，根据数据流的特点，给出一种基于 K-均值聚类和凝聚聚类的特异点发现方法。首先利用 K-均值聚类对数据流进行处理，以生成中间聚类结果，再用凝聚聚类对中间结果进行再次选择，最终找出可能存在的特异点。另外，首都师范大学的王妍等人提出了“基于 Voronoi 和空间自相关的离群点检测”方法，这是一种空间特异点检测算法。首先用 Voronoi 来计算空间对象间的邻近关系，然后在空间邻域内再利用空间自相关性计算局部 Moran 指数，将其作