主元分析中的平滑性

向 傀[1],周申培[1],李炳南[2]
(1.武汉理工大学自动化学院,湖北武汉 430070;2.合肥工业大学医学工程学院,安徽合肥 230009)

摘 要: 某些样本观测形如时间序列或离散信号,其本质为平滑曲线(即函数型数据),代表一个潜在的连续过程.在主元分析中引入平滑性,可更加全面地刻画样本观测中包含的连续动态特性.本文介绍了从离散样本过渡到连续曲线的平滑处理方法,陈述了线性平滑主元的基本框架——基函数空间下的多元统计.平滑曲线兼具幅度变异和相位变异,可通过配准分离两种变异.据此重点讨论了非线性平滑主元分析:既可采用混合数据形式,一并考察两种变异性;也可借助微分流形,在非欧氏空间描述相位变异.基于开源的步态数据集,给出了 3 组分析结果:未经配准的平滑主元分析;配准后的幅度变异分析和相位变异分析.最后,综述了平滑主元在生物信号处理中的典型应用.

关键词: 主元分析; 平滑; 函数型数据; 相位变异
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2014)03-0547-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.03.019

Smoothness in Principal Component Analysis: A Survey

XIANG Kui[1], ZHOU Shen-pei[1], LI Bing-nan[2]
(1. *School of Automation, Wuhan University of Technology, Wuhan, Hubei* 430070, *China*;
2. *School of Medical Engineering, Hefei University of Technology, Hefei, Anhui* 230009, *China*)

Abstract: Some of the sample observations, which seem like time series or discrete signals, are in fact smooth curves (functional data) corresponding to a latent continuous process. The smooth principal component analysis (PCA) focusing on functional data variation can fully characterize the dynamic features hidden in observations. The approaches smoothing discrete samples to continuous curves were introduced. The linear framework of smooth PCA was described as multivariate statistics in basis function spaces. The amplitude variation and phase variation embedded in smooth curves needed registration operations to separate themselves. The nonlinear framework of smooth PCA was discussed in two aspects: depicting two types of variation together with mixed data; depicting phase variation separately with differential manifolds in non-Euclidean space. Three groups of smooth PCA results were presented, which are raw gait data without registration, gait amplitude variation with registration and phase variation. Finally, the applications of smooth PCA in bio-signal processing were reviewed.

Key words: principal component analysis; smooth; functional data; phase variation

1 引言

主元分析(Principal Component Analysis, PCA)又称为主成分分析、主分量分析,是流传百年的降维算法,公认起源于 1901 年 Pearson 和 1933 年 Hotelling 的论文,Rao 在 1964 年的论文进一步解释和拓展了主元分析.Jolliffe 的专著《Principal Component Analysis》出版于 1986 年,2002 年再版[1],引用次数分别高达 7968 次和 6101 次.主元分析是多元统计学的核心内容,也深受工程技术人员青睐.在 Shlens[2] 的指南性文字中,充分展现了主元分析蕴含的数学严谨性,以及剖析复杂数据集的能力.

主元分析仍然有很多问题值得探讨.除了与高维相关的稀疏性以外[3],数据本身潜在的平滑性是另一个问题.近代,统计学似已形成相对完整的学科体系和知识框架,不过,计算机的普及为其提供了新的研究视野和方法改进[4].几乎所有统计方法都是以离散数据为出发点,在社会学、经济学研究中,这样的出发点基本是合适的.现代工业技术中,将连续信号经过采样并数字化,交由计算机处理,已渐成趋势.相形之下,传统的统计学并没有提供与之相符的解决方案.

收稿日期:2013-05-23;修回日期:2013-09-28;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61101022);湖北省自然科学基金(No.2012FFB05004);武汉理工大学自主创新研究基金(No.2012-II-017, No.2013-IV-063)

本质上,很多样本数据背后是一个连续过程.如果直接采用截面统计,则彻底忽视了样本之间的有序性;如果采用时间序列分析,虽可刻画有序性,但要求样本之间相互独立,这并不容易做到.如果用一个平滑的函数式来描述,既能体现系统本身的连续性,也符合能量有限的原理.此外,平滑处理可降低对采样密度和采样间隔的要求,且具有抑制噪声的作用.函数型数据分析(functional data analysis),正是契合这一需求而诞生的统计学分支.具体到函数型主元分析,可认为样本观测的平滑性要求主元和负荷向量也具有平滑性,是传统主元分析的平滑性延伸.因此,本文更多采用平滑主元的说法替代函数型主元.

2 引入平滑性

2.1 从样本点到曲线

函数型数据分析是关于随机曲线(或曲面)的统计学,每条曲线都是一个样本单元,对应了潜在随机过程的一次独立的平滑实现[5].曲线可以是随时间或空间连续变化的函数,现实中有很多关于函数型数据的实例:某地的气温曲线(以天为单位);行走的关节角度曲线(以步为单位);某人的手写体汉字;回转机械的速度曲线(以圈为单位).

早在 1960 年,法国数学家 Dieudonne 就阐述过函数型数据的思想."函数 f 是一个单独的对象,自身可以发生变化,通常可以看作一个大的函数空间中的一个点.实质上,经典和现代(数学)分析的一个主要区别在于:经典数学中,如果我们写下函数 $f(x)$,那么 f 是固定的,x 是变化的;现在 f 和 x 都是变化的;有时候,甚至是 x 固定 f 变化".1982 年,加拿大 McGill 大学的 Ramsay 发表了《When the data are functions》[6],进一步明确了数据为函数形式的问题.1992 年左右,Ramsay 和 Silverman 共同提出了"函数型数据分析"一词,随后于 1997 年出版了关于函数型数据分析的专著,2005 年再版[7].美国加州大学 Davis 分校的 Hans-Georg Müller、澳大利亚 Melbourne 大学的 Peter Hall 都是函数型数据分析领域非常活跃的研究者.有一些开源网站,提供与函数型数据分析相关的软件包,比如 Jim Ramsay 的 ftp 站点[8],Hans-Georg Müller 等人的网站 PACE[9].

从样本点过渡到曲线(函数型数据),平滑处理是首要步骤.平滑处理的方法很多,其中,B 样条最常用,它满足紧支(compact support)性,适合逼近某些非平滑的局部特征.另一种广泛使用的是傅里叶基,主要适用于周期信号,与 B 样条的特性恰好互补.核平滑也是一种重要的平滑方法,典型核函数包括高斯核、均匀核、二次核等等,它是最典型的局部加权函数.此外,多项式样条、三次样条、小波基等等,都可用于平滑处理.

虽然选用的基函数各不相同,但是平滑处理过程其实可以统一(核平滑稍有差异).假设一组相互独立的函数 ϕ_k 构成一个基函数系,可用 K 个基函数的线性组合逼近任意函数 x,形式如下:

$$x(t) = \sum_{k=1}^{K} c_k \boldsymbol{\phi}_k(t) \tag{1}$$

其中,c_k 为加权系数.也可以写成更加紧凑的矩阵形式:

$$x = \boldsymbol{C}^{\mathrm{T}}\boldsymbol{\Phi} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{C} \tag{2}$$

其中,$\boldsymbol{C}$ 和 $\boldsymbol{\Phi}$ 分别对应权系数和基函数向量.

平滑处理的效果,取决于基函数的选择,以及 K 值的大小.基函数系的特征与待展开函数 x 接近,可用较少的基函数获得满意的平滑效果,这种定性决策多依赖于经验.平滑效果往往与细节逼近相互冲突.K 值太小,细节损失过多,于后续分析不利.K 值增大,细节逼近效果自然好,但随之引入的高频成分也多,求导效果变差.因此,K 值不宜随意设定,可借助一个优化过程获得.下面从系数矩阵 $\boldsymbol{C}$ 的估计着手,讨论平滑效果的控制方法.

已知离散采样序列 $y_j, j = 1, \cdots, n$,系数 c 的估计可写成如下优化形式:

$$\arg\min_{c} = \sum_{j=1}^{n} \left[y_j - \sum_{k=1}^{K} c_k \boldsymbol{\phi}_k(t_j) \right]^2 \tag{3}$$

略去中间过程,可得系数矩阵 $\boldsymbol{C}$ 的最小二乘估计结果为:

$$\hat{\boldsymbol{C}} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y} \tag{4}$$

其中,$\boldsymbol{y}$ 是采样序列构成的矩阵.式(3)与普通最小二乘有细微区别,主要是 K 值待定.根据回归分析理论,均方误差包括方差(variance)和偏差(bias)两项.K 值选择其实也是偏差和方差的平衡问题,可采用变量选择方法来解决,也可采用罚函数方法.

Ramsay 和 Silverman[7] 设计了一种粗糙度(roughness)罚,专门处理 K 值相关的平滑问题.粗糙度罚定义为:

$$\begin{aligned} P_m(x) &= \int [\mathrm{d}^{(m)}x(s)]^2 \mathrm{d}s \\ &= \int [\mathrm{d}^{(m)}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{\phi}(s)]^2 \mathrm{d}s \\ &= \boldsymbol{C}^{\mathrm{T}}\left[\int \mathrm{d}^{(m)}\boldsymbol{\phi}(s)\mathrm{d}^{(m)}\boldsymbol{\phi}^{\mathrm{T}}(s)\mathrm{d}s\right]\boldsymbol{C} \end{aligned} \tag{5}$$

其中,$\mathrm{d}^{(m)}x$ 表示求 x 的 m 阶导数,可理解为曲率 $[\mathrm{d}^{(2)}x]^2$ 的推广形式.不难看出,x 的平滑度越低,粗糙度 P_m 越大.据此可将式(3)改写成如下形式:

$$\arg\min_{C} = (\boldsymbol{y} - \boldsymbol{\Phi C})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\Phi C}) + \rho P_2(x) \tag{6}$$

其中,ρ 是粗糙度罚参数,ρ 越大平滑程度越高.ρ 通常选一个非常小的正数,最优 ρ 值的选取,可借助交叉验证(cross-validation)或者广义交叉验证[10](generalized cross-validation)方法.令:

$$\boldsymbol{R}=\int \mathrm{d}^{(m)}\boldsymbol{\phi}(s)\mathrm{d}^{(m)}\boldsymbol{\phi}^{\mathrm{T}}(s)\mathrm{d}s \tag{7}$$

则 $P_m(x)=\boldsymbol{C}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{C}$.可得式(6)的解:

$$\hat{\boldsymbol{C}}=(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\rho\boldsymbol{R})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y} \tag{8}$$

某种意义上,式(8)可以看作一种岭回归形式.

上述讨论限于一般情况,还有一些特殊的函数型数据需要特殊处理.比如:河流的径流量,所有观测数据严格非负;人体身高的测量数据,非负,且单调递增(通常情况).上述基函数均不能确保取值非负或单调,Ramsay[11]提出用指数基解决这一问题.另外,小波很少用作平滑基函数,主要是因为小波基往往没有解析的导数形式.最近,Pigoli 和 Sangalli[12]提出了数值求导方法,可部分克服上述困难,并在多导联心电信号的平滑和求导中,成功使用小波基.

2.2　平滑主元的线性框架

在讨论平滑主元之前,首先阐述多元统计中一般性主元分析方法.假设 $\boldsymbol{x}$ 是由 p 维随机向量组成的矩阵,希望寻找一个线性组合 $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{x}$,具有最大方差.然后,在余下的方向上(与 $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{x}$ 正交),寻找另一个具有最大方差的线性组合 $\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{x}$.依次类推,选择的前 m 个组合

$$\sum_{i=1}^{m}\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{x} \tag{9}$$

可以代表 $\boldsymbol{x}$ 的方差结构,称之为主元.

寻找主元的过程,常常跟矩阵的特征分析紧密联系在一起.已知随机向量 $\boldsymbol{x}$ 的协方差矩阵$\boldsymbol{\Sigma}$,可以证明:主元 $f_k=\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{x}$ 中的向量$\boldsymbol{\alpha}_k$ 其实是矩阵 $\boldsymbol{\Sigma}$ 第 k 个最大特征值λ_k 对应的特征向量.$\boldsymbol{\alpha}_k$ 通常归一化为单位长度,主元 f_k 的方差为λ_k.实际分析过程中,常用样本协方差替代 $\boldsymbol{\Sigma}$.如果维数 p 过大,样本观测量 n 难以保证远大于p,则样本协方差与真实的协方差矩阵 $\boldsymbol{\Sigma}$ 相去甚远,由此引发了稀疏主元的问题.如果样本观测中的 p 维变量是有序的,或者是直接源于 p 个时间点的采样,则对应了本文讨论的平滑主元问题.

早在 1982 年,Dauxois 等人[13]就注意到了采样数据的主元分析问题,它的本质不是数据的分析,而是过程的分析.如果延续多元统计的思路,会有两方面的困难:无限维的问题;过程本身不能视作标量.1991 年,Rice 和 Silverman[14]在分析曲线样本的方差结构时,正式采用了平滑主元的方法.这一方法可以从式(1)演变而来:将无限维的函数曲线投影到基函数空间,在基函数空间中用多元统计的方法展开分析,将分析结果重新映射回到曲线所在的空间(Hilbert 空间).不难看出,这个基本框架的结构是线性的.至于其中出现的某些非线性问题,将在下一节讨论.

假设 x 是函数型数据样本(曲线),对应的平滑主元定义如下:

$$\begin{aligned} f&=\int\alpha x \\ &=\int\alpha(s)x(s)\,\mathrm{d}s \end{aligned}\quad \text{s.t.}\int\alpha(s)^2\mathrm{d}s=1 \tag{10}$$

协方差函数定义为:

$$v(s,t)=\frac{1}{N}\sum_{i=1}^{N}x_i(s)x_i(t) \tag{11}$$

则

$$\boldsymbol{V}\alpha=\lambda\alpha \tag{12}$$

其中,α 不再是特征向量,而是特征函数,λ 仍然是特征值.

上述平滑主元的求解,在函数型数据分析中已有许多研究[15],在此重点介绍基函数方法.根据式(2),式(11)可写成如下形式:

$$v(s,t)=\frac{1}{N}\boldsymbol{\Phi}(s)^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\Phi}(t) \tag{13}$$

假设特征函数 α 也可以写成基函数展开的形式$\alpha(s)=\boldsymbol{\Phi}^{\mathrm{T}}(s)\boldsymbol{B}$,则式(12)可以转化为:

$$\lambda\boldsymbol{\Phi}^{\mathrm{T}}(s)\boldsymbol{B}=\frac{1}{N}\int\boldsymbol{\Phi}^{\mathrm{T}}(s)\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{\mathrm{T}}(t)\boldsymbol{B}\mathrm{d}t \tag{14}$$

令矩阵 $\boldsymbol{W}=\int\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$,式(14)可简化为:

$$\lambda\boldsymbol{B}=\frac{1}{N}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{W}\boldsymbol{B} \tag{15}$$

如果选择的基函数 $\boldsymbol{\Phi}$ 是正交规范基,则 $\boldsymbol{W}$ 退化为单位矩阵,式(15)看上去类似一个普通的主元分析表达式,只需要对矩阵 $\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}/N$ 进行特征分析.如前所述,在基函数空间中,采用一般的多元统计方法,可实现平滑主元分析.不过,有一个细节不容忽视:在多元统计的特征分析中,p 维样本最多有 p 个不同特征值;函数型数据名义上是无限维,但是依据式(15)并不能获得无限多个不同特征值.其实,协方差矩阵 $\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}/N$ 的秩与样本数量有关,不同特征值的最大数目为 $N-1$.

虽然由式(15)得到的主元本身是平滑的,但平滑程度依赖于样本的平滑.Silverman[16]提出了一个实用的优化形式,使主元本身的平滑亦可直接控制.其基本形式如下:

$$\arg\max_{\alpha}=\frac{\mathrm{var}\left(\int\alpha x\right)}{\|\alpha\|^2+\rho P(\alpha)} \tag{16}$$

其中,$P(\alpha)$是类似式(5)的粗糙度罚,ρ 是平滑参数.美中不足的是,将负荷向量的范数 $\|\alpha\|$ 置于分母中,虽然有约束范数大小的作用,但不能确保 α 一定为单位长度.式(16)可获得唯一的主元分析结果,但主元的顺序有可能发生变化.Aguilera 和 Aguilera-Morillo[17]提出用 P 样条罚代替以上的粗糙度罚,即用 B 样条相邻系数 d 阶差分的平方作为罚函数,减轻了平滑主元分析的计算负担.

有关平滑主元分析的基本框架,Ocaña 等人[18]描述

了详细的算法流程和计算细节,Qi 和 Zhao[19]为上述框架提供了理论证明. Li 等人[20]基于边际(marginal)模型的 Bayesian 信息准则,专门研究主元个数的选择标准,适用于密集或稀疏的曲线样本集. Sawant 等人[21]根据式(15)中的系数矩阵 $\boldsymbol{C}$,辨识函数型数据中的离群点(outlier),提高主元分析的鲁棒性. 除了上述经典框架,Huang 等人[22]提出了秩 1 逼近方法,同样可用罚函数控制主元的平滑度.

为了更形象地说明平滑主元,下面给出一个关于步态数据分析的实例,更多的实例可参阅文献[23]. 该数据集源自美国加州 San Diego 的儿童医院,包括 39 名儿童的髋关节与膝关节角度采样,每人均保留一个完整步态的采样,且只考虑关节角在矢状面的投影[24]. 本文只展示了膝关节数据及其分析结果. 每个步态周期自下肢脚后跟触地开始,为描述方便,将步态周期归一化到[0,1]区间. 原始数据经平滑后得到的曲线样本如图 1(a)所示,总共包括 39 条不同颜色的曲线. 根据本节提供的基本框架,采用文献[13]提供的程序包,计算得到平滑主元,其中的一阶与二阶主元如图 1(b)、(c)所示.

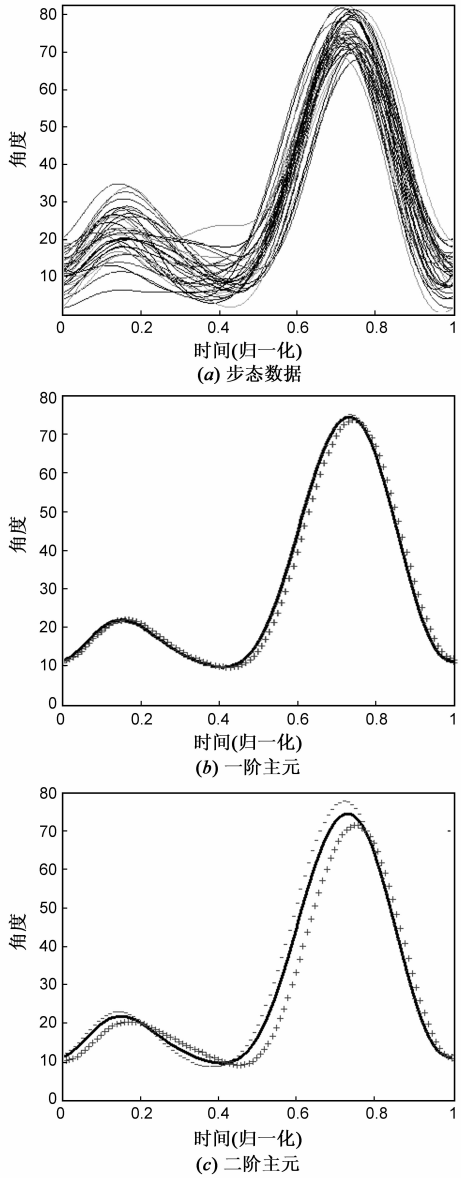


图1 步态数据的平滑主元分析

为便于观察,图 1(b)、(c)中主元的绘制方法为:蓝色实线为均值曲线,红色“+”和“-”组成曲线 s_+ 和 s_-,定义为

$$\begin{cases} s_+ = u(t) + \sqrt{\lambda}\alpha \\ s_- = u(t) - \sqrt{\lambda}\alpha \end{cases} \tag{17}$$

其中,$\sqrt{\lambda}$ 代表主元的均方差,即变异的大小;α 为负荷向量,即变异的方向. 从 s_+ 和 s_- 包围的区域可以看出变异发生的位置和大小. 需要指出的是,为了便于对主元的解释,上述分析过程中均采用了 varimax 旋转方法[1].

3 平滑带来的变化

3.1 曲线配准

对于曲线样本 $x(t)$,它不是一个数值,而是一个实体(entity),具有结构属性. 仅就样本均值而言,多元统计样本的均值是一个向量;函数型数据的均值是一条曲线. 在曲线样本集合上定义的截面统计不再是凸运算,运算结果不属于集合元素. 一个简单例子:假设曲线样本为同频率、不同幅度、不同相位的单周期正弦信号,如果采用截面统计求均值的方法,结果将不再是正弦信号. 直观解释是:曲线样本不仅有幅度变异,还有相位变异. 从物理角度考察,相位变异源于系统内部时钟与采样时钟的差别. 要想获得均值曲线的估计,需要实施配准(registration)处理,分离幅度变异与相位变异.

基本配准方法有 3 种:平移(shift)配准、地标(landmark)配准、连续配准[25]. 平移配准是其中最简单的一种,在曲线 x_i 中添加一个平移参数 δ_i,得到 $x_i(t+\delta_i)$,参数 δ_i 可通过以下方法确定:

$$\arg\min_{\delta} = \sum_{i=1}^{N} \int [x_i(t+\delta_i) - \hat{u}(t)]^2 \mathrm{d}s \tag{18}$$

其中,$\hat{u}(t)$ 表示样本集合 $\{x_i\}$ 对应的均值曲线. 地标配准则更多利用了曲线的特征点,比如极大值、极小值、过零点等等. 可以这样认为,同种地标点对应了系统内部时钟的同一时刻,配准时优先保证地标点相互对齐. 一个典型例子,心电信号的 QRS 波峰对应了心室的除极阶段,它是心电信号排列的主要基准. 准确识别出地标点并不容易,除了曲线本身以外,经常借助其导数形式,或是借助小波分析定位[26],以提高配准质量.

连续配准主要是借鉴了动态时间规整(dynamic

time warping)算法,动态时间规整在形状分析中使用较多.假设规整函数为 $h(t)$,通常设定为严格单调递增,这与时间不可反转有关.另外,$h(t)$本身最好是平滑的,符合函数型数据分析的宗旨.因此,经常将 $h(t)$写成如下参数形式[25]:

$$h(t) = c_0 + c_1 \int_0^t \exp T(u)\mathrm{d}u \tag{19}$$

其中,c_0和 c_1分别是平移和伸缩量,可使 $h(t)$与 t 在整个区间上两端对齐.指数形式保证了 $h(t)$的单调递增特性,也使得它连续可导.若 $T(u)=0$,系统时钟与采样时钟完全一致;若 $T(u)>0$,则 $h(t)>t$,说明系统时钟较慢,$h(t)$起加速作用;$T(u)<0$ 的情况可类推得知.在估计 $h(t)$时,可采用上文提到的粗糙度罚,控制 $h(t)$的平滑程度,如下所示:

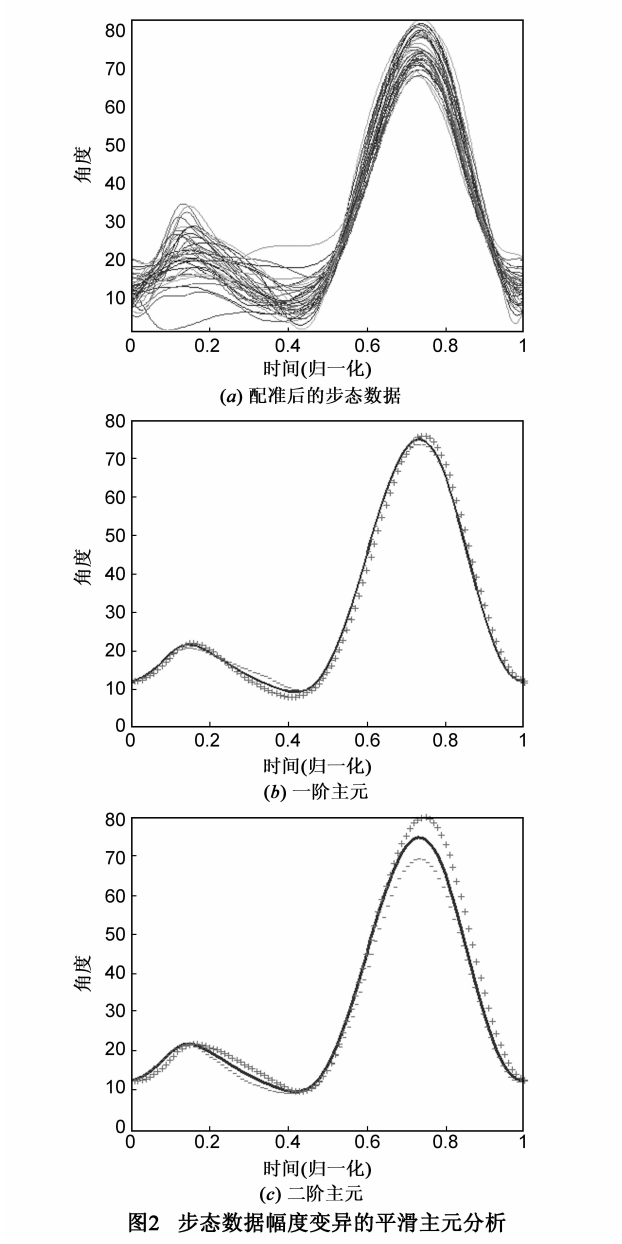
$$\arg\min_h = \sum_{i=1}^{N}\int_T [x_i(h_i(t)) - \hat{u}(t)]^2\mathrm{d}t + \rho\int_0^1 [h_i^{(p)}(t)]^2\mathrm{d}t \tag{20}$$

通过设置参数 ρ,即可调整 $h(t)$的平滑度.

在式(18)和(20)中,存在一个共性问题:$\hat{u}(t)$并非已知,其实 $\hat{u}(t)$本该是通过配准而获得的估计量.为此,必须采用一个迭代算法方能真正求解式(18)和(20).一般情况下,可采用截面均值作为 $\hat{u}(t)$的初值,完成一轮配准后,重新计算曲线样本的截面均值,更新 $\hat{u}(t)$,继续下一轮配准.当 $\hat{u}(t)$不再有明显变化时,即可停止迭代,多数情况下,该迭代过程经过 1~2 轮就会收敛.函数型数据分析领域,还有很多关于配准方法的研究,篇幅限制,此处不一一展开叙述.

依然采用 2.2 节的例子,增加配准步骤,获得如图 2 可类比的结果.采用连续配准方法,对图 1(*a*)中所有曲线样本配准,结果如图 2(*a*).配准结果存在一个不良倾向,配准过程过度依赖于幅度较大的曲线段.图 2(*a*)中,最大屈膝位置(关节角 70°~80°)的配准效果最好,而另一个较小屈膝位置(关节角 20°~30°)的配准效果明显差很多.究其原因,虽然配准的目标是相位对齐,但参考的依据仍然是幅度变化.

图 2 所示的主元分析结果,是剔除了相位变异特性的幅度变异分析.从图 2(*b*)、(*c*)可以看出,变异明确集中在两个时段:一是从第一个波峰下降的位置,二是最大波峰位置.步态信号的起点为脚跟触地,不难推断上述变异对应的生理学解释.第一个变异时段对应了前脚着地以后,人体重心前移,膝关节开始承重的过程;第二个变异时段对应了后脚悬空,作为摆动腿自由移动的过程.在配准之前实施主元分析,前 4 阶主元可以描述 88% 的变异性,配准之后,这一指标上升为 95%.从中可以得到一个结论,幅度变异和相位变异会相互串扰,可将它们先行分离,然后做独立表征.



(*a*) 配准后的步态数据
(*b*) 一阶主元
(*c*) 二阶主元

图2　步态数据幅度变异的平滑主元分析

3.2　非线性平滑主元

分离幅度和相位变异成分,可以更好地量化幅度变异,同时有助于窥寻相位变异中隐藏的信息.比如,Slaetsf 等人[27]在研究 Berkeley 生长数据时,根据幅度和相位信息进行聚类,发现了一类特殊个体:在经历了婴儿期的快速生长期之后,生长速度呈现明显下降.相比幅度变异,相位变异的量化难度很大,因为相位信息不适合在欧氏空间中刻画.两种变异的综合量化分析则更为复杂.

因为幅度和相位总是成对出现,可以考虑把它们组成混合数据(mixed data)的形式:$(\tilde{x},\theta)$,$\tilde{x}$ 代表配准后的幅度,θ 代表相位.在此基础上,合并考虑两者的主元分析问题.

$$\begin{pmatrix} \tilde{x} \\ \theta \end{pmatrix} = \sum_j f_j \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \tag{21}$$

其中，(α_j, β_j) 是第 j 阶主元，f_j 为主元得分，$f_j = \int (\tilde{x}\alpha_j + \theta\beta_j)$.把$(\tilde{x}, \theta)$记为 z，通过定义 z 的内积形式，可将混合数据的主元分析重新纳入到 2.2 提供的线性框架下.潜在的障碍是，如何平衡幅度和相位的变异性，因为两者的量纲和尺度都不具备可比性.为此，专门增加一个常数 c，将复杂的非线性问题线性化.令 $z = (\tilde{x}, c\theta)$来表达相位变异的权重，并将 z 的内积定义如下：

$$\langle z_1, z_2 \rangle = \int (\tilde{x}_1 \tilde{x}_2 + c^2\theta_1\theta_2) \tag{22}$$

Kneip 和 Ramsay[28]根据上述思路，发展了集配准和主元分析于一体的方法.将函数型数据的整体变异 v_t 分解为配准后的幅度变异 v_a 和相位变异 v_p：

$$\begin{cases} v_t = \mathrm{E}\int [x(t) - \mu(t)]^2 \mathrm{d}t \\ v_a = b\mathrm{E}\int [\tilde{x}(t) - \tilde{\mu}(t)]^2 \mathrm{d}t \\ v_p = b\int \tilde{\mu}^2(t)\,\mathrm{d}t - \int \mu^2(t)\,\mathrm{d}t \end{cases} \tag{23}$$

其中 $\tilde{x}(t)$和 $\tilde{\mu}(t)$是配准后的样本及其均值，系数 b 用来度量规整函数 h 与配准结果 $\tilde{x}(t)$之间的独立性.两种变异各自所占的份额，依赖于规整函数 $h(t)$的定义.重复一个配准加特征分析的迭代过程，直至 v_p/v_t的值不再增大.可以看出，迭代的终点在于相位变异所占比重最大化.

用 2.2 节的框架分析 Berkeley 生长数据，绘制一阶主元 v.s.二阶主元的曲线图，会呈现一个奇怪的“马蹄”形，说明不同主元之间存在非线性相依性.要想把这种“马蹄”形展开到一个线性结构上，就需要借助非线性降维方法，比如流形学习.

Chen 和 Müller[29]提出了函数型流形元分析(functional manifold component analysis)方法，包括了流形均值、流形模态和流形元，可以看作是平滑主元的非欧氏版本.相比前文提到的均值曲线，流形均值的定义方式更加全面.求解方法上，流形元分析主要借鉴流形学习中广泛使用的等距映射(isometric mapping)算法，解决测地线距离(geodesic distance)的经验性定义问题.

假设 M 是 d 维流形，φ 是一个双射，$\varphi: R^d \to M \subset L_2$，$\psi = \varphi^{-1}$是其表达映射，则有

$$\mu = \mathrm{E}[\psi(x)], \mu^M = \psi^{-1}(\mu) \tag{24}$$

其中，μ 是 d 维表达空间的均值，μ^M是 L_2范数空间的流形均值.若 M 是等距流形，对于所有等距的表达映射，μ^M都唯一确定，定义如下：

$$\mu^M = \arg\min_{x \in M} \mathrm{E}[d_g^2(z, x)] \tag{25}$$

其中，d_g 为测地线距离.借助多维尺度分析可得到 ψ 的估计，并通过分段衔接的办法获得 d_g 的估计.然后可仿照前述的线性框架，在流形上定义非线性的主元，即流形元.

Tucher 等人[30]提出的均方速率(square root slope)函数 $q: [0,1] \to R$，定义如下：

$$q(t) = \mathrm{sign}(x'(t))\sqrt{|x'(t)|} \tag{26}$$

可将规整函数 $h(t)$转化到 Hilbert 空间中分析，更好地刻画相位变异.对于任意 $q \in L_2$ 以及 $t \in [0,1]$，皆可获得一个函数 x，

$$x(t) = x(0) + \int_0^t q(s)\,|q(s)|\,\mathrm{d}s \tag{27}$$

$x \to (q, x(0))$是一个双射(bijection)，对 x 做时间规整 $x \oplus \gamma$ 可转化为：

$$(q, \gamma)(t) = q(\gamma(t))\sqrt{\gamma'(t)} \tag{28}$$

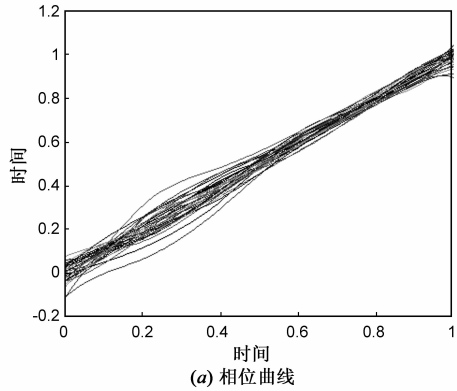
可以先对任意曲线的均方速率函数进行配准，然后把配准结果映射返回原空间.借助均方速率函数，可把复杂的几何空间简化为一个单位球面，任意两个规整函数 γ_1、γ_2 之间距离其实就是单位球面上的弧长.据此可以定义集合$\{\gamma_i\}$的均值，然后将其投影到切平面上做主元分析[30].此外，Lu 和 Marron[31]针对规整函数的变异分析，提出主嵌套球(principal nested spheres)的方法，比切平面上的主元分析更加简单有效.

在 3.1 节步态数据的配准之后，可以分离出相位变异，其规整函数 $h(t)$如图 3(a)所示.对 $h(t) - t$ 做主元分析，其一阶和二阶主元如图 3(b)、(c).从图中可以看出，相位变异主要集中在两个点，其一是身体重心从后往前转移的时段，其二是后脚跟触地的刹那.值得一提的是，在引入平滑主元之前，类似的分析未曾被关注过.关于幅度与相位变异的混合分析，以及基于流形的分析，相关方法尚未成熟，因此不做实例展示.

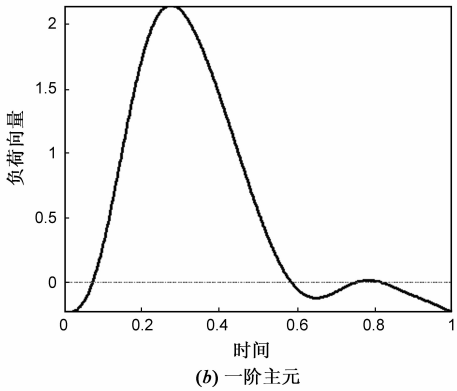
4 在生物信号中的应用

Ullah 和 Finch[32]利用 11 个文献数据库，检索了 1995～2010 年间有关函数型数据分析的英文文章.排除纯方法论述之后，共获得了 84 篇有关应用的文章，其中 75% 的文章发表于 2005 年以后，可见该领域的应用研究呈明显上升趋势.在 84 篇文章中，有关生物医学的研究占 21.4%，而使用平滑主元分析方法的高达 60.7%.因此本节专门以生物信号处理为背景，介绍平滑主元的实际应用.

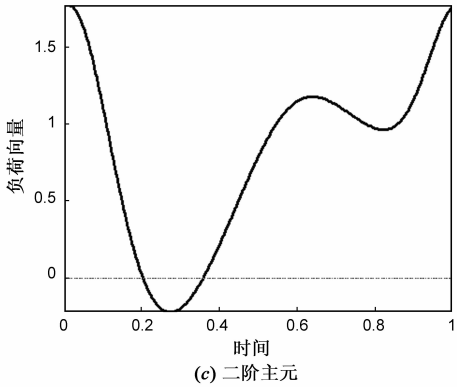
生物信号大致可分为生物电、生物力和生物影像共 3 种类型，就已经掌握的文献来看，有关生物力信号的应用最多，生物影像也有一些，而生物电则比较少见.无论是生物电还是生物力信号，其实都是函数型时



(*a*) 相位曲线



(*b*) 一阶主元



(*c*) 二阶主元

图3　步态数据相位变异的平滑主元分析

间序列. Kurtek 等人[33]专门讨论了生物信号的分割、排列与统计分析问题. 基于生物信号的周期性,将其分割为多个子序列,每个子序列对应一个循环(cycle),可视为一个样本. 运用配准方法,将各个子序列排列整齐. 采用均方速率函数方法,分析幅度变异和相位变异. 在基于步态和心电的疾病诊断中,上述方法的识别效果要明显优于传统基于横截面统计的方法. Coffey 等人[34]采用共同(common)平滑主元方法分析人体运动数据,以便更好地捕捉人体运动模式的转变. 对人体运动的长期监测获得数据集,可分组进行主元分析,或是整体进行主元分析. 前者获得的主元空间基坐标系可能并不一致,而后者难以捕捉到细微的组间变异. 在共通平滑主元中,构成每一组数据变异的是共同的因子,但不同因子对不同组的变异贡献大小不一样. 这种方法可以揭示伤病前后人体运动控制模式的改变.

Ryan 等人[35]用平滑主元研究了人的纵向弹跳运动,发现第三阶主元最为重要,它对应了人的牵张-收缩循环(stretch-shortening cycle)能力,尽管第三阶主元只表达了 7.9% 的变异性. Epifanio 等人[36]用平滑主元研究了坐下—起立过程中的膝关节弯曲角度和弯曲力矩,发现平滑主元分析有助于更好地区分关节炎患者和健康个体. Samadani 等人[37]针对人体运动数据的降维与分类问题,对比研究了 Fisher 投影、平滑主元与等距映射算法,发现平滑主元在分类精度和计算资源两项指标上具有优势. 林辉杰等人[38]用平滑主元研究了掷铁饼动作的运动数据,探讨了其中的运动协调量化问题,为该动作运动协调特征及运动成绩影响因素的探索提供了方法指导.

Tian[39]综述了函数型数据分析在影像处理中的应用,包括 3 个方面:降维(或特征提取)、空间(形状)分类、脑磁图反问题,同时指出 3 个方面的挑战:计算代价高、非高斯噪声、空间相关性. Viviani 等人[40]把平滑主元应用于功能核磁共振影像的处理中,用平滑函数取代像素点序列,可以更好地从噪声污染中恢复影像. O′Connor等人[41]把平滑主元用于处理医学影像的某一特定区域,获得某一个像素值分布的主密度(principal densities),用于解释某些重要特征(如肿瘤形状)的变异模态,对治疗效果做出正式评估. Jiang 等人[42]基于平滑主元,提出了一种累加的非参数随机效应模型,用于分析正电子发射断层成像(positron emission tomography),可降低后续参数估计的最小均方误差. 此外, Ormoneit 等人[43]将平滑主元用于运动捕捉数据的分析,三维的关节角序列可分割成运动循环的子序列,平滑主元有助于处理原始数据的噪声和丢点问题.

5　结论

平滑性约束广泛存在于工程系统中,发展平滑主元对于观测信号的分析处理非常必要. 从统计的视角,平滑主元可以认为是对连续过程的降维,有利于深入分析动态系统特性. 尤其在计算机的时代,平滑主元是衔接连续动态分析与数字信号处理的重要工具. 由于主元分析作为一种降维工具的基础性作用,使得平滑主元很容易延伸到回归分析、模式识别等研究方向中去,限于篇幅,这部分内容从略.

引入平滑处理,为分离幅度和相位变异创造了条件. 一则使幅度变异的量化更加纯粹,二则为相位变异的独立表征创造了条件. 在此之前,相位变异经常作为有害信息而丢弃. 不过,相位变异量化存在的困难较多. 有关信号的幅度和相位定义原本具有含混性,通过

配准来分离幅度和相位实非易事.相位变异并非存在于欧氏空间,简单的线性主元分析容易导致曲解.已经有研究开始引入流形学习之类的非线性降维方法,这也为平滑主元的研究注入了活力,后续发展更加值得期待.

参考文献

[1] I T Jolliffe. Principal Component Analysis (2nd edition)[M]. New York: Springer-Verlag, 2002.

[2] J Shlens. A Tutorial on Principal Component Analysis[R]. USA: Systems Neurobiology Laboratory, University of California at San Diego, 2005.

[3] 向馗,李炳南.主元分析中的稀疏性[J].电子学报,2012,40(12):2525 – 2532.
K Xiang, B Li. Sparsity in principal component analysis: A survey[J]. Acta Electronica Sinica, 2012, 40(12): 2525 – 2532. (in Chinese)

[4] 米子川,赵丽琴.函数型数据分析的研究进展和技术框架[J].统计与信息论坛,2012,27(6):13 – 20.
Z Mi, LZhao. The research development and technical framework of function data analysis[J]. Statistics & Information Forum, 2012, 27(6): 13 – 20. (in Chinese)

[5] H G Müller. Functional data analysis[A]. M Lovric. International Encyclopedia of Statistical Science[C]. Heidelberg: Springer Science Business Media, 2010.

[6] J O Ramsay. When the data are functions[J]. Psychometrika, 1982, 47(4): 379 – 396.

[7] J O Ramsay, B W Silverman. Functional Data Analysis (2nd Edition)[M]. New York: Springer-Vetlag, 2005.

[8] J O Ramsay. Functional Data Analysis[EB/OL]. http://www.psych.mcgill.ca/misc/fda/index.html, 2013 – 5 – 30.

[9] H G Müller. PACE: Principal Component by Conditional Expectation[EB/OL]. http://anson.ucdavis.edu/~ntyang/PACE/, 2012 – 12 – 17.

[10] T Hastie, R Tibshirani, J Friedman. The Elements of Statistical Learning Data Mining, Inference, and Prediction (2nd edition)[M]. New York: Springer-Verlag, 2002.

[11] J O Ramsay. Estimating smooth monotone functions[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1998, 60(2): 365 – 375.

[12] D Pigoli, L M Sangalli. Wavelets in functional data analysis: Estimation of multidimensional curves and their derivatives[J]. Computational Statistics & Data Analysis, 2012, 56(6): 1482 – 1498.

[13] J Dauxois, A Pousse, Y Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference[J]. Journal of Multivariate Analysis, 1982, 12(1): 136 – 154.

[14] J A Rice, BW Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1991, 53(1): 233 – 243.

[15] H L Shang. Visualizing and Forecasting Functional Time Series[D]. Melbourne: Monash University, 2010.

[16] B W Silverman. Smoothed functional principal components analysis by choice of norm[J]. The Annals of Statistics, 1996, 24(1): 1 – 24.

[17] A M Aguilera, M C Aguilera-Morillo. Penalized PCA approaches for B-spline expansions of smooth functional data[J]. Applied Mathematics and Computation, 2013, 219(14): 7805 – 7819.

[18] F A Ocaña, A M Aguilera, M Escabias. Computational considerations in functional principal component analysis[J]. Computational Statistics, 2007, 22(3): 449 – 465.

[19] X Qi, H Zhao. Some theoretical properties of Silverman′s method for smoothed functional principal component analysis[J]. Journal of Multivariate Analysis, 2011, 102(4): 741 – 767.

[20] Y Li, N Wang, R J Carroll. Selecting the number of principal components in functional data[J]. Journal of the American Statistical Association, 2013, 108(504): 1284 – 1294.

[21] P Sawant, N Billor, H Shin. Functional outlier detection with robust functional principal component analysis[J]. Computational Statistics, 2012, 27(1): 83 – 102.

[22] J Z Huang, H Shen, A Buja. Functional principal components analysis via penalized rank one approximation[J]. Electronic Journal of Statistics, 2008, 2: 678 – 695

[23] J O Ramsay, B W Silverman. Applied Functional Data Analysis: Methods and Case Studies[M]. New York: Springer, 2002.

[24] R A Olshen, E N Biden, M P Wyatt, et al. Gait analysis and the bootstrap[J]. Annals of Statistics, 1989, 17(4): 1419 – 1440.

[25] J O Ramsay, X Li. Curve registration[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1998, 60(2): 351 – 363.

[26] J Bigot. Landmark-based registration of curves via the continuous wavelet transform[J]. Journal of Computational and Graphical Statistics, 2006, 15(3): 542 – 564.

[27] L Slaets, G Claeskens, M Hubert. Phase and amplitude – based clustering for functional data[J]. Computational Statistics & Data Analysis, 2012, 56(7): 2360 – 2374.

[28] A Kneip, J O Ramsay. Combining registration and fitting for functional models[J]. Journal of the American Statistical Association, 2008, 103(483): 1155 – 1165.

[29] D Chen, H G Müller. Nonlinear manifold representations for functional data[J]. Annals of Statistics, 2012, 40(1): 1 – 29.

[30] J D Tucker, W Wu, A Srivastava. Generative models for functional data using phase and amplitude separation[J]. Computational Statistics & Data Analysis, 2013, 61(1): 50 – 66.

[31] X Lu, J S Marron. Principal nested spheres for time warped functional data analysis[J/OL]. Math ST arXiv: 1304.6789, April, 2013.

[32] S Ullah, C F Finch. Applications of functional data analysis: A systematic review[J]. BMC Medical Research Methodology, 2013, 13(1): 1 – 12.

[33] S Kurtek, W Wu, G E Christensen, et al. Segmentation, alignment and statistical analysis of biosignals with application to disease classification[J]. Journal of Applied Statistics, 2013, 40(6): 1270 – 1288.

[34] N Coffey, A J Harrison, O A Donoghue, et al. Common functional principal components analysis: A new approach to analyzing human movement data[J]. Human Movement Science, 2011, 30(6): 1144 – 1166.

[35] W Ryan, A Harrison, K Hayes. Functional data analysis of knee joint kinematics in the vertical jump[J]. Sports Biomechanics, 2006, 5(1): 121 – 138.

[36] I Epifanio, C Avila, A Page, et al. Analysis of multiple waveforms by means of functional principal component analysis: normal versus pathological patterns in sit-to-stand movement [J]. Medical & Biological Engineering & Computing, 2008, 46 (6): 551 – 561.

[37] A A Samadani, A Ghodsi, D Kulic. Discriminative functional analysis of human movements[J]. Pattern Recognition Letters, 2013, 34(15): 1829 – 1839.

[38] 林辉杰,严波涛,许崇高,等.基于函数型数据分析技术的运动协调量化方法应用研究[J].体育科学, 2012, 32 (9): 81 – 87.

H Lin, B Yan, C Xu, et al. Applied research on quantitative method of motor coordination basing on functional data analysis technique[J]. China Sport Science, 2012, 32(9): 81 – 87. (in Chinese)

[39] T S Tian. Functional data analysis in brain imaging studies[J/OL]. Frontiers in Psychology, 1: 35, Oct 8, 2010.

[40] R Viviani, G Grön, M Spitzer. Functional principal component analysis of fMRI data[J]. Human Brain Mapping, 2005, 24 (2): 109 – 129.

[41] E O′Connor, N Fieller, A Holmes, et al. Functional principal component analyses of biomedical images as outcome measures[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 2010, 59(1): 57 – 76.

[42] C R Jiang, J A D Aston, J L Wang. Smoothing dynamic positron emission tomography time courses using functional principal components[J]. NeuroImage, 2009, 47(1): 184 – 193.

[43] D Ormoneit, M J Black, T Hastie, et al. Representing cyclic human motion using functional analysis[J]. Image and Vision Computing, 2005, 23(14): 1264 – 1276.

作者简介

向　馗　男,1976 年 10 月出生于湖北省秭归县.现为武汉理工大学自动化学院副教授,从事生理信号处理和人 机协作方面的研究工作.
E-mail: xkarcher@126.com

周申培　女,1979 年 7 月出生于湖北省武汉市,现为武汉理工大学自动化学院副教授,从事模式识别和智能交通方面的研究工作.
E-mail: zhousp73@sina.com

李炳南(通讯作者)　男,1978 年 5 月出生于江苏省常州市.现为合肥工业大学医学工程学院教授、黄山青年学者.新加坡国立大学生物工程以及澳门大学电子工程双博士学位.在国内外发表学术论文 30 余篇.
E-mail: bingnan@live.com