HS主曲线的数学特性*)

王 真 苗夺谦 张红云
(同济大学计算机科学与工程系 上海200092)

摘 要 主曲线被定义作穿过多维数据分布"中间"的满足"自相合"的光滑曲线,它是第一主成分的非线性推广,第一主成分是对数据集的一维线性最优描述。HS主曲线强调非参数模型,对其参数无关性本文给出了具体证明。同时为了全面理解主曲线,本文以空间主曲线为例,分析了它的横截性质。

关键词 主曲线,自相合,非参数化,横截性

Mathematics Properties of HS Principal Curves

WANG Zhen MIAO Duo-Qian ZHANG Hong-Yun
(Department of Computer Science and Engineering, Tongji University, Shanghai 200092)

Abstract Principal Curves have been defined as 'self-consistent' smooth curves passing through the middle of a multidimensional data set. They are nonlinear generalizations of the first principal component, which are optimal linear 1-d generalizations of the data. HS principal curves emphasize on the nonparameteried model, in this paper we discuss and prove the nonparametric property . Besides this, the goal of this paper is to further contribute to the theoretical understanding of principal curves, we analysize the transveriality of the principal curves in the 3-d Eulid space.

Keywords Principal curves, Self-consistency, Nonparameteric property, Transveriality

1 引言

我们先来看一组多维随机变量 $X\in R^d$,假设其服从密度函数为 p 的概率分布,样本数为 n,记作 $X=\{X_1,X_2,\cdots,X_n\}$。找一条能最好拟合这组数据的直线,那么必定是第一主成分线[1]。若 X 的分布呈椭圆(球)状,则其第一主成分恰为穿过中心的主轴线。

早在1904年,Spearman 便提出了线性主成分分析的思想。近几十年来,许多学者开始着眼于将主成分向非线性推广的研究。例如,Gnanadesikan 和 Wilk (1966)[4]; Srivastava (1972)[5]; Yohai, Ackermann, 和 Haigh (1985)[6]; Koyak (1987)[7];以及 Gifi(1990)[8]等。他们有的尝试通过将观测量进行非线性变换来实现,有的提出非线性相关函数将低维线性空间映射到数据空间。

到1989年,Hastie 和 Stuetzle 对这个问题给出了新的解决办法,首次提出主曲线的思想。形象地讲,HS 主曲线就是一条穿过数据分布"中间"的光滑曲线。其定义中对"中间"的含义是通过"自相合性"来诠释的,即主曲线上任意一点是投影到这一点的所有数据点的均值。计算中,Hastie 强调非参数方法,换句话说,不事先给定曲线类型,而是从曲线簇中选择满足自相合的具有中间性的曲线。这一点使 HS 主曲线更易于描述现实世界,因为通常情况下我们需要从一堆看似杂乱无章的数据点中发现其内在联系和规律。而非参数方法正适应这一特点。那么为什么说 HS 主曲线是参数无关的呢?针对这一点,本文将给出理论上的论证。同时为了进一步加深对主曲线的理解,我们还将探讨空间主曲线的几何性质。

由于主曲线的这些性质和优点,自20世纪90年代以来在国外取得了较快的发展。虽然在主曲线的原理中使用了较复杂的数学,但由于其广泛的应用前景,在90年代后期已引起国外计算机科学家的关注,现在他们已报道了许多主曲线在计算机方面的应用,如线性对撞机中对电子束运行轨迹的控制、图像处理中辨识冰原轮廓、手写体的主曲线模板化和数据可听化等。

2 HS主曲线的参数无关性

从文[3]中我们可以知道 HS 主曲线的表示是依弧长参数化的,同时它又强调非参数模型,这似乎有些矛盾,其实两者非但不抵触,反而有因果承接关系。弧长参数又称作自然参数,它与曲线类型无关。针对这一点本节将给出详细阐述。

由于讨论中需要使用曲线的基本几何属性,下面简要引入曲线论[2]中的一些基本概念:

定义1 d 维欧氏空间中的一条曲线是一个连续函数 $r:I\to R^d$. $I=[a,b]$是实线上的一个闭区间。

曲线 $r(t)$可看成一个单参数 t 的 d 维函数,$r(t)=(r_1(t),r_2(t)\cdots r_d(t))$,其中$(r_1(t),r_2(t)\cdots r_d(t))$称为坐标函数。

定义2 若在 $t_0\in(t_1,t_2)$处 $\left|\frac{dr(t)}{dt}\right|\neq 0$,则对应于参数 t_0 的点称为曲线的正则点,如果一条曲线上每一点都是正则点,则称为正则曲线。

这里$\frac{dr(t)}{dt}$表示曲线 $r(t)$的切向量,也记作 $r'(t)$。其中

$r'(t)=(r_1'(t),r_2'(t)\cdots r_d'(t))$,

$\left|\frac{dr(t)}{dt}\right|=\sqrt{(r_1'(t))^2+(r_2'(t))^2+\cdots(r_d'(t))^2}$。

定义3 正则曲线 $r(t)$ 从点 $r(t_0)$ 到 $r(t)$ 的弧长表示为:

$s(t)=\int_{t_0}^{t}|r'(t)|dt$。

在这三条概念的基础上,下面介绍一下 HS 主曲线的定义[3]:设 $r(s)=(r_1(s),r_2(s),\cdots r_d(s))$是 R^d 上的一条光滑曲线($s\in I,r:I\to R^d$ 是一连续可微函数),若满足自相合性 $r(s)=E(X|s_r(X)=s)$,则 $r(s)$为 X 数据分布的主曲线。其中 s_r

*)本文得到国家自然科学基金项目(No. 60175016)和上海市教委"曙光计划"项目的资助。王 真 硕士研究生,研究方向为智能计算理论、主曲线。苗夺谦 教授,博士生导师,研究方向为人工智能、模式识别、数据挖掘、粗糙集理论、主曲线。

$(X)=\sup\{s:\|X-r(s)\|=\inf_{r}\|X-f(r)\|\}$，称为投影指标。

理论上说，每条正则曲线总可以用弧长作为它的参数，若$r(s)$是概率密度为p的数据分布的主曲线，在一些近似条件下，$r(s)$可认为是正则曲线。HS主曲线的投影指标隐含了序的性质，而序的特点利用了曲线的弧长作为参数来体现。HS主曲线是依弧长参数化的曲线，正是由于这一点，所以说它是非参数的。这听似矛盾，其实不然。

因为弧长是曲线本身的一个几何量，不受曲线表达式的影响，以下我们将通过定义3证明弧长是与参数无关的，也就是说假设已知曲线的类型，此曲线记作$r(t)$，若换作另一参数t^*，它所确定的弧长一致。

证明：设$r(t)=r(t(t^*))=r^*(t^*)$，其中新、旧参数之间的关系是$t=t(t^*)$，并且设$t\in(a,b)$，t^*对应的区间为$t^*\in(c,d)$。

因为$\frac{dt}{dt^*}\neq0$(否则t与t^*无关)，所以成立

$$\int_a^d\left|\frac{dr^*}{dt^*}\right|dt^*=\int_a^d\left|\frac{dr}{dt}\frac{dt}{dt^*}\right|dt^*=\int_a^d\left|\frac{dr}{dt}\right|\left|\frac{dt}{dt^*}\right|dt^* \tag{1}$$

(i) 若$\frac{dt}{dt^*}>0$，$\left|\frac{dt}{dt^*}\right|=\frac{dt}{dt^*}$，这时$a=t(c)$，$b=t(d)$则(1)式$=\int_a^d\left|\frac{dr}{dt}\right|\frac{dt}{dt^*}dt^*=\int_a^b\left|\frac{dr}{dt}\right|dt$；

(ii) 若$\frac{dt}{dt^*}<0$，$\left|\frac{dt}{dt^*}\right|=-\frac{dt}{dt^*}$，这时$a=t(d)$，$b=t(c)$，可以得到同样的结果。

这说明弧长的定义是与曲线的参数无关的。在这个意义上，我们说HS主曲线具有参数无关性，即事先不需要知道曲线的类型。

3 空间主曲线的横截性质

文[9]讨论了平面上主曲线的几何性质，提出平面上均匀分布的主曲线是某个微分方程的解，进而说明某些几何区域上均匀分布的主曲线有多条且相交。本节将讨论三维欧氏空间中主曲线，就其横截性质进行分析。我们首先把HS主曲线定义的背景和前提介绍一下，因为其性质也必定建立在这些基础上。

3.1 定义与说明

令$\Omega\in R^3$为一紧致空间，三维随机变量$X\in\Omega$服从概率密度为p的分布，有有限二阶矩。$r(s)$为此数据分布在紧支撑上的主曲线。则对于每一个观察点$X_i\in\Omega$，$r(s)$上至少存在一个投影点$r(s_r(X_i))$。由投影指标的定义(见第2节)可知，若样本点在曲线上获得多个相等的最小距离投影点时，取投影指标最大的一个，从而排除了存在歧义点的可能。这里歧义点是指数据点到曲线的投影不唯一，即具有多个相同距离的投影点。在这种情况下，投影指标$s_r(X)$不连续，Hastie已证明，在主曲线的计算中虽然存在歧义点，如果限定可微曲线的长度，则所有歧义点集Lebesgue测度为零，因此可以忽略。从而$s_r(X)$可认为是连续的。

为了叙述方便，我们约定数据点X的投影点为$g(X)$，$g:R^3\to r(s)$，即$g(X)=r(s_r(X))$。过$r(s)$上一点A，曲线的单位切向量记作$T(A)$，即$T(A)=r'(s)$。过A点$r(s)$的法平面记作$\gamma(A)$。我们认为概率密度p在Ω内部连续且严格为正。

3.2 横截性质

在上一小节约定的前提下，我们来分析空间主曲线的几何性质。文[9]曾研究了二维主曲线的诸多性质，通过分析比较我们发现三维Eulid空间中主曲线也有类似性质：

定理1 如果数据分布X的主曲线$r(s)$在概率密度p的紧支撑Ω上，则有

(1)$r(s)$的端点落在Ω的边界上；

(2)$r(s)$与Ω边界在其端点处正交；

(3)$r(s)$的端点是Ω的局部弱凸点。

要证明以上三条性质，我们先来看一个引理

引理1 若$r(s)\subset\Omega$是数据分布的一条主曲线，则任意一点$X\in\Omega$必落在其投影点$g(X)$所在的法平面上，即对$\forall X\in\Omega$，$X\in\gamma(g(X))$

证明：选取一点$X\in\Omega$，若$g(X)$是$r(s)$的内点，由于投影$g(X)$实现X到$r(s)$的距离，显然引理成立；若$g(X)=A$是主曲线的一个端点，下面将证明这种情形结论也成立。

$r(s)$切向量总与弧长增加方向一致(详见文[2])，我们选此方向为正向，那么所有投影至A的点Y满足$(Y-A)\cdot T(A)\geqslant0$。否则$Y$到$r(s)$的距离会严格小于$\|Y-A\|$。因此集合$g^{-1}(A)\cap\Omega$包含于$\Omega$中$\gamma(A)$及$\gamma(A)$外侧(沿正向的一侧)的区域中。假设引理不成立，则存在点$Y(g(Y)=A)$使$(Y-A)\cdot T(A)>0$，即落在$\gamma(A)$外侧的Ω区域H中。取一开集$Q\subset H$，基于p在紧支撑Ω上连续且严格为正，所以有：

(1)$g(Q)=A$；

(2)开集Q上$p>0$；

(3)对所有$Y\in Q$，满足不等式$(Y-A)\cdot T(A)>0$。

综合这三条可得出$E(X|g(X)=A)$是H的内点，这违背了自相合条件$E(X|g(X)=A)=A$(A在H边界上)。引理得证。

再来看定理1，若性质(1)不成立，$r(s)$的端点不落在Ω边界上，则过端点作$r(s)$的法平面γ_0，总存在投影至端点的Ω的边界点不在γ_0上，与引理矛盾。同理可推得性质(2)、(3)。至此我们说明了空间主曲线具备上述三条性质。

结束语 HS主曲线被定义作满足自相合性的穿过数据集合“中间”的光滑曲线，是线性主成分的自然推广，具有开创性意义。后来又有不少学者分别从不同方面对其进行改进，但相比之下，HS主曲线优点之一在于其参数无关性，本文着重在理论上论述了这一特性。另外，为了更深入理解主曲线的性质，本文就空间主曲线为例给出了几点说明。读者可利用微分流形方法将其向高维推广，用与主曲线正交的超平面代替法平面进一步研究探讨。

参考文献

1 张军平，王珏. 主曲线综述[J]. 计算机学报，2003，26(2)：129～146

2 王申怀，刘继志. 微分几何. 北京：北京师范大学出版社，1988

3 Hastie T, Stuetzle W. Principal curves. Journal of the American Statistical Association, 1988, 84(406): 502～516

4 Gnanadesikan R, Wilk M B. Data analytic methods in multivariate statistical analysis. In: Krisnaiah P R, ed. Multivariate analysis, 1966, 2

5 Srivastava J N. An information approach to dimensionality analysis and curved manifold clustering. In: Krisnaiah PR, ed., Multivariate analysis, 1972, 3

6 Yohai V J, Ackermann W, Haigh C. Nonlinear principal components. Quality and Quantity, 1985, 19: 53～69

7 Koyak R. On measuring internal dependence in a set of random variables. Annals of Statistics, 1987, 15: 1215～1228

8 Gifi A. Nonlinear Multivariate Analysis. New York: John Wiley, 1990

9 Duchamp T, Stuetzle W. Geometric properties of principal curves in the plane. Robust Statistics, Data Analysis, and Computer Intensive Mehtods. New York: Springer Press, 1996, 109: 135～152

10 Kegl B. Principal curves: Learning, design, and applications: [Ph D dissertation]. Concordia University. Canada, 1999

11 Delicado P. Another look at principal curves and surfaces. Journal of Multivariate Analysis, 2001, 77(1): 84～116