响应变量随机缺失的相依函数型单指标模型的 k 近邻估计

何文然,黄振生
南京理工大学 数学与统计学院,南京 210094

摘 要:针对具有 α 混合结构的函数型时间序列数据,当响应变量随机缺失时,利用函数型单指标模型进行统计建模,并采用 k 近邻方法对模型中未知参数和未知函数进行估计,与经典核方法相比,其数据适用性更强,可以提高估计效率;通过数值模拟和厄尔尼诺海平面温度数据,将 k 近邻方法和经典核方法进行比较,讨论 k 近邻方法与经典核方法对未知参数和未知函数的估计效果;从模拟结果可以看到:k 近邻方法对未知参数和未知函数的估计精度以及随样本增加的改善效果要优于经典核方法,在真实数据分析中,k 近邻对真实数据的精度拟合以及趋势拟合都表现优异;这些结果表明:在响应变量随机缺失的时间序列单指标模型中,采用 k 近邻方法对未知参数和未知函数进行估计,在精度上要优于经典核方法,同时在真实数据分析中,相比经典核方法,k 近邻方法能更好地拟合数据。

关键词:函数型单指标模型;α 混合;k 近邻估计;随机缺失
中图分类号:O212.7 **文献标识码**:A doi:10.16055/j.issn.1672-058X.2023.0006.013

k-nearest Neighbor Estimation for Dependent Function Single-indicator Models with Random Missing Response Variables

HE Wenran, HUANG Zhensheng
School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing 210094, China

Abstract: For functional time series data with α-mixed structure, when the response variables are randomly missing, the functional single indicator model is used for statistical modeling, and the k-nearest neighbor method is used to estimate the unknown parameters and unknown functions in the model. Compared with the classical kernel method, the method proposed in this paper has better data applicability and can improve the estimation efficiency. The k-nearest neighbor method was compared with the classical kernel method through numerical simulations and El Niño sea level temperature data to discuss the estimation effects of the k-nearest neighbor method and the classical kernel method on the unknown parameters and unknown functions. From the simulation results, it can be seen that the k-nearest neighbor method outperformed the classical kernel method in terms of accuracy of estimation of unknown parameters and unknown functions as well as improvement with increasing samples. Moreover, in the analysis of real data, the k-nearest neighbor method performed well in the accuracy fitting and trend fitting of real data. These results show that the k-nearest neighbor method is superior to the classical kernel method in terms of accuracy in estimating the unknown parameters and unknown functions in a single indicator model of a time series with random missing response variables. Meanwhile, in the real data analysis, the k-nearest neighbor method can better fit the data than the classical kernel method.

Keywords: functional single-indicator; α mixing; k-nearest neighbor estimation; missing at random

收稿日期:2022-06-21 **修回日期**:2022-10-07 **文章编号**:1672-058X(2023)06-0105-06
基金项目:全国统计科学研究重大项目(2018LD01).
作者简介:何文然(1998—),男,山东临沂人,硕士研究生,从事函数型数据研究.
引用格式:何文然,黄振生. 响应变量随机缺失的相依函数型单指标模型的 k 近邻估计[J]. 重庆工商大学学报(自然科学版),2023,40(6):105—110.
HE Wenran, HUANG Zhensheng. k-nearest neighbor estimation for dependent function single-indicator models with random missing response variables[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(6): 105—110.

1 引 言

过去几十年来,高维数据处理一般采用单指标模型进行降维,一方面可以避免维数灾难问题,另一方面能够保持多元回归中非参数方法的优势。该模型在生物医学、自然科学和经济金融诸多领域已体现出巨大的研究价值和广泛的实际应用前景。伴随着函数型数据分析的热潮,统计学者考虑将单指标模型降维的半参数优点和函数型数据结合,引入了函数型单指标回归模型[1]:

$$Y=r(\langle\theta,X\rangle)+\varepsilon \tag{1}$$

其中,函数 $r(\cdot)$ 是未知连接函数,$r:\mathbf{R}\to\mathbf{R}$;$\theta$ 是未知函数型单指标参数,取值于可分的 Hilbert 空间 H;ε 是随机误差项,满足 $E(\varepsilon|X)=0$;Y 是标量响应变量;X 是函数型协变量,取值于可分的 Hilbert 空间 H。

Ferraty 等[1]首先研究函数型单指标模型式(1),给出模型的核估计,并讨论了估计量的渐进性质,但并没有给出单指标函数的估计;随后,Ait-Saïdi 等[2]提出采用交叉验证方法估计未知的单指标参数,基于固定带宽,采用核估计得到未知的连接函数,并给出了估计量的渐近一致收敛速度;Attaoui 等[3]利用核方法研究函数型单指标模型的条件密度估计,给出估计量的逐点和一致几乎完全收敛速度,特别地,他们利用拟极大似然估计研究了单指标参数;Ferraty 等[4]基于泛函导数估计提出一种新的估计方法;Said 等[5]在强混合时间序列情况下,得到基于固定带宽的条件密度核估计量在一般条件下的均匀几乎完全收敛速度和渐近正态性,同时给出了估计量的置信区间;此外,Ding 等[6]研究一类函数型部分线性单指标模型并提出一种结合局部常数平滑的剖面最小二乘法来估计斜率函数和连接函数。近年来,Novo 等[7]第一次在独立条件下,在函数型半参数模型采用 k 近邻来估计未知连接函数,并讨论单指标已知与未知时估计量的渐近一致收敛速度,同时采用交叉验证来估计未知的单指标参数。但是,他们是在独立样本的条件下研究的,没有考虑相依情况下的函数型数据。

上述所有涉及的贡献都是在完全观测样本下发生的。然而,许多实际工作中,如市场调查、医学研究、信度检验等,一些观察结果可能不完整,通常被称为缺失数据,其统计分析因为内容很少或者缺失,变得非常困难和具有挑战性。随机缺失是最基本的,也是应用最广泛的关于缺失机理的假设。在解释变量为有限维时,可以在统计文献中找到许多这种情况的例子及其回归模型的统计推断。当解释变量为无限的情况或有函数型特征时,有很少的文献研究缺失数据的模型,仅有 Ferraty 等[8]基于独立同分布样本研究了响应变量随机缺失的函数型非参数回归模型;Ling 等[9]同样基于函数型非参数模型,利用函数型平稳遍历数据研究了未知估计量的渐近性质。近年来,Febrero-Bande 等[10]基于函数型线性回归模型,对于独立同分布样本下的标量响应随机缺失的函数型数据,提出一种效率更高的新模型估计方法;Ling 等[11]则对于响应变量随机缺失的强混合时间序列数据,研究了函数型单指数回归模型的未知参数和未知函数的估计,同时在一些正则条件下,得到了估计量的一致几乎完全收敛速度以及渐近正态性。

综上所述,函数型数据分析方法大多是在独立同分布的情形下研究的,与之相比,具有强混合时间序列的数据分析并没有受到广大学者的足够重视。函数型单指标模型作为一种半参数模型,同时具有参数模型和非参模型的优势,通常的估计方法一般为经典核方法,具有更强数据适用性的 k 近邻方法并没有被普及,且同时对于响应变量随机缺失的情形,参考文献至今未被研究。但是响应变量随机缺失的缺失数据,以及具有强混合结构的函数型时间序列数据是一类重要的亟待处理的问题。

受以上论文的启发,本文在强混合函数型时间序列数据和响应变量随机缺失下研究模型式(1)。利用具有局部窗宽的 k 近邻方法估计给出未知连接函数的估计量,改进了具有全局窗宽的函数型经典核方法。k 近邻方法可以自适应调整窗宽,对数据具有更好的适用性,利用模拟研究对比 k 近邻方法和函数型经典核方法的估计精度,验证所提模型和方法的有效性。同时,用两种估计方法对同一个实际例子进行分析拟合,通过实际数据拟合的好坏进一步说明 k 近邻估计方法的优越性。

文章的剩余部分如下:第一章给出了模型的估计方法,第二章进行了数值模拟,第三章给进行了实例分析,在第四章给出了结论。

2 函数型单指标模型和估计

首先给出强混合(α 混合)的定义,即若 $\{X_n,n\geqslant 1\}$ 为随机变量序列,且定义在概率空间 (Ω,A,P),取值在空间 (Ω',A') 上,定义 $-\infty\leqslant j\leqslant k\leqslant+\infty$,令 A_j^k 是由 X_s,$j\leqslant s\leqslant k$ 张成的 σ 域,若它的混合系数 $\alpha(n)$ 满足:$\alpha(n)=\sup_k \sup_{A\in A_{-\infty}^k} \sup_{B\in A_{k+n}^{+\infty}} |P(A\cap B)-P(A)P(B)|\to 0$,$n\to\infty$,则称 $X_n,n\geqslant 1$ 为 α 混合的随机变量序列。

考虑以下函数型单指标模型:

$$Y=r(\langle X_i,\theta\rangle)+\varepsilon_i,i=1,2,\cdots,n \tag{2}$$

假设 $\{(X_i,\delta_i,Y_i),1\leqslant i\leqslant n\}$ 是一列来自总体 (X,δ,Y) 的函数型数据样本,$Y_i\in\mathbf{R}$;$X_i\in H$;$\theta\in\theta(t)$,取值于可分的 Hilbert 空间 H,$t\in\mathbf{R}$;$d(\cdot,\cdot)$ 为空间 H 上的半度量,$d_\theta(x_1,x_2)=|\langle x_1-x_2,\theta\rangle|$;$\varepsilon_i$ 要满足 $E(\varepsilon_i|X_i)=0$;同时当 Y_i 缺失时,$\delta_i=0$,否则 $\delta_i=1$;假设 $P(\delta_i=1|Y_i,x_i)$

$=P(\delta_i=1|x_i)=p(x_i)$,$i=1,2,\cdots,n$。

类似于 Ling 等[11],$r(\langle\theta,x\rangle)$ 的 k 近邻估计量构造如下:

$$\hat{r}_n(\theta,x)=\frac{\sum_{i=1}^{n}(\delta_i Y_i K(H_{n,k,\theta}^{-1}(x)d_\theta(x,X_i)))}{\sum_{i=1}^{n}(\delta_i K(H_{n,k,\theta}^{-1}(x)d_\theta(x,X_i)))}=\frac{\hat{r}_{2n}(\theta,y,x)}{\hat{r}_{1n}(\theta,x)}$$

其中,

$$\hat{r}_{1n}(\theta,x)=\frac{1}{n}\sum_{i=1}^{n}(\delta_i\Delta_i)$$

$$\hat{r}_{2n}(\theta,y,x)=\frac{1}{n}\sum_{i=1}^{n}(\delta_i Y_i\Delta_i)$$

$$\Gamma_i(\theta,X_i)=\frac{K(H_{n,k,\theta}^{-1}(x)d_\theta(x,X_i))}{\sum_{i=1}^{n}\delta_i EK(H_{n,k,\theta}^{-1}(x)d_\theta(x,X_i))}$$

上式中包含的 $K(\cdot)$ 为核函数,而 $H_{n,k,\theta}(x)$ 为随机窗宽,定义如下:

$$H_{n,k,\theta}(x)=\min\{h\in R^+:\Sigma_{i=1}^{n}I_{B_\theta(x,h)}(x_i)=k\}$$

其中,$B_\theta(x,h)$ 为以 x 为中心,$h(h>0)$ 为半径的小球,$I_{B_\theta(x,h)}(\cdot)$ 为集合的示性函数。若 $H_{n,k,\theta}(x)=h_n(x)$,其中 $h_n(x)$ 为一列非负随机正序列,且随着 $n\to\infty$,$h_n(x)\to 0$,则类似于 Kudrasz 等[11]提出的经典核估计,窗宽依赖于固定点,表达式如下:

$$\hat{r}_n(\theta,x)=\frac{\sum_{i=1}^{n}(\delta_i Y_i K(h_{n,\theta}^{-1}(x)d_\theta(x,X_i)))}{\sum_{i=1}^{n}(\delta_i K(h_{n,\theta}^{-1}(x)d_\theta(x,X_i)))}$$

实际操作中,函数型单指标 θ 无法通过先验知识知晓,为此需要一个估计它的方法。这里和 Ling 等[11]一致,借用 Ding 等[7]的一个想法,采用剖面最小二乘法结合局部平滑常数技术来估计 θ。

下面给出估计流程。

步骤 1　构建含参数(θ)的目标损失函数:

$$Q(\theta)=\sum_{i:\delta_i=1}(Y_i-\hat{r}_n(\theta,x))^2$$

步骤 2　基于剖面最小二乘法,结合局部平滑常数技术来估计 θ。将 θ 分解为协方差函数主成分分解得到的基函数的累加和形式,同时确定主成分基函数的数目,最后将估计函数的问题变为估计基函数前系数的问题。

步骤 3　步骤 1 的含参(θ)目标函数,变为了未知量的基函数前系数,对损失函数求解最小值得到系数。

步骤 4　用得到的系数计算,得到 $\hat{r}_n(\theta,x)$ 以及 $\hat{\theta}$。

3　数值模拟

本节的目的是通过模拟对比 k 近邻方法(kNN)与经典核方法(NW),以验证本文所提方法的有效性。基于模型式(2),函数型解释变量 $X_i=X_i(t)$ 由以下函数生成:

$$x_i(t)=\alpha_i t^2+\sin\left(b_i\left(t-\frac{\pi}{3}\right)\right)+\cos\left(b_i\left(t-\frac{\pi}{3}\right)\right),t\in[0,\frac{\pi}{3}]$$

其中,α_i 独立同分布于 $N(0,\frac{\pi}{9})$,$b_i=1/4b_{i-1}+\xi_i$,b_i 中的参数 ξ_i 独立同分布于 $N(0,1)$,$b_0\sim N(0,1)$。另外,每个 $x_i(t)$ 都是$[0,\pi/3]$区间上 100 个等间隔的时间点上观测到的。图 1 是样本量为 300 时解释变量的曲线。

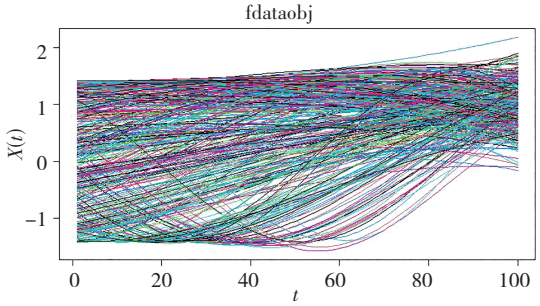

图 1　曲线 $x_{i=1,\cdots,200}(t_j)$,$t_{j=1,\cdots,100}\in[0,\pi/3]$

Fig. 1　Curve $x_{i=1,\cdots,200}(t_j)$,$t_{j=1,\cdots,100}\in[0,\pi/3]$

函数型单指标参数构造如下:

$$\theta(t)=\sin\left(3t-\frac{\pi}{3}\right),t\in\left[0,\frac{\pi}{3}\right]$$

本文的 $\theta(t)$ 估计方法同 Ling 等[11]一致,用积分平方误差 F_{RISE} 作为评价指标来评价 $\theta(t)$ 估计的好坏。

$$F_{\mathrm{RISE}}(\hat{\theta}(t))=\int_0^{\pi/3}(\hat{\theta}(t)-\theta(t))^2\mathrm{d}t$$

用平均均方误差 F_{AMSE} 来评价结果的好坏,同时核函数设置为 $K(u)=\frac{3}{4}\times\left(\frac{12}{11}-u^2\right)$。平均均方误差的表达式为

$$F_{\mathrm{AMSE}}=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_i-\hat{Y}_i)^2$$

其中,m_i 为每次生成的样本数,$\hat{Y}_i$ 为 Y_i 的估计值。

本文的缺失机制,参考 Ling 等[11],缺失机制满足:

$$p(x)=exgit\left(\frac{6\alpha}{\pi}\int_0^{\frac{\pi}{3}}x^2(t)\,\mathrm{d}t\right)$$

其中,$p(x)=P(\delta=1|X=x)$,对任意 $w\in\mathbf{R}$,有 $exgit(w)=\mathrm{e}^w/(1+\mathrm{e}^w)$。这里参数 α 控制缺失率,当 α 增加时,缺失率下降,这里选取 $\alpha=2$。

如表 1 所示,模拟研究了单指标 θ 已知、未知情形下,k 近邻方法与经典核方法预测效果的好坏,评价指

标为平均均方误差;同时也考虑 k 近邻方法与经典核方法对未知参数 θ 的估计优劣。模拟时,将样本分为训练集与预测集,用训练集来训练带宽,其中经典核方法为全局最优带宽,k 近邻方法为自适应窗宽,它可以基于样本得到一个个局部窗宽,对样本的适应性更优。为了方便,预测集样本量统一设置为 100。表 1 中的 n 为训练集样本量,同时对每个样本重复 200 次。

表 1 两种方法在不同样本量下的积分平方误差和平均均方误差

Table 1 The integral square error and mean square error of the two methods under different sample sizes

误差项目	n		
	100	200	300
F_{AMSE}(kNN)预测 θ 未知	0.003 691	0.002 162	0.001 797
F_{AMSE}(NW)预测 θ 未知	0.006 973	0.004 525	0.004 572
F_{AMSE}(kNN)预测 θ 已知	0.002 831	0.001 802	0.001 49
F_{AMSE}(NW)预测 θ 已知	0.005 148	0.004 02	0.003 789
$F_{\mathrm{RISE}}(\hat{\theta})$($k$NN)	0.134 291	0.122 782	0.074 221
$F_{\mathrm{RISE}}(\hat{\theta})$(NW)	0.213 972	0.130 81	0.122 524
F_{AMSE}(kNN)训练 θ 未知	0.002 964	0.001 666	0.001 199
F_{AMSE}(NW)训练 θ 未知	0.005 477	0.003 965	0.003 927
F_{AMSE}(kNN)训练 θ 已知	0.002 323	0.001 361	0.000 999
F_{AMSE}(NW)训练 θ 已知	0.004 392	0.003 487	0.003 323

从表 1 可以看出:在单指标 θ 未知情况下,预测集中,k 近邻方法的估计精度要优于经典核方法。在训练样本量为 100 时,相比于经典核方法,k 近邻方法对应的平均均方误差下降了 47%,在训练样本量为 300 时,下降了 60%。这说明随着样本量增加,k 近邻方法比经典核方法估计效果更优。

在单指标 θ 已知情况下,同样在预测集中,k 近邻方法的估计精度也要优于经典核方法,同样地,随着样本量的增加,k 近邻方法的改进幅度要大于经典核方法。在训练样本量为 100 时,相比于经典核方法,k 近邻方法对应的平均均方误差下降了 45%,在训练样本量为 300 时,下降了 60%,且相比于 θ 未知的时候,k 近邻方法与经典核方法估计精度都得到了提高。因此,对 θ 估计效果的好坏也是影响最终估计效果的一个重要因素。

从表 1 可以看出:在估计 θ 时,k 近邻方法的表现也要优于经典核方法,同时随着样本量的增大,k 近邻方法对 θ 估计效果的提升也要高于经典核方法。在训练样本量为 100 时,相比经典核方法,k 近邻对应的平均均方误差下降了 37%,在训练样本量为 300 时,下降了 39%。

在训练集中,k 近邻方法的表现也同样优于经典核方法。

在训练集中,当 θ 未知时,可以看到 k 近邻方法的估计精度要优于经典核方法。随着样本量的增加,k 近邻方法的改进幅度要大于经典核方法。在训练样本量为 100 时,相比于经典核方法,k 近邻方法对应的平均均方误差下降了 46%,在训练样本量为 300 时,下降了 69%。

在训练集中,当 θ 已知时,可以看到 k 近邻方法的估计精度同样也要优于经典核方法。随着样本量的增加,k 近邻方法的改进幅度要大于经典核方法。在训练样本量为 100 时,相比于经典核方法,k 近邻方法对应的平均均方误差下降了 47%,在训练样本量为 300 时,下降了 70%。

综上可以看出:k 近邻方法基于样本本身自适应调整带宽,其表现要优于经典核方法,因为核方法的带宽是全局固定带宽,并不会针对样本本身进行自适应调整;同时 k 近邻方法的表现随着样本量的增加,估计效果相较于经典核方法提升明显。基于样本自适应调整带宽的 k 近邻方法依赖于样本本身,随着样本量的增加,其效果更优。

图 2 为 k 近邻方法与经典核方法在预测集的平均均方误差箱线图,考虑训练集样本量为 200 的情形,进行了 200 次独立重复实验。

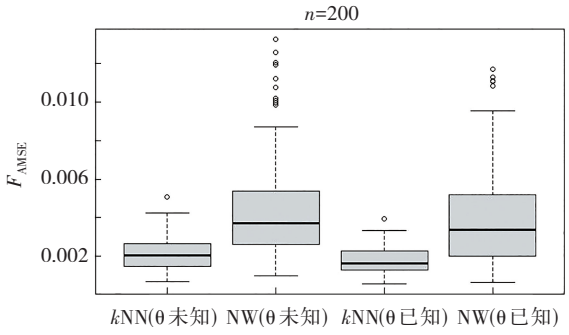

图 2 kNN 与 NW 预测集平均均方误差

Fig. 2 Mean square error of kNN and NW prediction sets

图 2 中,4 个箱线图从左到右依次是 θ 未知时 k 近邻方法对应的平均均方误差,θ 未知时经典核方法对应的平均均方误差,θ 已知时 k 近邻方法对应的平均均方误差,θ 已知时经典核方法对应的平均均方误差。

从图 2 左边两个图,可以明显看出:无论是 θ 未知还是已知,k 近邻方法对应的平均均方误差均明显优于经典核方法,k 近邻方法对应的平均均方误差要更集中且整体数值也要小;同时,在 θ 已知时,k 近邻方法以及经典核方法对应的平均均方误差相较于 θ 未知时都得到了一定改善;无论 θ 已知还是未知,相比于 k 近邻方法,经典核方法对应的平均均方误差的离群值要多一些。

综上,在预测集上,从 k 近邻方法与经典核方法对应的平均均方误差的箱线图中,可以得到 k 近邻方法要优于经典核方法的结论。

图 3 为 k 近邻与经典核方法在训练集的平均均方误差箱线图,同样考虑是样本量为 200 的情形,进行了 200 次独立重复实验。

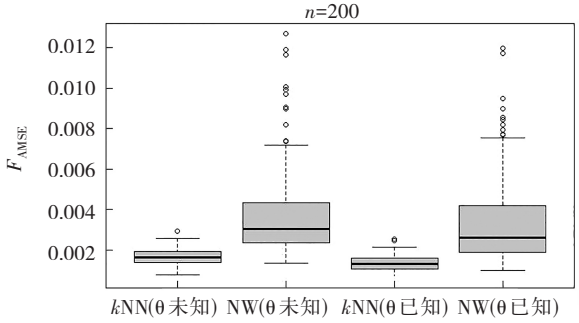

图 3　kNN 与 NW 训练集平均均方误差

Fig. 3　Mean square error of kNN and NW training sets

图 3 中,4 个箱线图从左到右依次是 θ 未知时 k 近邻方法对应的平均均方误差,θ 未知时经典核方法对应的平均均方误差,θ 已知时 k 近邻方法对应的平均均方误差,θ 已知时经典核方法对应的平均均方误差。

从图 3 可以看出:在训练集中,无论 θ 未知还是已知,k 近邻方法对应的平均均方误差同样是优于经典核方法,k 近邻方法对应的平均均方误差的集中性以及在数值大小上的表现与在预测集上一样优于经典核方法;无论 θ 是已知还是未知,k 近邻方法的离群值要少于经典核方法。

综上,在训练集上,从 k 近邻方法与经典核方法对应的平均均方误差的箱线图中,可以得到 k 近邻方法优于经典核方法的结论。

图 4 为 k 近邻方法与经典核方法在训练集中估计效果的对比图,横坐标为真值,纵坐标为估计值,其中蓝色点为 k 近邻方法,黑色点为经典核方法,这里 θ 是未知的。从图中可以明显看出:k 近邻方法估计效果要优于经典核方法,其对应点明显靠近 $y=x$ 直线,即图 4 中红色的直线。

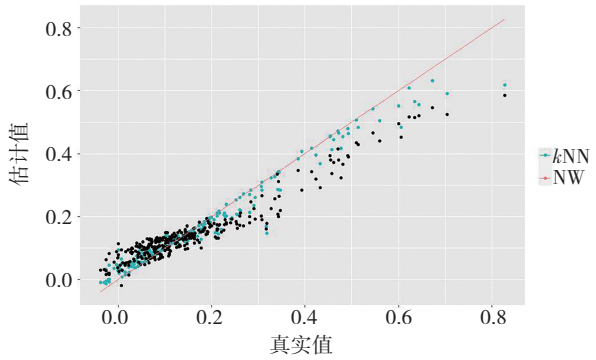

图 4　kNN 与 NW 在训练集中比较

Fig. 4　Comparison between kNN and NW

4　实例分析

这部分进行实例分析,分析 EL Nino 地域(0~100 s,800~900 w),时间跨度为 1982-01—2016-12 的海平面月温度数据。数据来源:http://www.cpc.ncep.noaa.gov/data/indices/。本节的目的是利用真实数据比较 k 近邻与经典核估计的优劣,比较 k 近邻方法与经典核方法对真实数据的预测效果,同时采用与上一节模拟相同的随机缺失机制,对数据的处理方式参考 Ling 等[15]。

首先,将 816 个月的海平面月温度数据$\{z_i, i=1,2,\cdots,816\}$转化为函数型数据:将 816 个月的温度离散化数据分割成 68a 的温度曲线数据,将其表示为 $x_i=\{v_j(t), 12(j-1)+4<t<12j+4\}, j=1,2,\cdots,68$(图 5)。

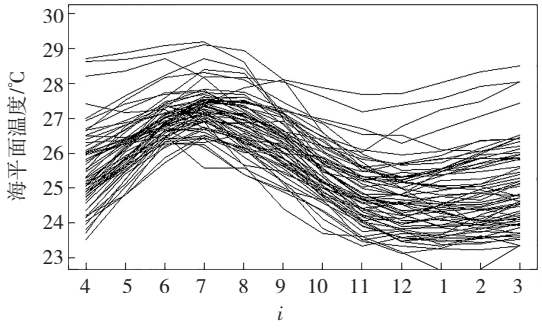

图 5　温度曲线

Fig. 5　Curves of temperature

响应实变量可以表示为 $Y_i(s)=\{v_{12(j+s)}, s=1,2,\cdots,12; j=1,2,\cdots,67\}$。现在可以建立样本量为 67 的相依样本$(x_i, Y_i(s))_{i=1,2,\cdots,67}$,其中 x_i 为函数型数据,$Y_i(s)$为实值。

将 67 个样本观测值$(x_i, Y_i(s))_{i=1,2,\cdots,67}$分成两个部分,一部分是学习样本$(x_i, Y_i(s))_{i=1,2,\cdots,66}$用于建立模型,另一部分是检验样本$(x_{67}, Y_{67}(s))$。建模过程核函数的选取与数值模拟的核函数一致。

最后用 k 近邻方法和经典核方法预测第 68 个数据,结果如图 6 所示:红色折线为 k 近邻方法的预测值,蓝色折线为经典核方法的预测值,黑色折线为数据真值。

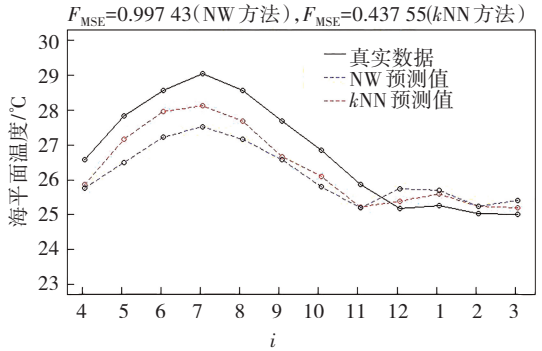

图 6　kNN 与 NW 预测比较

Fig. 6　Comparison between kNN and NW prediction

可以看到利用 k 近邻方法得到的均方误差(0.437 55)要小于经典核方法(0.997 43),在真实数据中,k 近邻方法的表现同样保持优异。同时从曲线贴合度来

看,k 近邻方法优于经典核方法,即 k 近邻方法估计的预测值更接近于真实值。

从曲线的变化趋势上来看,k 近邻方法也要优于经典核方法。气温在12月进入拐点,在这之后到3月气温保持相对平稳不变,虽然在11月和12月,k 近邻方法和经典核方法预测的气温走势都与真实气温数据走势相反,但是在12月到1月以及2月到3月,经典核方法预测的气温走势与真实数据是相反的,此时 k 近邻方法预测的气温走势与真实气温数据一致,更能说明真实气温的走势情况。由此可以看出:k 近邻方法预测的气温走势相比于经典核方法,更接近于真实数据。

在每个月份的估计效果上,k 近邻方法都优于经典核方法,其中5月、6月、7月、8月优越性表现更为明显。

这说明 k 近邻方法用数据本身自适应调整带宽能够很好地改进估计效果及精度,相比于全局固定带宽的经典核方法表现要好很多。

5 结 论

经过上述分析可以看出:数值模拟中无论是训练集还是预测集,相对于经典核方法,采用 k 近邻方法可以明显改进估计效果,同时对于单指标 θ 的估计,k 近邻方法比经典核方法要好,且随着样本量的增加,k 近邻的提升效果要明显优于经典核方法。

从实例分析来看:k 近邻方法在精确度以及稳定性上也都优于经典核方法,这说明采用 k 近邻对时间数据进行估计,其表现更加优异。这是因为核估计好坏主要受带宽的选取影响,k 近邻方法基于数据本身,自适应调整带宽,可以更好地改进估计效果,而经典核方法采用的带宽为全局最优带宽,无法依托于数据本身进行自适应调整,对数据的适应性较差。从模拟和实例分析的结果来看:自适应调整带宽的 k 近邻方法的估计效果比经典核方法的估计效果要好。

综上可以得到:k 近邻方法在响应变量随机缺失的时间序列单指标模型中表现优异,无论是模拟还是实例分析,其估计效果都明显优于经典核方法。

参考文献(References):

[1] FERRATY F, PEUCH A, VIEU P. Modèle à indice fonctionnel simple[J]. Comptes Rendus Mathematique, 2003, 336(12): 1025—1028.

[2] AIT-SAIEDI A, FERRATY F, KASSA R, et al. Cross-validated estimations in the single-functional index model[J]. Statistics, 2008, 42(6): 475—494.

[3] ATTAOUI S, LAKSACI A, SAID E O. A note on the conditional density estimate in the single functional index model[J]. Statistics & Probability Letters, 2011, 81(1): 45—53.

[4] FERRATY F, PARK J, VIEU P. Estimation of a functional single index model[C]//Recent advances in functional data analysis and related topics. Physica-Verlag HD, 2011: 111—116.

[5] ATTAOUI S. Strong uniform consistency rates and asymptotic normality of conditional density estimator in the single functional index modeling for time series data[J]. AStA Advances in Statistical Analysis: A Journal of the German Statistical Society, 2014, 98(3): 257—286.

[6] DING H, LIU Y H, XU W C, et al. A class of functional partially linear single-index models[J]. Journal of Multivariate Analysis, 2017, 161: 68—82.

[7] NOVO S, ANEIROS G, VIEU P. Automatic and location-adaptive estimation in functional single-index regression[J]. Journal of Nonparametric Statistics, 2019, 31(2): 364—392.

[8] FERRATY F, SUED M, VIEU P. Mean estimation with data missing at random for functional covariables[J]. Statistics, 2013, 47(4—6): 688—706.

[9] LING N, LIANG L, VIEU P. Nonparametric regression estimation for functional stationary ergodic data with missing at random[J]. Journal of Statistical Planning Inference, 2015, 162: 75—87.

[10] FEBRERO-BANDE M, GALEANO P, GONZÁLEZ-MANTEIGA W. Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random[J]. Computational Statistics & Data Analysis, 2019, 131: 91—103.

[11] LING N, CHENG L, VIEU P, et al. Missing responses at random in functional single index model for time series data[J]. Statistical Papers, 2022, 63(2): 665—692.

[12] A N L K, B P V. Uniform consistency of kNN regressors for functional variables[J]. Statistics & Probability Letters, 2013, 83(8): 1863—1870.

[13] FERRATY F, VIEU P. Nonparametric functional data analysis[M]. New York: Theory and Practice. Springer, 2006.

[14] EZZAHRIOUI M, OULD-SAÏD E. Asymptotic normality of a nonparametric estimator of the conditional mode function for functional data[J]. Journal of Nonparametric Statistics, 2008, 20(1): 3—18.

[15] LING N, MENG S, VIEU P. Uniform consistency rate of kNN regression estimation for functional time series data[J]. Journal of Nonparametric Statistics, 2019, 31(2): 451—468.

责任编辑:李翠薇