主曲线成分分析

苏 菡[1),2)] 黄凤岗[1)] 贾迪野[1)]

[1)](哈尔滨工程大学计算机科学与技术学院，哈尔滨 150001) [2)](四川师范大学计算机科学与技术学院，成都 610016)

摘 要 广泛应用的第一主成分是对数据集的一维线性最优描述，主曲线是第一主成分的非线性推广。线性主成分分析是一种线性分析方法，而数据通常是非线性的。用线性方法分析非线性数据在分析能力上常常是受限的。为此在对线性主成分分析非线性数据研究的基础上，提出了一种新的非线性成分分析方法，即主曲线成分分析。该方法从数据本身出发进行非线性分析，强调非参数特性，能有效地建模非线性数据。实现主曲线成分分析时，采用了改进的神经网络建模方法，该建模方法以其较强的近似性能很好地表达了非线性关系。仿真实验结果表明，主曲线成分分析能很好地解决非线性主成分问题，应用前景广阔。

关键词 图像分析 主曲线 主曲线成分 神经网络

中图法分类号：TP391 **文献标识码**：A **文章编号**：1006-8961(2005)04-0499-06

Principal Curve Component Analysis

SU Han[1),2)], HUANG Feng-gang[1)], JIA Di-ye[1)]

[1)](*School of Computer Science and Technology, Harbin Engineering University, Harbin* 150001)

[2)](*School of Computer Science and Technology, Sichuan Normal University, Chendu* 610016)

Abstract The first linear principal component is the optimal linear 1-d summarization of the data. Principal curves are nonlinear generalizations of the first linear principal component. Principal component analysis is a linear method, but the most data are nonlinear. Sometimes the linear principal component analysis works inadequately when the data are nonlinear. In this paper, a new nonlinear analytic method, principal curve component analysis (PC^2A) is proposed. This method can model nonlinear data effectively, which analyzes the data from its inherence and emphasizes the non-parametric characteristic. And the method uses the advanced neural network to model data. This is an excellent approach for expressing the nonlinear relationship because of its universal approximation property. Experimental results show that principal curve component analysis is excellent for solving nonlinear principal component problem, and it has great applications potentials.

Keywords image analysis, principal curves, principal curve component, neural network

1 引 言

主成分分析(PCA)已在数字图像处理和计算机视觉等计算机的很多研究领域以及生物、化学、工业、天文、经营管理等领域得到了广泛的应用，现在，PCA仍然是一个研究热点。它是一种有效的线性分析方法，能对数据降维。对线性数据，PCA能保证降维后信息损失最小；然而，对占多数的非线性数据，这种线性分析方法受到较大的限制，用它分析数据可能会造成大量信息丢失。这就需要改进线性分析方法来克服缺陷，即将线性主成分分析进行非线性推广。文献[1]曾指出，目前“向非线性”推广是数据统计分析的研究主流，但存在着不同的技术路线。最经典的研究分为两类：一是根据统计学习理论中的核技术，将数据集合映射到特征空间，在特征空间计算其主成分，即核主成分分析(KPCA)，文献[1]指出该技术路线的本质仍是线性主成分分析；二是从数据本身的分布出发，找到能最好描述数据内在结构的概率分布模型，如矢量量化、主曲线、生成式拓扑映射等。

本文提出了主曲线成分分析(PC^2A, principal curve component analysis)，它是一种从数据本身出发

收稿日期：2004-06-30；**改回日期**：2004-09-09

第一作者简介：苏菡(1979～)，女，教师。2004 年于哈尔滨工程大学获工学硕士学位，现为哈尔滨工程大学图像处理与模式识别专业博士研究生。主要研究方向为生物特征识别、图像处理、视频处理及模式识别。E-mail：susuhan@163.com

进行非线性推广的新分析方法，强调非参数分析，对分析非线性数据相当有效。本文还给出了求取主曲线成分的具体方法，该方法是利用神经网络建模，从数据集中提取主曲线成分并由主曲线成分重构数据点。大量实验结果表明，主曲线成分分析能很好地解决非线性主成分问题，在较小的损失下描述数据。

2 主成分与主曲线

主曲线[1,2]是通过 m 维数据分布或数据“云”的“中间”，并且满足自相合的无参数光滑曲线，这些曲线给出了数据的概貌，是很有效的特征提取工具。其理论基础是寻找嵌入高维空间的非欧氏低维流形。现在广泛应用的第一主成分是对数据集的 1 维线性最优描述，主曲线正是第一主成分的非线性推广。在数学上，主成分和主曲线均是求解一个函数族中满足给定目标函数的最优函数问题，主成分分析限制函数族为一次多项式。

线性主成分分析可定义为

$$\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{\lambda}_i + \boldsymbol{e}_i, \boldsymbol{x}_i \in \boldsymbol{X}$$

其中，$\boldsymbol{x}_i$ 为 m 维数据点，$\boldsymbol{\lambda}_i$ 是生成数据分布的 k 维隐变量，即数据的内在分布变量，$\boldsymbol{A}$ 为 $m \times k$ 维矩阵，描述了数据集与内在分布变量之间的线性关系，$\boldsymbol{e}_i$ 是独立同分布的噪声。主成分分析就是从数据集合中求出 $\boldsymbol{\lambda}$。

如果数据集合 $\boldsymbol{X}$ 是一些非线性数据，大量实验证明线性主成分分析不能很好地描述它们，因此，主曲线应运而生，其形式为

$$\boldsymbol{x}_i = \boldsymbol{f}(\boldsymbol{\lambda}_i) + \boldsymbol{e}_i, \boldsymbol{x}_i \in \boldsymbol{X}$$

其中，$\boldsymbol{\lambda}_i$ 是投影指标，$\boldsymbol{f}(\cdot)$ 表示了投影指标与数据点之间的 k 维矢量函数，是与内在分布变量有非线性关系的描述，进行主曲线成分分析也是要从数据集合中求出 $\boldsymbol{\lambda}$。

3 主曲线理论

Hastie 和 Stuetzle 第 1 次提出了主曲线，曲线的形状是由数据点决定的，可以较好地描述非线性数据。

定义 1（HS 主曲线） 光滑的曲线 $\boldsymbol{f}(\boldsymbol{\lambda})$ 如果满足以下要求，它就是 $\boldsymbol{X}$ 集合的主曲线。

（1）$\boldsymbol{f}$ 不自相交，

（2）$\boldsymbol{f}$ 在有界子集 $\mathbf{R}^d$ 中是有限长的，

（3）$\boldsymbol{f}$ 是自相合的。自相合是指，曲线上的每一点是投影至该点的数据点的条件均值。即

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{\lambda}) &= E(\boldsymbol{X} \mid \boldsymbol{\lambda}_f(\boldsymbol{X}) = \boldsymbol{\lambda})\\ \boldsymbol{\lambda}_f(\boldsymbol{x}) &= \boldsymbol{f}^{-1}(\boldsymbol{x}) = \sup_{\lambda}\{\boldsymbol{\lambda}: \| \boldsymbol{x} - \boldsymbol{f}(\boldsymbol{\lambda}) \| \\ &= \inf_{\omega} \| \boldsymbol{x} - \boldsymbol{f}(\omega) \| \}\end{aligned}$$

其中，$\boldsymbol{\lambda}$ 是投影指标，$\boldsymbol{\lambda}_f(\boldsymbol{x})$ 是与 $\boldsymbol{x}$ 和 $\boldsymbol{f}(\boldsymbol{\lambda})$ 之间的最小正交距离对应的 $\boldsymbol{\lambda}$。当数据点在曲线上获得多个相等的最小距离投影点时，取投影坐标最大的一个。曲线 $\boldsymbol{f}(\boldsymbol{\lambda})$ 可以看成是单变量 $\boldsymbol{\lambda}$ 的函数，$\boldsymbol{f}(\boldsymbol{\lambda}) = (f_1(\boldsymbol{\lambda}), f_2(\boldsymbol{\lambda}), \cdots, f_m(\boldsymbol{\lambda}))$，$(f_1(\boldsymbol{\lambda}), f_2(\boldsymbol{\lambda}), \cdots, f_m(\boldsymbol{\lambda}))$ 是坐标函数。

主曲线本质是嵌入欧氏空间的 1 维流形，它的一般化描述为：令 $\boldsymbol{D}$ 为数据域，$\boldsymbol{F}$ 为函数集，对于每个 $f \in \boldsymbol{F}$，$f: \boldsymbol{D} \to \mathbf{R}^m$，称 f 的值域为由 f 生成的流形 M_f，即 $M_f = f(\boldsymbol{D}) = \{f(\boldsymbol{x}): \boldsymbol{x} \in \boldsymbol{D}\}$。在 $\boldsymbol{F}$ 中全部函数所生成的所有流形的集合定义为 $\boldsymbol{M} = \{M_f: f \in \boldsymbol{F}\}$。流形的距离函数定义为数据集合 $\boldsymbol{X}$ 和 $\boldsymbol{M}$ 之间的期望距离，即 $\Delta(\boldsymbol{M}) = E[\Delta(\boldsymbol{X}, \boldsymbol{M})]$。

HS 主曲线描述了数据分布的中间性、自相合以及无参数特征，但是其原始定义及以后的一些应用仍然存在不足。如 HS 主曲线的存在性至今未得到证明（除了椭圆分布等一些特例），还存在模型偏差和估计偏差，收敛性问题等。为了克服 HS 主曲线的不足，Banfield 和 Raftery 主曲线在算法上改进了 HS 主曲线，解决了在闭主曲线下由于曲率过大的问题；针对模型偏差，Tibshirani 引入了半参数方法，重新定义的主曲线基于混合模型；Kégl 引入了有长度约束的主曲线概念；Morales 从微分流形角度研究了几何数据拟合问题，指出主曲线是流形拟合中的特例；Smola 提出了寻找具有多种正则项的正则主流形；Chang 和 Ghosh 结合生成式拓扑映射定义了概率主曲线及概率主曲面，在保持 HS 主曲线自相合特点的同时采用参数模型将主曲线延伸到较高维主流形。这些主曲线都在不同方面解决了 HS 主曲线存在的问题。

4 主曲线成分分析

对任意向量集 $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$，$\boldsymbol{X} \in \mathbf{R}^m$，$\boldsymbol{x}_i = (x_i(1), x_i(2), \cdots, x_i(m))^{\mathrm{T}}$，$i = 1, \cdots, n$，本文提出的主曲线成分分析，是将 $\boldsymbol{f}(\boldsymbol{\lambda})$ 的投影指标 $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \cdots, \boldsymbol{\lambda}_n)$ 作为数据的主曲线成分。$\boldsymbol{f}(\boldsymbol{\lambda})$ 是依弧长参数化的曲线，$\boldsymbol{\lambda}$ 描述了序结构。

对于已知分布的数据集，通过直接求主曲线

$E[\boldsymbol{X}|\lambda_f(\boldsymbol{X})]=f(\lambda)$,自然得到 λ;对未知分布的点集,通过估计 $E[\boldsymbol{X}|\lambda_f'(\boldsymbol{X})]$ 来找到其主曲线成分。为了去除数据内部的相关性,可以先对它进行 PCA 变换,将变换后不相关矢量的主曲线成分作为数据集的主曲线成分。$s_i=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i, i=1,\cdots,n$,其中,$s_i$ 为主成分,$\boldsymbol{A}$ 为变换矩阵,$\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_m)$,特征向量按降序排列。$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$ 经 PCA 变换为

$$\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n)$$

$$\boldsymbol{s}_i=(s_i(1),s_i(2),\cdots,s_i(m))$$

再对 $\boldsymbol{S}$ 进行分析,可得到 $f(\lambda)=E[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}|\lambda_f(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X})]$ 的 λ。

利用特征向量 $\boldsymbol{a}_i, i=1,\cdots,m$ 为正交向量,即 $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{a}_j=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases}$ 的特性,可以将数据集进一步主曲线成分分析[3]。即将 k 个正交向量构成子空间,在子空间上进行主曲线分析,其中,$k\geqslant 1$。

$k=1$ 时,该分析就是上述的分析方法;$k=2$ 时,按子空间的性质,将各特征向量组合张成特征子空间。在构造特征子空间时,依特征向量的降序成对组合。即由 $\boldsymbol{a}_1,\boldsymbol{a}_2$ 构成子空间 1,$\boldsymbol{a}_3,\boldsymbol{a}_4$ 构成子空间 2,以此类推,$\boldsymbol{a}_{m-1},\boldsymbol{a}_m$ 构成子空间 d,$d=\frac{m}{k}$。若 d 不能整除 k,则从后向前找 k 个向量构成最后一个子空间,这样进一步避免了信息丢失。其中,子空间 1 中的样本集

$$\boldsymbol{S}_1=\left\{\begin{pmatrix}s_1(1)\\ s_1(2)\end{pmatrix},\begin{pmatrix}s_2(1)\\ s_2(2)\end{pmatrix},\cdots,\begin{pmatrix}s_n(1)\\ s_n(2)\end{pmatrix}\right\}$$

子空间 2 中的样本集

$$\boldsymbol{S}_2=\left\{\begin{pmatrix}s_1(3)\\ s_1(4)\end{pmatrix},\begin{pmatrix}s_2(3)\\ s_2(4)\end{pmatrix},\cdots,\begin{pmatrix}s_n(3)\\ s_n(4)\end{pmatrix}\right\}$$

以此类推子空间 d 中的样本集

$$\boldsymbol{S}_d=\left\{\begin{pmatrix}s_1(m-1)\\ s_1(m)\end{pmatrix},\begin{pmatrix}s_2(m-1)\\ s_2(m)\end{pmatrix},\cdots,\begin{pmatrix}s_n(m-1)\\ s_n(m)\end{pmatrix}\right\}$$

由 $\boldsymbol{S}_1$ 生成主曲线 $\boldsymbol{f}_1$,将 $\boldsymbol{S}_1$ 中的样本投影到 $\boldsymbol{f}_1$ 上,得 $\boldsymbol{S}_1'=\{\lambda_1(1),\lambda_2(1),\cdots,\lambda_n(1)\}$,由 $\boldsymbol{S}_2$ 生成主曲线 $\boldsymbol{f}_2$,将 $\boldsymbol{S}_2$ 中的样本投影到 $\boldsymbol{f}_2$ 上,得 $\boldsymbol{S}_2'=\{\lambda_1(2),\lambda_2(2),\cdots,\lambda_n(2)\}$,同理得 $\boldsymbol{S}_d'=\{\lambda_1(d),\lambda_2(d),\cdots,\lambda_n(d)\}$。

主曲线成分为 $\boldsymbol{\lambda}=(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2,\cdots,\boldsymbol{\lambda}_n)^{\mathrm{T}}$,其中,$\lambda_i$ 为各子空间主曲线的投影指标。因此通过参数化成对主成分 $s(2i-1),s(2i)$ 为 $\lambda_{2i-1}, i=1,\cdots,\frac{m}{k}$,完全可以用 $\boldsymbol{\lambda}$ 近似代替 $\boldsymbol{X}$,达到特征压缩的目的。

经过主曲线成分分析后,集合 $\boldsymbol{X}$ 中的数据点 $\boldsymbol{x}\approx\sum_{i=1}^{d}\boldsymbol{f}_i(\boldsymbol{\lambda}_i)$。主曲线成分分析是无参数的,它可以方便地对各种数据分布建模,对未知分布的数据进行分析;其重构数据的误差较 PCA 小。另外,当主曲线是一条直线时,它就是线性主成分。主曲线是距离函数的临界值,因为,$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}D^2(\boldsymbol{x},\boldsymbol{f}_\varepsilon)_{\varepsilon=0}=0$,其中,$\varepsilon$ 是扰动因子,$\boldsymbol{f}_\varepsilon$ 是一 $\boldsymbol{f}_0=\boldsymbol{f}$ 的光滑曲线族,$D^2(\boldsymbol{x},\boldsymbol{f})=Ed^2(\boldsymbol{x},\boldsymbol{f}),d(\boldsymbol{x},\boldsymbol{f})=\|\boldsymbol{X}-\boldsymbol{f}(\lambda_f)\|$。这点与线性分析方法相同,当距离函数 $\boldsymbol{f}_\varepsilon$ 是一光滑直线族时,线性主成分也是其临界值。

非线性 PCA 与主曲线的主要区别在于主曲线允许投影指标 $\lambda_f(\boldsymbol{x})$ 在某个特殊点上不连续,而非线性 PCA 要求其是连续的,这就使得非线性 PCA 可能只找到了次优解 $(\hat{\boldsymbol{f}},\hat{\lambda})$,$\boldsymbol{x}$ 的投影不再是 $\hat{\boldsymbol{f}}$ 到 $\boldsymbol{x}$ 的最近点:$\|\boldsymbol{x}-\hat{\boldsymbol{f}}(\hat{\lambda}(\boldsymbol{x}))\|>\inf_{\lambda}\|\boldsymbol{x}-\hat{\boldsymbol{f}}(\lambda)\|$。

5　PC²A 的具体实现

在求主曲线成分时,首先需要求得数据点集对应的主曲线,其求解算法如下:假设已知 $\boldsymbol{X}$ 的数据分布。

(1) 初始化　令初始曲线 $\boldsymbol{f}^{(0)}(\lambda)=\bar{\boldsymbol{X}}+a\boldsymbol{\lambda}$ 为 $\boldsymbol{X}$ 的第一主成分线,其中,a 是 $\boldsymbol{X}$ 的第一线性主成分,$\bar{\boldsymbol{X}}$ 是 $\boldsymbol{X}$ 的均值。但 $\boldsymbol{f}^{(0)}$ 也可以是值较小的随机矢量。设 $j=0$;

(2) 投影　对所有的 $\boldsymbol{x}\in\boldsymbol{X}$,求 $\lambda_{f^{(j)}}(\boldsymbol{x})=\max\{t:\|\boldsymbol{x}-\boldsymbol{f}(\lambda)\|=\min_{\tau}\|\boldsymbol{x}-\boldsymbol{f}(\tau)\|\}$;

(3) 求期望　求取新的主曲线 $\boldsymbol{f}^{(j+1)}(\lambda)=E[\boldsymbol{X}|\lambda_{f^{(j)}}(\boldsymbol{X})=\lambda]$。计算距离函数 $D^2(\boldsymbol{X},\boldsymbol{f}^{(j+1)})=E\lambda_{f(X)}E[\|\boldsymbol{X}-\boldsymbol{f}(\lambda_f(\boldsymbol{X}))\|^2|\lambda_f(\boldsymbol{X})]$ 如果 $|D^2(\boldsymbol{X},\boldsymbol{f}^{(j)})-D^2(\boldsymbol{X},\boldsymbol{f}^{(j+1)})|/D^2(\boldsymbol{X},\boldsymbol{f}^{(j)})$ 小于某个阈值,则停止;否则,令 $j=j+1$,转到第 2 步投影。

若数据分布未知,数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\subset\mathbf{R}^n$,投影指标 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 按升序排列,曲线 $\boldsymbol{f}$ 表示成连续投影点形成线段 $(f(\lambda_i),f(\lambda_{i+1}))$ 构成的多边形曲线。距离函数等于 $\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{f}^{(j+1)}(\lambda_i^{(j)})\|^2$。

主曲线本身是流形的 1 维形式[4],然而流形不是欧氏空间,流形内的点并没有坐标,在求取主曲线时,可以利用将流形与欧氏空间对应的同胚映射 $\varphi_\alpha:\boldsymbol{U}_\alpha\rightarrow\boldsymbol{V}_\alpha$,把流形与 m 维欧氏空间的非空开集对应,其中,$\alpha\in\boldsymbol{\Lambda}$,$\boldsymbol{\Lambda}$ 为指标集。也就是它可以局部欧

氏空间化，主曲线的每一个局部开集 U_α 都和欧氏空间的一个开集 V_α 同胚。对于主曲线上的某点 p，通过同胚映射 φ_α，点 p 映射为 V_α 中的点 $x=(x_1,x_2,\cdots,x_m)$。

用以上求主曲线的方法可以逐步求得主曲线成分，为了更直观化，在实际计算中，运用改进的神经网络[5]方法来获得相关值。

定义 λ 为主曲线成分，对于点集 $X=\{x_1,x_2,\cdots,x_n\}$，其主曲线成分为 $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_n)^{\mathrm{T}}$，记 $\boldsymbol{\lambda}(1)=(\lambda_1(1),\lambda_2(1),\cdots,\lambda_n(1))^{\mathrm{T}}$。选取一个主曲线成分时，$X$ 可表为 $X=f(\boldsymbol{\lambda}(1))+E_1$，其中，$E_1$ 是误差。当然 X 可以用多个主曲线成分表示，即 $X=F(\boldsymbol{\lambda})+E,\boldsymbol{\lambda}=(\boldsymbol{\lambda}(1),\boldsymbol{\lambda}(2),\cdots,\boldsymbol{\lambda}(i))^{\mathrm{T}},i=1,2,\cdots,m$ 为主曲线成分矩阵，E 是误差。

$F(\cdot)$ 是一非线性函数，在众多的建模方法中，神经网络具有较通用的近似特性。文献[5]采用改进的神经网络建模 $F(\cdot)$，求取主曲线成分和重构数据建模。Hornik 曾证明只用 1 层 S 形单元隐含层的神经网络能较好地近似目标。故求取主曲线成分可采用 3 层前馈神经网络。其输入为 m 维原始数据或经过 PCA 变换后的数据，输出是主曲线成分。区别于 Dong 提出的神经网络建模，这里的主曲线成分输出个数可选，结构如图 1 所示。

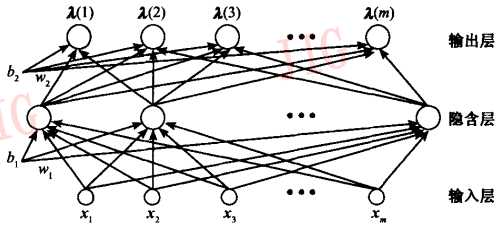


图 1 3 层神经网络求主曲线成分

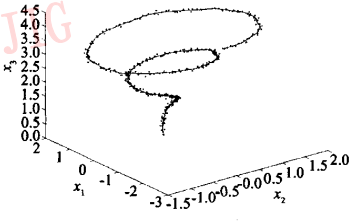
Fig. 1 PC^2 by 3 layers feed-forward neural network

神经网络训练时运用梯度下降法。隐含层节点个数对神经网络的学习至关重要，本文依照 Stone 提出的交叉验证策略决定不同维数据所需的隐含层节点个数。

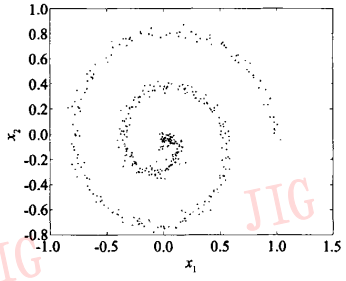
6 实验结果与分析

为了说明采用主曲线成分分析能有效地描述和分析非线性数据，给出了两个典型的例子。计算主曲线时，采用了均方误差 $MSE^{(j)}=E\{\|X-f[\lambda_f(X)]\|\}$ 作停止判断。

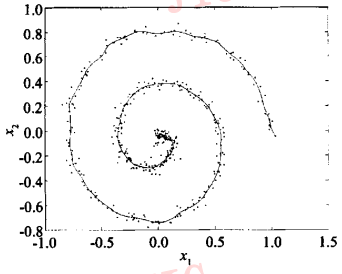
图 2 为对 400 个 3 维加噪声数据点进行主曲线



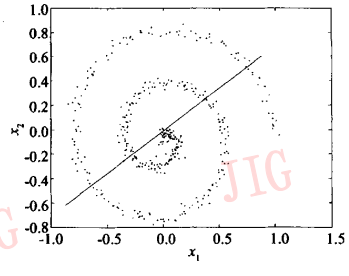
(a) 原始数据主曲线分析



(b) 投影至 2 维坐标系



(c) 投影后主曲线分析



(d) 投影后第一主成分线

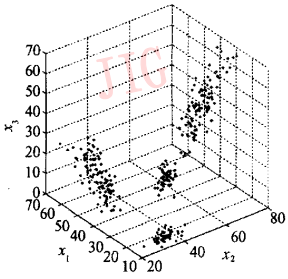
图 2 3 维数据主曲线分析

Fig. 2 PC^{2}A of the 3D space data

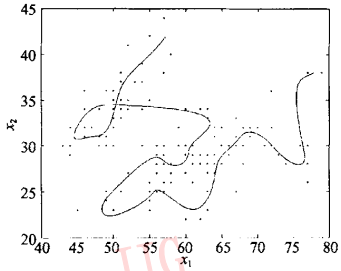
成分分析。为了便于观察,将数据点投影至 2 维坐标系中。

计算得,投影至 2 维坐标系后,采用 PCA 产生的 MSE 为 6.268,采用 PC^2A 产生的 MSE 为 0.359。

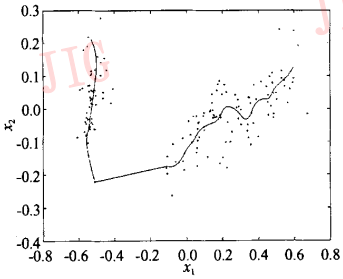
图 3 和表 1 是主曲线成分分析和线性主成分对 Iris 数据库的分析。这个数据库包含了 150 个 4 维矢量。



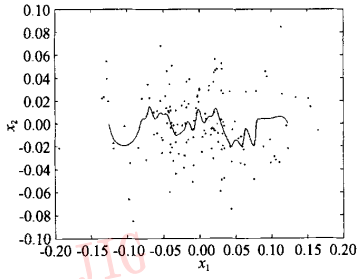
(a) Iris 原始数据



(b)未经 PCA 的 PC^2A



(c) PCA 后第 1 子空间



(d) PCA 后第 2 子空间

图 3　主曲线成分分析 Iris 数据集

Fig. 3　Results of PC^2A on the Iris dataset

表 1　主曲线成分分析和主成分分析重建误差

Tab. 1　Reconstruction Error using PC^2A and PCA

MSE	成分数目			
	1	2	3	4
PCA	0.899	0.426	0.130	0
PC^2A(未 PCA)	0.324			
PC^2A(PCA 后)	$0.482^{(1)}$	$0.520^{(2)}$		

注:上角(1)表示前两个特征子空间上进行的主曲线成分分析;
上角(2)表示后两个特征子空间上进行同样操作

表 1 是直接用主曲线、PCA 后主曲线成分和分别用 1 个、2 个、3 个线性主成分重建这些数据的误差对比,其中,在进行子空间构造时,k 取 2。可以看到,仅用一条主曲线直接描述数据比用两个线性主成分更好地再现了数据。图 3(b)是数据未去相关性,直接由主曲线进行分析的结果;图 3(c)和图 3(d)是经过 PCA 后形成两个子空间,再进行主曲线成分分析的结果。可以看出在数据点相关性较大时先进行 PCA 有优势。但是,当数据间相关性不大时,不进行 PCA 的主曲线成分分析可以节省时间,同时减小了分析误差。

从上面的图 3、表 1 中可以看出,采用线性主成分对非线性数据分析的效果不是很好,其压缩误差大,信息保持量不高,而采用主曲线成分分析方法,则效果明显不同,它能更准确的分析非线性数据。

7　结　论

主曲线成分分析是 PCA 的非线性推广,它与 KPCA、ICA 从不同角度描述数据内在结构。大量实验结果表明,主曲线能很好地描述数据,主曲线成分能在较小损失的情况下进行数据降维和特征提取,特别是分析稀疏的高维非线性数据更有效。由此,采用主曲线成分分析线性和非线性数据是很有意义的。但主曲线涉及到较复杂的数学知识,将其直接推广到曲面分析存在一些问题,如主曲线作为流形,存在如何局部欧式空间化的问题;主曲线理论不能直接用于 2 维以上主曲面,主曲线分析方法也就不

能简单地推广到曲面上等。下一步的研究将放在主曲线分析推广到高维主流形分析上,这是一个很富挑战性的研究课题。

参考文献(References)

1 Zhang J P, Wang J. An overview of principal curves[J]. Chinese Journal of Computers, 2003, **26**(2): 129 ~ 146. [张军平, 王珏. 主曲线研究综述[J]. 计算机学报, 2003, **26**(2): 129 ~ 146.]

2 Hastie T. Principal curves and surfaces [D]. USA: Stanford University, 1984.

3 Chang Kui-yu, Ghosh Jodeep. A unified model for probabilistic principal surfaces[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, **23**(1): 22 ~ 41.

4 Ouyang G Z. Calculus on Manifolds [M]. Shanghai: Shanghai scientific & Technical Publishers, 1988: 101 ~ 110. [欧阳光中. 流形上的微积分[M]. 上海: 上海科学技术出版社, 1988: 101 ~ 110.]

5 Dong D, McAvoy T J. Nonlinear principal component analysis-based on principal curves and neural networks[J]. Computers Chemical Engineering, 1995, **20**(1): 65 ~ 78.