文章编号:1671-9352(2026)04-0084-08 DOI:10.6040/j.issn.1671-9352.0.2024.438

基于深度与融合类信息的函数型数据重构方法

黄介武,陈星悦*,王淋杰,饶文康
(贵州民族大学数据科学与信息工程学院, 贵州 贵阳 550025)

摘要:针对部分观测函数型数据,提出一种基于深度与融合类信息的重构方法。运用基于深度的重构方法以及从 K 均值聚类中获取的样本曲线类间信息,在不同分类情形下对每条部分观测样本曲线进行重构。然后,利用自加权集成学习算法动态赋权,将各类别下的重构曲线融合,得到最终的重构曲线。数值模拟和实例分析表明:当样本中部分观测样本曲线占比较大时,所提方法在均方预测误差准则下优于基于深度的重构方法及正则化回归方法;而在部分观测样本曲线占比较小时,正则化回归方法表现更优。

关键词:部分观测函数型数据;函数型数据重构;数据深度;类信息
中图分类号:O212.1 **文献标志码**:A
引用格式:黄介武,陈星悦,王淋杰,等. 基于深度与融合类信息的函数型数据重构方法[J]. 山东大学学报(理学版),2026,61(4):84-91,101.

A depth-based and fusion class information reconstruction method for functional data

HUANG Jiewu, CHEN Xingyue*, WANG Linjie, RAO Wenkang
(School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, Guizhou, China)

Abstract: A depth-based and fusion class information reconstruction method is proposed for partially observed functional data. By applying the depth-based reconstruction method and the inter-class information of sample curves derived from K-means clustering, each partially observed sample curve is reconstructed under different classification scenarios. Then, with the weights dynamically assigned by the self-weighted ensemble learning algorithm, final reconstructed curves are obtained by combining the reconstructed curves of each class. Simulation studies and case analysis show that the proposed method outperforms the depth-based reconstruction method and the regularized regression method under the mean-square prediction error criterion when the proportion of partially observed sample curves in the sample is large. Conversely, the regularized regression method performs better when the proportion of partially observed sample curves is small.

Key words: partially observed functional data; reconstruction of functional data; data depth; class information

0 引言

函数型数据广泛应用于生物医学、经济学、社会科学等领域,具有高维(无限维)、高相关性等特征,通常以曲线、曲面、图像等形式呈现。用于处理和分析函数型数据的统计工具称为函数型数据分析[1]。函数型数据分析通常假设所有函数都在同一定义域上被完全、密集或稀疏地观测到,然而在实际应用中,往往存在因设备故障、参与者不配合或实际操作不当等原因,导致函数在定义域的某些子域上未被观测到而缺失的情况,即存在部分观测函数型数据(partially observed functional data, POFD)的情况,这种缺失在智能交通、医学和经济学等领域尤为常见。数据的缺失增加了统计分析的复杂性和难度,同时使得一些传统的函数型数据分析方法不再适用。因此,如何科学有效地处理部分观测函数型数据是亟待解决的问题。

收稿日期:2024-12-18; **网络出版时间**:2025-07-16
基金项目:贵州省教育厅自然科学研究项目(黔教技[2023]012 号,黔教技[2023]061 号)
第一作者:黄介武(1977—),男,教授,硕士生导师,博士,研究方向为统计模型与统计计算. E-mail:846221886@qq.com
*通信作者:陈星悦(2001—),女,硕士研究生,研究方向为统计模型与统计计算. E-mail:2418231562@qq.com

目前,已有许多学者对部分观测函数型数据的统计分析方法进行了研究。James 等[2]基于混合效应模型对部分观测函数型数据进行主成分分析。Chiou 等[3]将条件期望引入函数型主成分分析,以估计交通流量数据的缺失部分,并且基于函数型主成分分析方法对部分观测函数型数据中的异常值进行了检测。Kraus[4]给出部分观测函数型数据的均值函数和协方差函数的估计方法,同时基于岭正则化方法提出一种用于部分观测函数型数据缺失部分重构的线性算子。Kraus 等[5]证明岭重构方法[4]的渐近最优性。Kneip 等[6]为了避免观测部分和重构部分之间的人为跳跃,基于局部线性核提出一种新的重构算子,进而生成光滑的重构结果,并证明该重构算子的渐近最优性。Li 等[7]在考虑样本曲线之间相关性的基础上,结合 K—L 展开、子空间投影技术以及函数主成分分析方法,提出一种“聚类+插补”一体化的方法,用于估算交通流量轨迹中的未观测部分。高海燕等[8]通过聚类分析引入样本类信息,并利用类内样本数据的相关性进行缺失值插补,最后通过集成学习策略将样本数据的多个插补结果融合,提出一种融合类信息的函数型矩阵填充方法,用于处理智能交通系统中的交通流量数据缺失问题。杨玉杰等[9]基于部分观测函数型数据,探讨部分函数线性分位数回归模型的参数估计方法,并通过实例验证其方法的有效性。Liebl 等[10]针对非完全随机缺失的情况,利用微积分基本定理,提出部分观测函数型数据的均值函数和协方差函数的估计方法。以上重构方法大多依赖于合适的协方差估计,然而当样本数据来自多个总体或完整数据稀缺时,往往难以保证协方差估计的精度,从而限制重构方法的有效性。为此,Elías 等[11]提出一种基于深度的重构方法。该方法通过最大化拟重构曲线的深度,构建一组能表征拟重构曲线形状和大小的样本曲线子集(包络),并将这些曲线投影到缺失区域以实现重构。基于深度的重构方法有效降低协方差估计选取对重构精度的影响,显著提升部分观测样本曲线占比较大时的重构精度。然而,该方法也存在没有充分考虑样本曲线间的相关性、包络冗余引发过拟合风险等不足。

针对上述问题,本文以基于深度的重构方法为基础,通过 K 均值聚类挖掘样本曲线之间的时空相关性及潜在变化模式,再利用样本曲线之间的相关性重构曲线,并在整体学习框架下集成融合拟重构曲线的多个重构结果,构建一种基于深度与融合类信息(depth-based and fusion class information, C-Depth-based)的部分观测函数型数据重构方法,以进一步提高重构的精度。

1　C-Depth-based 重构方法

1.1　部分观测函数型数据

设 $X(t)$ 是平方可积的可分 Hilbert 空间 $L^2[a,b]$ 上的连续随机函数,$X_i(t)$, $i=1,2,\cdots,n$ 是 $X(t)$ 的 n 个独立样本,其中 t 表示时间或其他变量。不失一般性,取 $[a,b]=[0,1]$。当 $X_i(t)$ 在 $[0,1]$ 的非空紧子集 O_i 被观测到,而在 $M_i=[0,1]\setminus O_i$ 即 O_i 的补集没有被观测到时,分别称 O_i 与 M_i 为观测区域和缺失区域,并分别称 $M_i\neq\varnothing$ 与 $M_i=\varnothing$ 时对应的样本曲线 $X_i(t)$ 为部分观测函数型数据和完整观测函数型数据。

参照文献[12],在接下来的讨论中,假设 $X_i(t)$, $i=1,2,\cdots,n$ 是由 $X(t)$ 的生成机制 P 随机生成的 n 个独立样本,O_i, $i=1,2,\cdots,n$ 是由观测区域 O 的生成机制 Q 随机生成 n 个独立的样本,且 P 和 Q 相互独立,即部分观测函数型数据为完全随机缺失,其中 O_i 可以由 $[0,1]$ 的多个子集构成。

为了方便叙述,简记 $X_i(t)$ 为 X_i,并将样本曲线的全体及部分观测样本曲线的全体分别记为 $S=\{X_i\mid i=1,2,\cdots,n\}$ 和 $S^M=\{X_l\mid l\in L\}$,其中 $L=\{1\leqslant l\leqslant n\mid M_l\neq\varnothing\}$。同时,将任意样本曲线 X_i 观测到的部分即 $\{X_i(t),t\in O_i\}$ 与未观测到的部分即 $\{X_i(t),t\in M_i\}$ 分别记为 (X_i,O_i) 和 (X_i,M_i)。

1.2　基于深度的重构方法

对任意部分观测样本曲线 $X_l\in S^M\subset S$, Elías 等[11]提出的基于深度的重构方法主要利用部分观测函数型数据相关深度的知识[13]及有关部分观测函数型数据的包络与投影等思想[14]。数据深度或统计深度函数是测量一个数据有多靠近其相应概率分布中心或其所在数据集中心的函数,即测量一个数据在其相应概率分布或其所在数据集下中心度的函数。一般来说,数据深度越大,数据越靠近其相应概率分布中心或其所在数据集中心,深度提供了一种数据由中心向外排序的半序方法。

对来自随机变量 X 的样本 x,记 $\mathscr{F}$ 为实数域 $\mathbf{R}$ 上所有概率分布的集合,$F\in\mathscr{F}$ 为 X 的概率分布,则其深度(单变量统计深度函数)定义为 $\mathbf{R}\times\mathscr{F}\to[0,1]$ 的一个函数 $D(x,F(x))$,其中 $D(x,F(x))$ 在 x 的取值为 F

的中位数时达到最大,而随着 x 远离 F 的中位数而减小,例如

$$D(x,F(x))=2\{F(x)[1-F(x)]\}。\tag{1}$$

常见的数据深度还有半空间深度[15]、单纯形深度[16]、马氏深度[17]等。

对函数型数据 $X(t)$,即随机函数 $X(t):[0,1]\to\mathbf{R}$,记 $D(\cdot)$ 为一单变量统计深度函数,$w(t)$ 为 $[0,1]\to[0,+\infty)$ 的一权重函数,满足 $\int_0^1 w(t)\mathrm{d}t=1$,则积分泛函深度(integrated functional depth, IFD)[18-19]可表示为

$$d_{\mathrm{IFD}}(X(t),P)=\int_0^1 D(X(t),P_t)w(t)\mathrm{d}t,$$

其中 $P_t=P\{X(t)\leqslant x\}$ 为 $X(t)$ 的边际概率分布。

对部分观测函数型数据,记 $Q(t)=P\{t\in O\}$ 为观测区域 O 涵盖观测点 t 的概率,不失一般性,假设对 $\forall t\in[0,1]$ 都有 $Q(t)>0$。同时记 ϕ 为定义在 $[0,1]$ 上有界且使得 $\int_O\phi(Q(t))\mathrm{d}t>0$ 几乎处处成立的连续函数,则 Elías 等[13]基于 IFD 提出的部分观测积分泛函深度(partially observed integrated functional depth, POIFD)可表示为

$$d_{\mathrm{POIFD}}((X(t),O),P\times Q)=\int_O D(X(t),P_t)w_\phi(t\mid O)\mathrm{d}t,$$

其中 $w_\phi(t\mid O)=\dfrac{\phi(Q(t))}{\int_O\phi(Q(t))\mathrm{d}t}$ 为权重函数。

对有限部分观测函数型数据,记 $J(t)=\{1\leqslant j\leqslant n\mid t\in O_j\}$ 并假设 $J(t)\neq\varnothing$,$q(t)=\#J(t)$ 为在观测点 t 有观测值的样本曲线个数,则任意样本曲线 (X_i,O_i) 在 O_i 上的样本部分观测积分泛函深度可表示为

$$d_{\mathrm{POIFD}}((X_i(t),O_i),P\times Q)=\frac{\int_{O_i}D(X_i(t),F_{J(t)})q(t)\mathrm{d}t}{\int_{O_i}q(t)\mathrm{d}t},$$

其中 $F_{J(t)}$ 为样本 $\{X_j(t)\mid j\in J(t)\}$ 的经验分布函数。

接下来,根据文献[11,13]等给出以下部分观测函数型数据中包络、投影等的定义。

定义 1 对任意的 $X_l\in S^M$,若 $T_l=\{(X_j,O_j)\}$ 描述了 (X_l,O_l) 的形状、大小等特征,则称 T_l 为 (X_l,O_l) 的一个包络,其中 j 来自 $\{1\leqslant j\leqslant n\mid j\neq l\}$ 的一个子集且满足 $\lambda(O_j\cap O_l)>0$,$\lambda(\cdot)$ 为定义在实数域 $\mathbf{R}$ 上的一勒贝格测度。若记 T_l 中元素的下标集为 J_l,则 T_l 可记为 $T_l=\{(X_j,O_j)\mid j\in J_l\}$。

定义 2 对任意的 $X_l\in S^M$,设 $T_l=\{(X_j,O_j)\mid j\in J_l\}$ 和 $T_l'=\{(X_j,O_j)\mid j\in J_l'\}$ 为 (X_l,O_l) 的 2 个包络,若

$$\lambda\Big(\{t\in O_l\mid \min_{j\in J_l(t)}X_j\leqslant X_l\leqslant\max_{j\in J(t)}X_j\}\Big)>\lambda\Big(\{t\in O_l\mid \min_{j\in J_l'(t)}X_j\leqslant X_l\leqslant\max_{j\in J'(t)}X_j\}\Big),$$

则称包络 T_l 大于包络 T_l',其中 $J_l(t)=\{j\in J_l\mid t\in O_j\}$,$J_l'(t)=\{j\in J_l'\mid t\in O_j\}$。

定义 3 对任意的 $X_l\in S^M$,$X_j\in S$,$j\neq l$,若 $\lambda(O_j\cap O_l)>0$,则 X_j 与 X_l 之间的平均欧氏距离,即两者在共同观测区域 $O_j\cap O_l$ 上对应部分间的平均欧氏距离定义为

$$\|(X_j,O_j)-(X_l,O_l)\|=\frac{\sqrt{\int_{O_j\cap O_l}|X_j(t)-X_l(t)|^2\mathrm{d}t}}{\lambda(O_j\cap O_l)}。$$

定义 4 对任意的 $X_l\in S^M$,记 $\delta=\min_{j\in J_l}\|(X_j,O_j)-(X_l,O_l)\|$,并假设对任意的 $t\in O_l$ 都有 $J_l(t)\neq\varnothing$,则称包络 T_l 中样本曲线的加权平均

$$\hat{X}_l^\theta(t)=\frac{\sum_{j\in J_l(t)}\omega_jX_j(t)}{\sum_{j\in J_l(t)}\omega_j}$$

为包络 T_l 在 $[0,1]$ 上的指数加权投影,其中

$$\omega_j=\exp\left(\frac{-\theta\|(X_j,O_j)-(X_l,O_l)\|}{\delta}\right),$$

θ 为调整包络中样本曲线在投影中重要性的调节参数。易知,投影曲线 $\hat{X}_l^{\theta}(t)$ 包含 2 个部分:一部分是由 $\{(X_j, O_j \cap O_l) \mid j \in J_l\}$ 加权平均所得,记为 $(\hat{X}_l^{\theta}(t), O_l)$;一部分由 $\{(X_j, O_j \cap M_l) \mid j \in J_l\}$ 加权平均所得,记为 $(\hat{X}_l^{\theta}(t), M_l)$。

实际应用中,当 $\hat{X}_l^{\theta}(t)$ 在 O_l 即 $\bigcup\limits_{j \in J_j}(O_j \cap O_l)$ 可计算得到时,参数 θ 可通过式(2)求解得到:

$$\theta = \arg\min_{v} \sum_{l \in L} \| (X_l, O_l) - (\hat{X}_l^{v}(t), O_l) \|^2 。 \tag{2}$$

对拟重构的部分观测样本曲线 X_l,基于深度的重构过程包括以下 2 步:一是根据 $S \backslash \{X_l\}$ 中样本曲线与 X_l 平均欧氏距离的大小,由近到远,依次从 $S \backslash \{X_l\}$ 中选择样本曲线,通过动态迭代更新,构建 (X_l, O_l) 越来越大的包络 T_l,并在每次迭代中,计算 X_l 在 $T_l \cup \{X_l\}$ 中的深度,选择临近2次迭代中深度较大的包络作为新的包络,以得到具有文献[11]中性质 P1—P3 且 X_l 具有最深深度的最终包络;二是将 (X_l, O_l) 的最终包络 T_l 在 $[0,1]$ 上投影得到 X_l 的估计 $\hat{X}_l^{\theta}(t)$,简记为 $\hat{X}_l(t)$,并将 $\hat{X}_l(t)$ 的 $(\hat{X}_l(t), M_l)$ 部分作为 (X_l, M_l) 的估计,以此完成部分观测样本曲线的重构。基于深度的重构算法详见文献[11]。

1.3　C-Depth-based 重构方法的提出

记样本曲线集 $S = \{X_i \mid i = 1,2,\cdots,n\}$ 中部分观测样本曲线的占比为 c,任意部分观测样本曲线 $X_l \in S^M$ 的缺失比即 $\lambda(M_l)/\lambda([0,1])$ 为 p_l。易知,在 c 较大、p_l 较大或样本曲线为非光滑曲线等情形下,由于样本信息相对较少等原因,因此部分观测函数型数据的重构变得相对困难。考虑上述情形下,基于深度的重构方法在均方预测误差(mean squared prediction error, MSPE)意义下相对正则化回归方法(regularized regression method, Reg.regression)[4] 的优良性,同时考虑不同样本曲线之间的潜在差异,本文先对 S 中样本曲线进行聚类,并基于类间信息,运用基于深度的重构方法对 X_l 进行重构,再将不同聚类情形下,即 X_l 属于不同类间时得到重构结果融合,即类信息融合,提出 C-Depth-based 重构方法。

C-Depth-based 重构方法实现步骤如下。

步骤 1　对 S^M 中的每条曲线 X_l 采用基于深度的重构方法得到 X_l 在 $[0,1]$ 上的估计 $\hat{X}_l(t)$,并将 $(\hat{X}_l(t), M_l)$ 部分作为 (X_l, M_l) 的估计,得到重构后完整的曲线。

步骤 2　将所有重构所得的曲线与 $S \backslash S^M$ 中的样本曲线组成新的曲线集 $S^* = \{X_i^*(t) \mid i = 1,2,\cdots,n\}$,并将 $X_i^*(t)$, $i = 1,2,\cdots,n$ 的样本均值函数和样本自协方差函数分别记为

$$\mu(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^*(t)$$

和

$$C_X(t,s) = \frac{1}{n}\sum_{i=1}^{n} [X_i^*(t) - \mu(t)][X_i^*(s) - \mu(s)] 。$$

步骤 3　运用函数型主成分分析法,计算 S^* 中每条曲线 $X_i^*(t)$ 的 R 个函数型主成分得分

$$\xi_{ir} = \int_0^1 (X_i^*(t) - \mu(t))\phi_r(t)\mathrm{d}t, \quad r = 1,2,\cdots,R,$$

其中,$\phi_r(t)$ 为 $X(t)$ 的样本自协方差函数 $C_X(t,s)$ 的第 r 个特征函数,R 为 $X_i^*(t)$ 基于主成分基所得 K-L 展开式 $X_i^*(t) = \mu(t) + \sum\limits_{r=1}^{\infty} \xi_{ir}\phi_r(t)$ 中函数项 $\sum\limits_{r=1}^{\infty} \xi_{ir}\phi_r(t)$ 的截断个数。通常考虑使得累计贡献率(变差解释百分比) $\sum\limits_{r=1}^{R} \lambda_r / \sum\limits_{r=1}^{\infty} \lambda_r$ 较大时,如 $\sum\limits_{r=1}^{R} \lambda_k / \sum\limits_{r=1}^{\infty} \lambda_r \geqslant 85\%$ 时的 R 作为函数项的截断个数,其中 λ_r 为与特征函数 $\phi_r(t)$ 对应的非负特征值。

步骤 4　设聚类数为 $k(k = 1,2,\cdots,K)$,由步骤 3 计算所得 S^* 中每条曲线 $X_i^*(t)$ 的 R 个函数型主成分得分

$$\begin{aligned}
X_1^*(t):&\quad \xi_{11},\quad \xi_{12},\quad \cdots,\quad \xi_{1R},\\
X_2^*(t):&\quad \xi_{21},\quad \xi_{22},\quad \cdots,\quad \xi_{2R},\\
&\quad \vdots \qquad \vdots\\
X_n^*(t):&\quad \xi_{n1},\quad \xi_{n2},\quad \cdots,\quad \xi_{nR}。
\end{aligned}$$

运用 K 均值聚类方法,对 S^* 中的曲线进行聚类,得到不同聚类数 k 下 S^* 中曲线的聚类结果,对应地,得到不同聚类数 k 下 S 中原始样本曲线的聚类结果。

步骤 5 记聚类数为 k 时,S 中原始样本曲线的聚类结果为 $C_{k1},C_{k2},\cdots,C_{kk}$。对拟重构曲线 X_l,依次在聚类数 $k=1,2,\cdots,K$ 情形下,基于类间信息,采用基于深度的重构方法进行如下估计:

1) 当 $k=1$ 时,聚类结果即为 S,此时基于 S 中的样本曲线,直接采用基于深度的重构方法,即得聚类数 $k=1$时 X_l 在$[0,1]$上的估计 $\hat{X}_l^1(t)$;

2) 当 $k=2$ 时,聚类结果即为 C_{21},C_{22},此时记 X_l 所属的类为 $C_{2\cdot}$,其中 $C_{2\cdot}$ 表示 C_{21}或 C_{22},则认为 X_l 与 $C_{2\cdot}$ 中的曲线变化轨迹等相似度高,此时,基于 $C_{2\cdot}$ 中的样本曲线,采用基于深度的重构方法,即得聚类数 $k=2$ 时 X_l 在$[0,1]$上的估计 $\hat{X}_l^2(t)$;

3) 类似地,当 $k=K$ 时,聚类结果即为 $C_{K1},C_{K2},\cdots,C_{KK}$。此时 X_l 所属的类为 $C_{K\cdot},C_{K\cdot}$ 为 $C_{K1},C_{K2},\cdots,C_{KK}$中的一个,则基于 $C_{K\cdot}$ 中的样本曲线,采用基于深度的重构方法,即得聚类数 $k=K$ 时 X_l 在$[0,1]$上的估计 $\hat{X}_l^K(t)$。

从而得到了 K 个聚类情形下 X_l 的 K 个估计结果 $\hat{X}_l^1(t),\hat{X}_l^2(t),\cdots,\hat{X}_l^K(t)$。

步骤 6 计算 X_l 的 K 个估计结果在 O_l 部分的均方误差(mean squared error, MSE),

$$E_{\mathrm{MSE}}=\|(X_l,O_l)-(\hat{X}_l^k,O_l)\|^2,\quad k=1,2,\cdots,K,$$

得到 K 个估计结果的 E_{MSE}估计值,并将其中最小 E_{MSE}及其所对应的估计结果分别记为 e_1 和 $\hat{X}_l^{D_1}$。

步骤 7 为充分利用不同聚类数下得到的估计结果和降低聚类数 k 的影响,采用自加权集成学习算法动态赋权对 X_l 的任意 $w(w=2,3,\cdots,K)$个估计结果进行融合,假设任选 w 个估计结果对应的聚类数所组成的集合记为 D_w,得到融合后的结果 $\hat{X}_{l,w}^*$

$$\hat{X}_{l,w}^*=\frac{\sum\limits_{k\in D_w}\varphi_{k,l}\hat{X}_l^k}{\sum\limits_{k\in D_w}\varphi_{k,l}},$$

其中 $\varphi_{k,l}=1/2(\|(X_l,O_l)-(\hat{X}_l^k,O_l)\|^2)$为权重。将得到 C_K^w 个融合结果,计算每个融合结果在 O_l 部分的 E_{MSE}估计值,将最小 E_{MSE}值所对应的融合结果视为融合个数为 w 时的最终融合结果,并将最小 E_{MSE}值与最终融合结果分别记为 e_w 和 $\hat{X}_l^{D_w}$。

直到 $w=K$ 时,将得到 $K-1$ 个融合结果 $\hat{X}_l^{D_2},\hat{X}_l^{D_3},\cdots,\hat{X}_l^{D_K}$ 及其所对应的 E_{MSE}值 $e_2,e_3,\cdots,e_K$。

步骤 8 将 $e_1,e_2,\cdots,e_K$ 中的最小值 $e_{\cdot}$ 所对应的融合结果 $\hat{X}_l^{D_{\cdot}}$ 作为 X_l 的最终估计结果,其中$\hat{X}_l^{D_{\cdot}}\in\{\hat{X}_l^{D_1},\hat{X}_l^{D_2},\cdots,\hat{X}_l^{D_K}\}$,并将$(\hat{X}_l^{D_{\cdot}}(t),M_l)$部分作为$(X_l,M_l)$的最终估计,得到最终重构后完整的曲线 $\hat{X}_l^*$。

2 数值模拟

为了验证 C-Depth-based 方法在重构部分观测函数型数据时的有效性,下面选取式(1)定义的 $D(\cdot)$为深度函数,通过数值模拟将 C-Depth-based 方法与基于深度的重构方法及正则化回归方法做比较研究。

2.1 实验设计

首先根据文献[13],对模型 $X(t)=\mu(t)+\varepsilon(t)$,其中均值函数 $\mu(t)$是一个协方差为 $\rho_\mu(s,t)=\sigma\mathrm{e}^{-(2\sin(\pi|s-t|)^2/l^2)}$的中心高斯过程,随机误差函数 $\varepsilon(t)$是一个协方差为 $\rho_\varepsilon(s,t)=\alpha\mathrm{e}^{-\beta|s-t|}$ 的中心高斯过程,$s,t\in[0,1]$,设置参数$\beta=5$, $\alpha=0.5$, $\sigma=3$, $l=0.5$,在$[0,1]$上随机生成 200 个等距的点以作为任意样本曲线的估计,即以这 200 个点所在曲线作为样本曲线。为了解样本容量对重构效果的影响,这里设定样本容量 $n=200$和 $n=600$,即由上述方法分别模拟生成 200 条和 600 条样本曲线。

其次设样本曲线中部分观测样本曲线的占比 c 分别为 0.5、0.75、0.9、1。$c=1$ 意味着样本曲线中没有一条曲线被完整观测到。

对任意部分观测样本曲线,假设其在[0,1]的 m 个互不相交的子区间上有观测数据,而其余子区上没有观测数据,观测比为 $1-p$,部分观测样本曲线由样本曲线删除相应子区间上的数据生成。这里设置 $m=8$,同时为分析缺失比对重构方法的影响,设置 $p=0.25,0.5$。

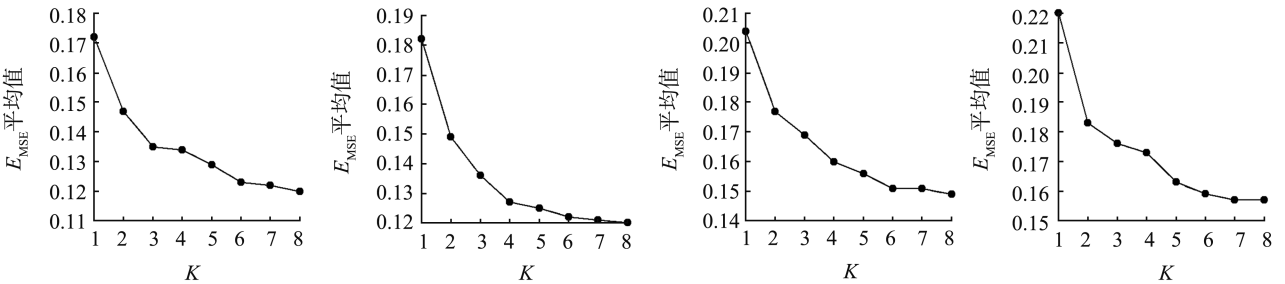
最后,采用均方预测误差(mean squared prediction error, MSPE)作为重构效果评价标准,其计算公式如下:

$$E_{\mathrm{MSPE}}=\frac{1}{n^*}\sum_{l\in L}\|(X_l,M_l),(\hat{X}_l^*,M_l)\|^2。$$

其中 n^* 表示样本中部分观测样本曲线个数。

1.5　聚类数 K 的选取

在 C-Depth-based 重构方法中,聚类数 K 的选取对重构效果存在一定的影响。K 较小意味着簇划分相对较粗,从而可能导致重构误差相对较大。而 K 较大意味着簇内样本相对较少,从而可能导致簇内样本不能有效捕捉拟重构曲线的真实特征。实际工作中,可先计算 K 取不同值时每条样本曲线的 E_{MSE},再求出所有样本曲线 E_{MSE}的平均值,并绘制 E_{MSE}的平均值随 K 变化的曲线,选择 E_{MSE}的平均值下降变化的拐点作为最佳聚类数。在本节模拟分析中,$n=200$ 时,将 K 的初始取值范围设置为 1—8,参照肘部法则,计算并绘制不同条件下 E_{MSE}的平均值随 K 变化的曲线如图 1 所示。可知 E_{MSE}的平均值随 K 的增大而逐渐减小,而当 K 大于等于 6 时,E_{MSE}的平均值下降趋于平稳,因而取 6 为最佳聚类数,以避免过拟合等问题。$n=600$ 时,类似地,可得 9 为最佳聚类数。



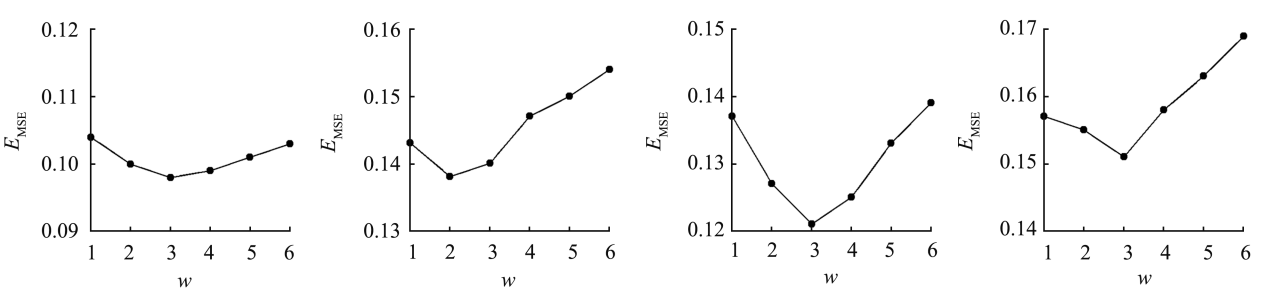
(a) c=0.50, p=0.25　(b) c=0.75, p=0.25　(c) c=0.90, p=0.25　(d) c=1.00, p=0.25

图 1　不同条件下 E_{MSE}的平均值随聚类数 K 的变化曲线

Fig.1　Curve of the mean value of E_{MSE} with the number of clusters K under different conditions

2.3　融合数对重构效果的影响

当最佳聚类数确定为 K 时,对任意一条拟重构曲线,由基于深度的重构方法,可以得到 K 条重构曲线。那么,从 K 条重构曲线中选择几条重构曲线或怎样选择重构曲线进行融合,以得到最终的重构曲线,这对重构效果有着至关重要的影响。对 $n=200$, $p=0.5$, $c=0.75$ 的情形,给出了 4 条拟重构曲线基于不同融合数所得重构曲线的 E_{MSE},如图 2 所示。融合后所得重构曲线的 E_{MSE}随着融合数的增加呈现先减小后增大的趋势,这说明当融合数较小或较大时,由于引入信息不够或冗余,因此重构效果不能达到最佳。同时,由图 2 可知,对不同的拟重构曲线,融合的最佳点也不一定相同,如对拟重构曲线 1、2,其最佳融合数分别为 3、2,因此,对不同的拟重构曲线,动态选择融合数对最终重构结果至关重要。



(a) 变化曲线1　(b) 变化曲线2　(c) 变化曲线3　(d) 变化曲线4

图 2　不同样本曲线的 E_{MSE}随融合数的变化曲线

Fig.2　Curve of E_{MSE} estimates with the number of fusions for different sample curves

2.4 结果分析

C-Depth-based 方法、基于深度的重构方法及正则化回归方法在不同情形下的 E_{MSPE} 估计值见表 1。对所有模拟情形,C-Depth-based 方法在均方预测误差意义下均优于基于深度的重构方法。当部分观测函数型数据中部分观测样本曲线占比较大及部分观测样本曲线缺失比较大,如 $c=0.9$, $p=0.5$ 时,C-Depth-based 方法所得重构曲线的均方预测误差相对最小,即 C-Depth-based 方法表现最优。正则化回归方法则在部分观测样本曲线占比较小时表现更好。同时,由表 1 可知,随着样本量得增加,3 种重构方法所得曲线的均方预测误差均呈现减小趋势,这说明,增加样本容量,可有效提高重构精度。而随着 p 与 c 增大,不同重构方法下的均方预测误差则呈增大趋势,这说明,缺失比和部分观测样本曲线占比越大,重构难度越大。

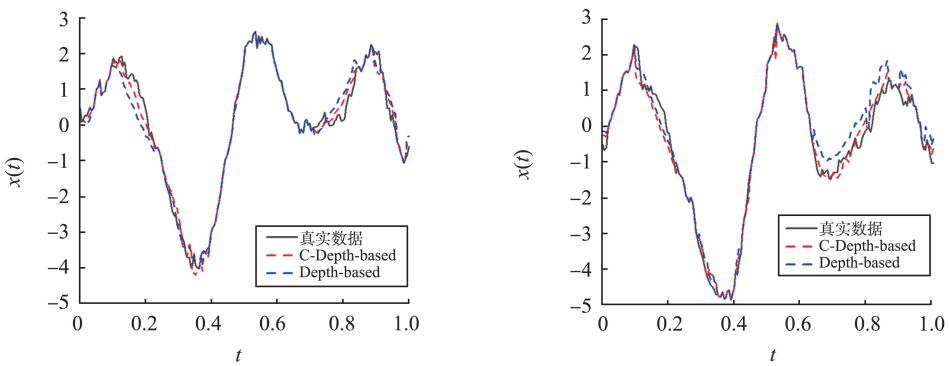
表 1 不同方法下 E_{MSPE} 估计值

Table 1 E_{MSPE} estimates value under different methods

n	方法	$c=0.50$		$c=0.75$		$c=0.90$		$c=1.00$	
		$p=0.25$	$p=0.50$	$p=0.25$	$p=0.50$	$p=0.25$	$p=0.50$	$p=0.25$	$p=0.50$
200	Reg.regression	**0.054**	**0.106**	**0.059**	**0.133**	**0.101**	0.196		
	Depth-based	0.173	0.191	0.178	0.197	0.212	0.224	0.218	0.232
	C-Depth-based	0.114	0.133	0.121	0.146	0.158	**0.184**	**0.160**	**0.172**
600	Reg.regression	**0.060**	**0.097**	**0.053**	**0.121**	**0.119**	0.171		
	Depth-based	0.160	0.176	0.171	0.183	0.190	0.215	0.196	0.221
	C-Depth-based	0.113	0.124	0.121	0.134	0.141	**0.167**	**0.142**	**0.164**

注:空白表示该方法不能提供重构。

为更直观地展示 C-Depth-based 方法在 $c=1$ 极端情况下的优越性,图 3 给出 $n=200$, $c=1$, $p=0.5$ 情形下下 2 条拟重构曲线与 2 种重构方法下所得重构曲线的对比图。从图 3 可以看出,利用 C-Depth-based 方法得到的重构曲线较基于深度的重构方法得到的重构曲线几乎在每一个观测点都更贴近于真实曲线。



(a) 重构曲线1 (b) 重构曲线2

图 3 重构曲线对比

Fig.3 Comparison of reconstruction curves

3 实例分析

为进一步验证所提方法的有效性,本章选取 2014 年北京 PM 2.5质量浓度的数据(https://archive.ics.uci.edu/datasets/Beijing+PM 2.5+Data)进行实证分析。该数据集包含 2014 年北京每天共 365 条 PM 2.5质量浓度测量数据,每条数据包含 24 个观测值。同时,为做好与真实数据的比较分析,选取其中 200 条完整数据作为实际分析样本数据。

3.1 实验设计

为模拟实际应用中可能出现的情形,本节做如下设定。假设 200 条样本曲线中部分观测样本曲线的占比 c 分别为 0.5、0.75、0.9 和 1,部分观测样本曲线的缺失比 p 分别为 0.25、0.5。而部分观测样本曲线则由完整样本曲线删除其观测区间[0,24]上互不相交的随机子区间上对应的观测数据生成,这里随机子区间的个数设置为 3。这种模拟方式能够较好地反映现实世界中因设备故障、数据传输中断等原因导致的局部数据缺失问题。

3.2　结果分析

北京 PM2.5 质量浓度数据在不同方法下的重构效果见表 2。在均方预测误差准则下,C-Depth-based 方法在所有模拟情形下均优于基于深度的重构方法。当部分观测函数型数据中部分观测样本曲线占比较大及部分观测样本曲线缺失比较大,如 $c=0.9$, $p=0.5$ 时,C-Depth-based 方法表现均优于其他 2 种方法。而在部分观测样本曲线占比较小、缺失比较小时,正则化回归方法表现更优。实例分析结果与数值模拟结果一致。

表 2　不同方法下 PM2.5 质量浓度数据重构的 E_{MSPE} 估计值

Table 2　E_{MSPE} estimates value for PM2.5 mass concentration data reconstruction under different method

方法	$c=0.5$		$c=0.75$		$c=0.90$		$c=1.00$	
	$p=0.25$	$p=0.50$	$p=0.25$	$p=0.50$	$p=0.25$	$p=0.50$	$p=0.25$	$p=0.50$
Reg.regression	**8.202**	**8.454**	**10.228**	11.526	**11.243**	13.352		
Depth-based	10.646	11.356	13.333	14.051	17.151	17.838	17.157	18.896
C-Depth-based	8.361	9.217	11.155	**11.168**	12.689	**13.186**	**15.375**	**16.609**

注:空白表示该方法不能提供重构。

4　结语

针对部分观测函数型数据,本文基于深度和融合类信息的重构思想,提出一种新的部分观测样本曲线重构方法。该方法充分利用样本曲线之间的相关性,同时降低重构过程中对协方差估计的依赖,为部分观测函数型数据的重构提供一种有效的解决方案。数值模拟和实例分析结果显示,当部分观测函数型数据中部分观测样本曲线占比较大及部分观测样本曲线缺失比较大时,该方法在均方预测误差意义下均优于基于深度的重构方法和正则化回归方法,且对部分观测样本曲线占比为 1 的极端情况也具有良好的适用性。

参考文献:

[1] RAMSAY J O, SILVERMN B W. Functional data analysis[M]. New York: Springer, 2005: 19-34.

[2] JAMES G M, HASTIE T J, SUGAR C A. Principal component models for sparse functional data[J]. Biometrika, 2000, 87: 587-602.

[3] CHIOU J M, ZHANG Y C, CHEN W H, et al. A functional data approach to missing value imputation and outlier detection for traffic flow data[J]. Transportmetrica B(Transport Dynamics), 2014, 2(2):106-129.

[4] KRAUS D. Components and completion of partially observed functional data[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2015, 77(4):777-801.

[5] KRAUS D, STEFANUCCI M. Ridge reconstruction of partially observed functional data is asymptotically optimal[J]. Statistics & Probability Letters, 2020, 165(20):108813.

[6] KNEIP A, LIEBL D. On the optimal reconstruction of partially observed functional data[J]. The Annals of Statistics, 2020, 48(3):1692-1717.

[7] LI P L, CHIOU J M. Functional clustering and missing value imputation of traffic flow trajectories[J]. Transportmetrica B: Transport Dynamics, 2021, 9(1):1-21.

[8] 高海燕,马文娟,薛娇. 融合类信息的函数型矩阵填充方法与应用[J]. 统计与决策,2023,39(23):40-45.
GAO Haiyan, MA Wenjuan, XUE Jiao. Functional matrix completion method with class information and its application[J]. Statistics & Decision, 2023, 39(23):40-45.

[9] 杨玉杰,凌能祥. 不完全观测的部分函数型线性分位数回归模型及应用[J]. 山东大学学报(理学版),2025,60(3):100-106.
YANG Yujie, LING Nengxiang. Partially linear quantile regression model and empirical research for incomplete functional data[J]. Journal of Shangdong University(Natural Science), 2025, 60(3):100-106.

(下转第 101 页)