软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(9):2002−2017 [doi: 10.13328/j.cnki.jos.004635]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

时序数据曲线排齐的相关性分析方法*

姜高霞, 王文剑

(山西大学 计算机与信息技术学院,山西 太原 030006)

通讯作者: 王文剑, E-mail: wjwang@sxu.edu.cn, http://scit.sxu.edu.cn/SchoolTeacherD.aspx?id=326

摘 要: 时序数据是数据挖掘的一类重要对象.在做时序数据分析时,若不考虑数据的时差,则会造成相关性的误判.所以,时序数据存在相关性和时差相互制约的问题.通过对时序数据的相关性和协同性进行研究,给出了双序列的相关性判定方法和曲线排齐方法.首先,从时间弯曲的角度分析了两类相关性错误产生的原因及其特点;然后,根据相关系数的渐近分布得到相关系数在一定显著性水平上的界,将两者综合得到基于时移序列相关系数特征的相关性判定方法;最后,提出一种基于相关系数最大化的曲线排齐模型,其适用范围比 AISE 准则更广.模型采用光滑广义期望最大化(S-GEM)算法求解时间弯曲函数.在构造数据和真实数据上的数值实验结果表明:该相关性判别方法在伪回归识别中,比常规的 3 种相关系数以及 Granger 因果检验更有效;提出的 S-GEM 算法在大多数情况下明显优于连续单调排齐法(CMRM)、自模型排齐法(SMR)和极大似然排齐法(MLR).该文考虑的是双序列的线性相关问题和函数型曲线排齐方法,这些结果可为回归分析的相关性判定和时间对齐提供理论基础,并为多序列相关性分析和曲线排齐提供参考方向.

关键词: 伪回归;时间弯曲;相关性;曲线排齐;曲线光滑

中图法分类号: TP311

中文引用格式: 姜高霞,王文剑.时序数据曲线排齐的相关性分析方法.软件学报,2014,25(9):2002−2017. http://www.jos.org.cn/1000-9825/4635.htm

英文引用格式: Jiang GX, Wang WJ. Correlation analysis in curve registration of time series. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):2002−2017 (in Chinese). http://www.jos.org.cn/1000-9825/4635.htm

Correlation Analysis in Curve Registration of Time Series

JIANG Gao-Xia, WANG Wen-Jian

(School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China)

Corresponding author: WANG Wen-Jian, E-mail: wjwang@sxu.edu.cn, http://scit.sxu.edu.cn/SchoolTeacherD.aspx?id=326

Abstract: Time series data is an important object of data mining. In analysis of time series, misjudgment of correlation will occur if time lags are not considered. Therefore, there exists mutual restraint between correlation and time lags in time series. Based on the exploration of correlation and simultaneousness of time series, the correlation identification and curve registration methods for double sequences are given in this paper. Concretely, the study investigates the reasons and characteristics of two types of errors in correlation analysis in the view of time warping, and then deduces the correlation coefficient's bounds in a certain significance level by its asymptotic distribution. Further, a correlation identification method based on time-lag series is proposed. Finally, the curve registration model of maximizing the correlation coefficient is presented with a broader application than AISE. Smoothing-generalized expectation maximization (S-GEM) algorithm is used to solve the time warping function of the new model. The experimental results on simulated and real data demonstrate that the proposed correlation identification approach is more effective than 3 correlation coefficients and Granger causality test in recognition of spurious regression. The registration method provided is obviously performed better than the classical

* 基金项目: 国家自然科学基金(61273291, 71031006); 山西省回国留学人员科研基金(2012-008); 中国民航信息技术科研基地开放基金(CAAC-ITRB-201305)

收稿时间: 2014-01-20; 修改时间: 2014-04-22; 定稿时间: 2014-06-09

continuous monotone registration method (CMRM), Self-modeling registration (SMR) and maximum likelihood registration (MLR) in most situations. Linear correlation of double series and functional curve registration are considered here, and the results can provide the theoretical basis for correlation identification and time alignment in regression and reference direction for correlation analysis and curves registration of multiple series.

Key words: spurious regression; time warping; correlation; curve registration; curve smoothing

时序数据是数据挖掘中最常见的数据类型之一,在许多领域都有应用,如某河流的逐月径流量、当地月平均气温和降水量、我国的居民消费价格指数(CPI)和国内生产总值(GDP)、地震发生时多个观测点得到的波形序列等.对这些时序数据进行分析,可以得到一些有益的结论,如:通过研究河流历史流量、气温和降水量等特征,可以有效提高洪灾预测水平;利用 CPI 和 GDP,可以分析国家或地区的通胀程度和经济发展势头;根据多地地震波序列,可准确定位震源、震级等[1].有些非时序数据也可经过转化利用时序数据分析方法来处理,如:用树叶的边缘到质心的距离来描述其特征,取不同角度可得到一列数据,进而可判别树叶的类型[2].

在进行时序数据分析时,如果不考虑数据的时差,则很容易受直觉或偏见的影响,造成相关性的误判;但考虑不相关序列的时差是没有意义的.也就是说,判断序列相关性时需要考虑时差,既要考虑时差又要求数据具有相关性,所以序列之间的相关性和时差相互制约.目前,时序数据相关分析面临一些难题,如数据关系较为复杂、数据中含有噪声、缺失数据或异常数据等[3].同质数据(来源或属性相同的数据,如多地采集的同一地震的地震波数据)具有天然的相似性,不需要判断相关性,也不存在相关性和时差的制约问题,多用于分类或聚类.而对于异质数据(来源或属性不同的数据,如降水量与河流径流量、CPI 和 GDP)需要判断其相关性,若具有相关性再做回归分析等.因此,时序数据相关分析的主要对象是异质数据.

与二分类问题或假设检验中的两类错误类似,在对异质数据做相关性分析或数据回归时容易犯两类错误:

(1) 认为相关的数据不具有相关性;

(2) 认为不相关的数据具有相关性并做回归分析.

前者在实际应用中经常出现,比如太阳的仰角与地面温度、降水量与河流径流量,如果按照时间对照研究两组数据的相关性可能得到不相关的结论,但实际上,如果将两者在时间上进行平移,将出现高度相关性.上述第 2 类错误也会出现,如将过去 20 年中国的 GDP 与某人 20 岁之前的身高做回归,肯定显著正相关,而这其实是没有意义的、不合逻辑的回归,称为伪回归(nonsense regression 或 spurious regression).因此,在对数据进行分析之前,如果不考虑数据的相关性,第 1 类相关性错误会造成数据潜在信息的浪费,第 2 类相关性错误可能对后续的分析产生误导.多组数据的相关性,可以借助两组数据的相关性来给定.

第 1 类相关性错误发生的主要原因是两组时序数据发生时间弯曲(time warping),这时只要做时间上的转换,即可实现两者的协同一致.现实中,一般是非线性时间弯曲或动态时间弯曲(dynamic time warping,简称 DTW)[4,5],这就需要借助函数型数据分析(FDA)的方法,将时序数据转化为函数数据做时间校正,一般称为曲线排齐(curve registration)或曲线配准(curve alignment).Kneip 和 Gasser[6]将极值等作为特征点排齐(landmark registration),但对于特征点不明显的曲线不大适合,而且特征点的选取对结果影响较大.更一般化的方法是:确定一个目标函数或曲线,将其他曲线的局部特征与之对齐或最小化一些度量(如各曲线与目标曲线的均方距离)[7].Ramsay 等人[8]提出连续单调排齐方法(continuous monotone registration method,简称 CMRM),保证了时间弯曲函数的连续性和一致单调性.Wang 与 Gasser[9−11]提出一种基于动态时间弯曲模型的曲线排齐方法.Kneip 等人[12]采用关于时间弯曲函数的局部非线性调整方式来排齐曲线.Liu 和 Müller[13]在每个观测函数对应于潜在的二维随机过程的框架下,给出相应的随机排齐方法.Rønn[14],Gervini 和 Gasser[15]均采用非参数极大似然法(NPMLE)进行曲线排齐.James[16]提出基于所有曲线片段等同化的函数对齐方法.Liu 和 Yang[17]将曲线排齐和聚类融合起来,并用 EM 算法求解相应模型,直接得到排齐和聚类结果.较为常见的一种曲线排齐准则是最小化平均曲线误差平方积分(average integrated squared error,简称 AISE),其形式为

$$\min_{h_i(t)} \frac{1}{k}\sum_{i=1}^{k}\int_T \{x_i[h_i(t)]-x_0(t)\}^2 \mathrm{d}t,$$

其中,$x_0(t)$为基准函数,$x_i(t)$为需要排齐的函数,$h_i(t)$为各函数对应的时间弯曲函数,k 为排齐函数的个数.

以上这些曲线排齐方法多用于处理同质并发数据,但对于异质数据,可能因为量纲不同或负相关而对结果产生不利影响,所以并不适用.目前,关于专门处理异质数据的曲线排齐方法还未见公开报道.

第 2 类相关性错误(也称伪回归)实际上只是数据上的相关,并非现实的逻辑相关,其研究领域主要集中在计量经济方面.就一般的数据相关性问题,只能从分析伪回归的统计特点加以识别;至于数据在现实中的相关性,还需要根据所选择数据的背景知识去分析.导致伪回归的原因较多,Granger 和 Newbold[18]指出:对非平稳时间序列进行回归时,容易产生伪回归现象.但 Phillips[19]研究表明,平稳时间序列也可能出现类似错误.刘汉中[20,21]对平稳过程的伪回归进行研究,认为残差项未知形式的自相关,是导致伪回归的主要原因,而且残差项往往呈现出与数据过程阶数相同的自相关.一般回归中,如果具有很高的拟合度且具有很低的 DW(Durbin- Watson)统计量时,极有可能是伪回归.Jin 等人[22]指出:许多金融时序数据是具有厚尾特征的无限方差序列,此时,DW 统计量并非依概率收敛到零.现实数据的复杂多样性加剧了伪回归判断的难度.

同质数据具有天然的相似性,可直接做曲线排齐.对于异质时序数据相关性和时差的相互制约问题,本文固定时差判断各个时移序列的相关性,在序列相关的基础上,再通过曲线排齐细化时差函数.异质数据做相关性判断时,一方面由于样本相关系数与总体相关系数存在偏差,故研究总体相关系数的上下界;另一方面,为防止出现两类相关性错误,从其发生的主要原因出发,研究两类相关性错误的特征并提出相应的相关性判断方法.适用于异质数据的曲线排齐方法同样适用于同质数据,但适用于同质数据的准则(如 AISE)并不适用于异质数据(量纲不统一和负相关等).因此,主要根据异质数据的特点提出基于相关系数(绝对值)最大化的曲线排齐准则,并采用 S-GEM 算法求解.

1 曲线排齐相关分析方法

由于在解决实际问题中只能拿到样本数据,当使用样本估计总体时会产生偏差,因此,本文首先利用样本相关系数推断总体相关系数在一定显著性水平上的界;同时,为防止两类相关性错误的发生,本文研究了两种错误下时移序列相关系数的特征,并据此排除两类相关性错误.综合上述两个方面,可得到两组时序数据的相关性判定方法.

1.1 相关性判定

1.1.1 具有时间弯曲相关序列的相关性判定

为了判定序列相关性,需要推断总体相关系数的上下界.本文由关于样本相关系数的两个渐近分布得到总体相关系数在一定显著性水平的上下界,然后结合第 1 类相关性错误的特点,得到具有时间弯曲相关序列的相关性判定方法.

1) 相关系数的界

Pearson 相关系数是衡量序列相关性时最常用的度量方式.若有两组对应数据$\{(x_i,y_i),i=1,2,\ldots,n\}$($n$ 为样本量)是来自二元正态总体 $(x,y)\sim N(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$ 的样本,则样本相关系数为

$$\hat{\rho}(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\cdot\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}} \tag{1}$$

其中,$\bar{x},\bar{y}$ 分别为 X,Y 的样本均值.

样本相关系数 $\hat{\rho}(X,Y)$ 可以作为两个正态总体(X,Y)的相关系数ρ的无偏和一致估计量,但相关系数有一个明显的缺点,即它接近于 1 的程度与数据组数 n 相关,这容易给人一种假象.因为当 n 较小时,相关系数的波动较大,对有些样本相关系数的绝对值易接近于 1,特别是当 n=2 时,相关系数的绝对值总为 1;当 n 较大时,相关系数的绝对值容易偏小.有不少学者给出关于样本相关系数、样本量和二元正态总体相关系数的分布结果.

在(*X*,*Y*)为二元正态总体且ρ=0 的假设下,有如下分布:

$$T=\frac{\sqrt{n-2}\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sim t(n-2).$$

当$\rho=\rho_0$时,Fisher 给出了一种较为复杂的 $\hat{\rho}$ 的概率密度函数,适当变换后,得到如下渐近分布:

$$z=\frac{\phi(\hat{\rho})-\phi(\rho)}{2\sqrt{n-3}}\overset{n\to\infty}{\sim}N(0,1) \tag{2}$$

其中,$\phi(x)=\ln\frac{1+x}{1-x}$,当样本量较大时,可由样本相关系数对总体相关系数进行估计.

文献[23]证明了在二元正态总体中抽取 *n* 个样本,则有以下渐近分布:

$$\sqrt{n}(\hat{\rho}-\rho)\overset{n\to\infty}{\sim}N(0,(1-\rho^2)^2),\text{即}\frac{\sqrt{n}(\hat{\rho}-\rho)}{(1-\rho^2)}\overset{n\to\infty}{\sim}N(0,1) \tag{3}$$

本文分别依据上述两个渐近分布估计总体相关系数.

由于公式(2)中$\phi(x)$为单调增函数,可知:

- 当 $\rho\geqslant\hat{\rho}$ 时:

$$P\left\{\rho\leqslant\phi^{-1}\left[\phi(\hat{\rho})+2z_{1-\frac{\alpha}{2}}\cdot\sqrt{n-3}\right]\right\}=1-\alpha \tag{4}$$

- 当 $\rho\leqslant\hat{\rho}$ 时:

$$P\left\{\rho\geqslant\phi^{-1}\left[\phi(\hat{\rho})-2z_{1-\frac{\alpha}{2}}\cdot\sqrt{n-3}\right]\right\}=1-\alpha \tag{5}$$

其中,$\phi^{-1}(x)=\frac{e^x-1}{e^x+1}$;$Z_\alpha$为标准正态分布的$\alpha$分位点,即 $P(x\leqslant Z_\alpha)=\alpha$;随机变量 $x\sim N(0,1)$.

本文在公式(3)的基础上进一步推断总体相关系数的界,即:

- 当 $\rho\geqslant\hat{\rho}$ 时:

$$P\left\{\frac{-\sqrt{n}-\sqrt{n+4\sqrt{n}z_{1-\frac{\alpha}{2}}\cdot\hat{\rho}+4z^2_{1-\frac{\alpha}{2}}}}{2z_{1-\frac{\alpha}{2}}}\leqslant\rho\leqslant\frac{-\sqrt{n}+\sqrt{n+4\sqrt{n}z_{1-\frac{\alpha}{2}}\cdot\hat{\rho}+4z^2_{1-\frac{\alpha}{2}}}}{2z_{1-\frac{\alpha}{2}}}\right\}=1-\alpha \tag{6}$$

- 当 $\rho\leqslant\hat{\rho}$ 时:

$$P\left\{\frac{\sqrt{n}-\sqrt{n-4\sqrt{n}z_{1-\frac{\alpha}{2}}\cdot\hat{\rho}+4z^2_{1-\frac{\alpha}{2}}}}{2z_{1-\frac{\alpha}{2}}}\leqslant\rho\leqslant\frac{\sqrt{n}+\sqrt{n-4\sqrt{n}z_{1-\frac{\alpha}{2}}\cdot\hat{\rho}+4z^2_{1-\frac{\alpha}{2}}}}{2z_{1-\frac{\alpha}{2}}}\right\}=1-\alpha \tag{7}$$

综合公式(4)~公式(7),当α=0.05 时,有如下近似:

- 当 $\rho\geqslant\hat{\rho}$ 时:

$$\inf_{\alpha=0.05}\rho=\max\left\{\frac{-\sqrt{n}-\sqrt{n+8\sqrt{n}\cdot\hat{\rho}+16}}{4},\hat{\rho},-1\right\} \tag{8}$$

$$\sup_{\alpha=0.05}\rho=\min\left\{\frac{-\sqrt{n}+\sqrt{n+8\sqrt{n}\cdot\hat{\rho}+16}}{4},\phi^{-1}[\phi(\hat{\rho})+4\sqrt{n-3}],1\right\} \tag{9}$$

- 当 $\rho\leqslant\hat{\rho}$ 时:

$$\inf_{\alpha=0.05}\rho=\max\left\{\frac{\sqrt{n}-\sqrt{n-8\sqrt{n}\cdot\hat{\rho}+16}}{4},\phi^{-1}[\phi(\hat{\rho})-4\sqrt{n-3}],-1\right\} \tag{10}$$

$$\sup_{\alpha=0.05}\rho=\min\left\{\frac{\sqrt{n}+\sqrt{n-8\sqrt{n}\cdot\hat{\rho}+16}}{4},\hat{\rho},1\right\} \tag{11}$$

图 1 为在不同样本规模 n 和样本相关系数下总体相关系数的上下界.从图中可以看出:本文给出的上下界曲线具有以下特点:

(1) 样本规模越大,上下界越紧凑;

(2) 样本规模相同时,上下界曲线中心对称;

(3) 相关系数绝对值越大,上下界越紧凑.

以上特点容易由公式(8)~公式(11)证明.

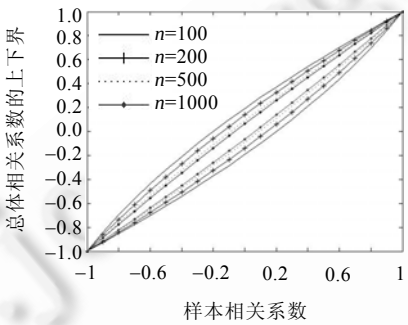


Fig.1 Population correlation coefficient's bounds (significance level α=0.05)

图 1 总体相关系数的界(显著性水平α=0.05)

2) 相关性判定方法

为方便描述序列的相关特征,首先给出时移序列(time-lag series)的定义.

假设两序列$(X,Y)=\{(x_i,y_i),i=1,2,\ldots,n\}$,定义如下序列为时移序列:

$$\begin{aligned}(X_t,Y_{t+m})&=\{(x_i,y_{i+m}),i=1,2,...,n-m\},1\leqslant m<n,m\in N^+\\(X_t,Y_{t-m})&=\{(x_i,y_{i-m}),i=m+1,2,...,n\},1\leqslant m<n,m\in N^+\end{aligned} \tag{12}$$

对于第 1 类回归错误,即认为相关的序列不具有相关性,直接考虑初始序列,其相关性必然较小;如果考虑其时移序列的相关性,则一定存在 $m_0(1\leqslant|m_0|<<n,|m_0|\in N^+)$使得 (X_t,Y_{t+m_0}) 的相关系数较大.

由此得到具有时间弯曲相关序列的相关性判定方法:若时移序列的相关系数随 m 的变化而变化,并在 m_0 处达到最大值,即曲线图 $\rho=\hat{\rho}(m)$ 呈现明显的上凸现象,可根据公式(8)~公式(11)估计总体相关系数的范围.如果$|\rho(m_0)|>\rho_0$(即超过给定阈值如 0.6),认为时移序列 (X_t,Y_{t+m_0}) 存在相关性,可做曲线排齐及回归分析等.

1.1.2 伪回归的相关性判定

在给出伪回归的相关性判定之前,本文首先研究伪回归的主要原因及其时移序列的相关系数特点,并根据这种特点做出针对性的判断.前面提到,残差项往往呈现出与数据过程阶数相同的自相关,可见,多数情况下伪回归是由残差的自相关性或序列的自相关性引起的.本文分别以序列一阶自相关和残差一阶自相关为前提给出伪回归的相关性判定方法.

(1) 自相关序列的相关性判定

假设两序列$(X,Y)=\{(x_i,y_i),i=1,2,\ldots,n\}$均是平稳一阶自相关序列,即

$$\begin{cases}X_{t+1}=c+aX_t+\varepsilon_{X_t}\\Y_{t+1}=d+bY_t+\varepsilon_{Y_t}\end{cases} \tag{13}$$

其中,a,b,c,d 均为常数; ε_{X_t} 和 ε_{Y_t} 是相互独立的正态随机变量, $\varepsilon_{X_t}\sim N(0,\sigma_X^2),\varepsilon_{Y_t}\sim N(0,\sigma_Y^2)$.

由公式(1)得(X_t,Y_t)的相关系数为

$$\hat{\rho}(X_t,Y_t)=\frac{E(X_tY_t)-E(X_t)\cdot E(Y_t)}{\sqrt{D(X_t)\cdot D(Y_t)}}=\frac{Cov(X_t,Y_t)}{\sqrt{D(X_t)\cdot D(Y_t)}} \tag{14}$$

其中,$E(\cdot),D(\cdot),Cov(\cdot)$分别表示期望、方差和协方差函数.

时移序列(X_t,Y_{t+1})的相关系数为

$$\begin{aligned}\hat{\rho}(X_t,Y_{t+1})&=\frac{E(X_tY_{t+1})-E(X_t)\cdot E(Y_{t+1})}{\sqrt{D(X_t)\cdot D(Y_{t+1})}}\\&=\frac{E(dX_t+bX_tY_t+X_t\varepsilon_{Y_t})-E(X_t)\cdot E(d+bY_t+\varepsilon_{Y_t})}{\sqrt{D(X_t)\cdot D(d+bY_t+\varepsilon_{Y_t})}}\\&=\frac{b[E(X_tY_t)-E(X_t)\cdot E(Y_t)]}{\sqrt{D(X_t)\cdot[D(bY_t)+D(\varepsilon_{Y_t})]}}\\&=\frac{b}{|b|}\cdot\frac{E(X_tY_t)-E(X_t)\cdot E(Y_t)}{\sqrt{D(X_t)\cdot\left[D(Y_t)+\frac{\sigma_Y^2}{b^2}\right]}}\end{aligned} \tag{15}$$

实际问题中,一般 $b>0$,此时,一阶时移序列的相关系数与原序列的相关系数之比为

$$\frac{\hat{\rho}(X_t,Y_t)}{\hat{\rho}(X_t,Y_{t+1})}=\sqrt{1+\frac{\sigma_Y^2}{b^2D(Y_t)}} \tag{16}$$

从公式(16)可以看出:当 $\sigma_Y^2/b^2D(Y_t)$ 较小时,时移序列相关系数与原序列相关系数差异不大;随机项 ε_{Y_t} 的方差越小,Y_t的方差越大,这种差异越不明显.一般情况下,$\sigma_Y^2<<D(Y_t)$,否则,Y_t接近随机序列.因此,一阶自相关序列的时移序列相关系数随 m 变化不大.

由上述自相关序列的时移序列相关系数特点得到如下伪回归相关性判别方法:假设时间序列(X_t,Y_t)的时移序列为(X_t,Y_{t+m}),其相关系数随 m 的改变无明显变化(如 $\max|\hat{\rho}(m_i)-\hat{\rho}(m_j)|<0.1$),则认为两序列无相关性.即便其相关系数很高,做出的回归也是伪回归.

(2) 残差自相关序列的相关性判定

假设两序列$(X,Y)=\{(x_i,y_i),i=1,2,\ldots,n\}$的回归模型如下:

$$Y_t=\theta X_t+\varepsilon_t \tag{17}$$

其中,θ为常数;残差序列ε_t一阶自相关,即

$$\varepsilon_{t+1}=c\varepsilon_t+\mu_t \tag{18}$$

其中,c 为残差序列自相关系数;μ_t为白噪声序列,且 $\mu_t\sim N(0,\sigma_\mu^2)$.

由公式(17)、公式(18)可得 $E(Y_{t+1}-Y_t)=c\theta\cdot E(X_{t+1}-X_t)$,根据归纳法知 $E(Y_{t+m}-Y_t)=(c\theta)^m\cdot E(X_{t+m}-X_t)$,当然有:

$$E(Y_{1+m}-Y_1)=(c\theta)^m\cdot E(X_{1+m}-X_1).$$

若令 $Y_t'=Y_t-Y_1,X_t'=X_t-X_1$,则 $E(Y_t')=(c\theta)^{t-1}\cdot E(X_t')$,故有:

$$\begin{cases}Y_t'=(c\theta)^{t-1}\cdot X_t'+\varepsilon_t'\\Y_{t+1}'=(c\theta)^t\cdot X_{t+1}'+\varepsilon_{t+1}'\end{cases}.$$

如果 X_t具有自相关性,则 X_t' 也有自相关性,设为 $X_{t+1}'=c'X_t'+\varepsilon_{X_t}'$,则:

$$Y_{t+1}'=(c\theta)^t\cdot(c'X_t'+\varepsilon_{Xt}')+\varepsilon_{t+1}'=c'(c\theta)^t\cdot X_t'+(c\theta)^t\varepsilon_{X_t}'+\varepsilon_{t+1}'=c'c\theta\cdot Y_t'-c'c\theta\varepsilon_t'+(c\theta)^t\varepsilon_{X_t}'+\varepsilon_{t+1}'.$$

即:

$$E(Y_{t+1}')=c'c\theta\cdot E(Y_t'),E(Y_{t+1})=c'c\theta\cdot E(Y_t)+(1-c'c\theta)Y_1.$$

上述推导表明:当残差序列一阶自相关时,若有一个序列一阶自相关,则另外一个序列也必定一阶自相关.这与文献[21]中的结论"如果数据过程的生成机制都是自回归模型时,随机干扰项往往呈现出与数据过程阶数相同的自相关"相吻合.由自相关序列的特点知,残差自相关序列的时移序列相关系数对 m 的变化不敏感,因此,残差自相关序列的相关性判定方法可参照自相关序列的相关性判定方法.

由第 1.1.1 节和第 1.1.2 节的分析,本文关于相关性判定方法可总结如下:

1) 如果时移序列的相关系数都比较小,则两序列无相关性或相关性不明显;
2) 如果时移序列的相关系数都比较大,但变化不明显,则两序列无相关性或相关性不明显,即出现伪回归现象;
3) 如果时移序列的相关系数出现明显上凸现象(正相关),且在极值点处总体相关系数(由样本相关系数推断)比较大,则两序列为具有时间弯曲的相关序列.

1.2 曲线排齐方法

通过时移序列的相关系数,可以判定序列间是否具有相关性.若两序列存在相关性,但具有时间偏差,就需要曲线排齐方法将其对齐,消除相位上(时间轴上)的差异.对于异质数据,当采用 AISE 准则时,其结果会随量纲变化而变化,因此,需要提出一种无量纲的准则来排齐异质数据组成的曲线.

1.2.1 基于函数型相关系数的曲线排齐模型

Pearson 相关系数是一种描述两序列相关性或相似性的无量纲度量方式,但仅适用于描述离散数据的相关性.连续函数的相关性可用内积表示,同时,为使内积取值规范化,再除以两个函数的范数.异质数据组成的曲线排齐准则可通过函数型相关系数来构建:

$$\max_{h(t)} |\rho(x_1^*, x_2)| = \max_{h(t)} \left| \frac{\int_T x_1^*(s)x_2(s)\mathrm{d}s}{\sqrt{\int_T [x_1^*(s)]^2\mathrm{d}s}\sqrt{\int_T [x_2(s)]^2\mathrm{d}s}} \right| \tag{19}$$

其中, $x_1^*(t) = x_1[h(t)]$ 表示排齐后的函数.鉴于连续型函数的运算复杂性以及优化准则的高维特征,本文给出对应的离散化形式.

假设两个函数型数据 $x_1(t)$和 $x_2(t)$在采样时间点 $T=(t_1,t_2,\ldots,t_n)$处的样本序列为 $x_1(T)=[x_1(t_1),x_1(t_2),\ldots,x_1(t_n)]$和 $x_2(T)=[x_2(t_1),x_2(t_2),\ldots,x_2(t_n)]$,现要将函数 $x_1(t)$对照 $x_2(t)$做曲线排齐.令$\Delta=(\delta_1,\delta_2,\ldots,\delta_n)$为在时间点 T 处 $x_1(t)$相对于 $x_2(t)$的偏移量,即时间弯曲函数满足 $h(T)=T+\Delta$,则经过排齐的时序样本变为 $x_1(T+\Delta)=[x_1(t_1+\delta_1),\ldots,x_1(t_n+\delta_n)]$,排齐后的两组函数型数据的样本序列应当具有较高的相关性.曲线排齐问题可转化为求解:

$$\max_{\Delta} |\rho[x_1(T+\Delta), x_2(T)]|.$$

一般时间弯曲函数具有一致单调性,即满足 $t_{i-1}+\delta_{i-1}<t_i+\delta_i<t_{i+1}+\delta_{i+1}$.然而,偏移向量杂乱无序会造成时间弯曲函数不满足一致单调性,故限定 $\delta_i^{k+1} \in (bndl_i, bndr_i)$,其中, $bndl_i = t_{i-1} + \delta_{i-1}^k - t_i, bndr_i = t_{i+1} + \delta_{i+1}^k - t_i, \delta_i^k$,表示第 k 次迭代时δ_i 的值.具体实现时,可将 δ_i^{k+1}的搜索区间缩小为闭区间$[bndl_i+p\cdot(bndr_i-bndl_i), bndr_i-p\cdot(bndr_i-bndl_i)]$,其中,$p$ 为(0,0.5)内的常数.

最终,曲线排齐问题变为求解下面的约束优化问题:

$$\begin{cases} \Delta^* = \arg\max_{\Delta} |\rho[x_1(T+\Delta), x_2(T)]| \\ \text{s.t. } \delta_i \in [bndl_i + p\cdot(bndr_i - bndl_i), bndr_i - p\cdot(bndr_i - bndl_i)] \end{cases} \tag{20}$$

最后,将时间偏移向量Δ^*转化为函数形式,得到时间偏移函数 $d(t)$,相应的时间弯曲函数为 $h(t)=d(t)+t$.

1.2.2 模型求解——S-GEM 算法

文献[1,17]中采用 EM 算法求解曲线排齐优化问题,但当参数的维数较高时,难以求解 Q 函数的极大化问题.为克服此问题,本文将问题(20)的目标函数作为 Q 函数(即 EM 算法中对数似然函数的期望),将推广的 EM 算法(广义期望最大化算法 GEM)用于求解式(20);由于时间弯曲函数和时间偏移函数具有较好的光滑性,为加快其收敛速度,每进行完一次Δ^*的更新,都做一次 P 样条光滑处理,以增强时差向量的光滑性,而且 P 样条具有正则项,可防止过度优化造成的时差函数不稳定问题.因此得到求解模型(20)的光滑广义期望最大化方法(S-GEM),其主要步骤如下:

输入:两组在时间 $T_0=(t_{01},t_{02},\ldots,t_{0m})$上具有时间弯曲的相关时序数据 TS_1, TS_2;

Step 1. 初始化时差向量$\Delta_0=zeros(1,n)$,迭代容许误差 eps.

Step 2. 时序数据函数化.将 TS_1,TS_2 转化为函数型数据 $x_1(t)$和 $x_2(t)$,并在 T_0 内均匀取 n 个点 $T=(t_1,t_2,\ldots,t_n)$,得到光滑序列$\{x_1(t_i)\}$和$\{x_2(t_i)\}$(i=1,2,…,n),其中,$t_1=t_{01}$,$t_n=t_{0m}$.

Step 3. 用广义期望最大化求时差向量.记第 k 次迭代时差向量为 $\Delta^k=(\delta_1^k,\delta_2^k,...,\delta_n^k)$,进行 n–2 次条件极大化(假定起始点无时差,即 $\delta_1^k=\delta_n^k=0$),得到 $\Delta^{k+1}=(\delta_1^{k+1},\delta_2^{k+1},...,\delta_n^{k+1})$.

Step 4. 时差向量的光滑处理:采用 P 样条拟合关于序列Δ^{k+1} 的函数 $d^{k+1}(t)$,并将拟合值替代原始值,即:

$$\Delta^{k+1}=d^{k+1}(t_i),i=1,2,\ldots,n.$$

Step 5. 重复 Step 3 和 Step 4,直到收敛($|\Delta^{k+1}-\Delta^k|<eps$).

输出:时差函数 $d(t)=d^K(t)$或时间弯曲函数 $h(t)=d(t)+t$.

与很多函数型数据分析方法类似,本文提出的算法可处理数据量较大的数据排齐问题,即使存在数据缺失或异常数据,也能充分利用当前信息;另外,算法经过光滑处理,能够快速收敛到极值.重要的是,算法的运行时间或时间复杂度主要依赖于抽样个数,而与原始样本个数无关.S-GEM 算法的复杂度分析见表 1.

Table 1 Complexity of S-GEM algorithm

表 1 S-GEM 算法的复杂度分析

步骤	时间复杂度	空间复杂度
Step 1	$O(n)$	$O(n)$
Step 2	函数化: $O[(m+d-2)^3]$ 取值: $O(nd^2)$	函数化: $O[(m+d-2)^2]$ 取值: $O(n)$
Step 3	$O[n\times fm]$	$O(n)$
Step 4	光滑: $O[(n+d-2)^3]$ 取值: $O(nd^2)$	光滑: $O[(n+d-2)^2]$ 取值: $O(n)$
总体	$O[(m+d-2)^3+k((n+d-2)^3+n\times fm)]$	$O[(m+d-2)^2+(n+d-2)^2]$

其中,m 为初始样本量,n 为均匀采样量,d 为函数化的阶数(最高次数减 1),fm 为 Step 3 中条件极大化的平均时间复杂度,k 为总体迭代次数.由于 fm 和 k 与精度、参数和问题本身有关,故将其单独设出.数据函数化或光滑时采用 P 样条,样条函数的系数向量估计时最复杂的部分在样条基函数矩阵求逆,时间复杂度和空间复杂度分别为基函数个数的 3 次和 2 次.一般,m^3 为有限计算量,相对于迭代过程可略去;d 取值较小,本文实验中 d=4,此时算法时间复杂度和空间复杂度分别为 $O[k(n^3+n\times fm)]$和 $O[m^2+n^2]$.

2 实验结果与分析

本文在模拟数据(7 种人造时序数据和两组 Sinc 函数)和真实数据(先行指数与一致指数,GDP 与城镇就业人数、城镇居民家庭人均可支配收入)上对所提出的相关性判定方法及曲线排齐方法进行验证.对于上述存在时间弯曲的相关数据,采用本文方法做了曲线排齐,一方面分析方法对参数的敏感性,另一方面与已有方法进行对比分析.

2.1 相关性判断

2.1.1 模拟数据

(1) 伪回归判定

关于伪回归判定问题,选择时序数据的 7 种主要数据产生过程(data generating process,简称 DGP)(见表 2),分别通过时移序列的相关系数来判别伪回归,表 3 为伪回归检测中设置的参数值.

Table 2 Data generating process

表 2 数据产生过程

模型序号	名称	模型
1	AR(1): (无漂移)一阶自回归	$X_t=a_1X_{t-1}+\varepsilon_{X_t}$
2	AR(1)+Drift: 带漂移一阶自回归	$X_t=c+a_1X_{t-1}+\varepsilon_{X_t}$
3	I(1): (无漂移)一阶单整	$X_t=X_{t-1}+\varepsilon_{X_t}$

Table 2 Data generating process (Continued)

表 2 数据产生过程(续)

模型序号	名称	模型
4	I(1)+Drift: 带漂移一阶单整	$X_t = c + X_{t-1} + \varepsilon_{X_t}$
5	IMA(1,1): 一阶单整移动平均	$X_t = X_{t-1} + u_1\varepsilon_{X_{t-1}} + \varepsilon_{X_t}$
6	ARMA(1,1): 自回归移动平均(1,1)	$X_t = a_1X_{t-1} + u_1\varepsilon_{X_{t-1}} + \varepsilon_{X_t}$
7	ARMA(2,2): 自回归移动平均(2,2)	$X_t = a_1X_{t-1} + a_2X_{t-2} + u_1\varepsilon_{X_{t-1}} + u_2\varepsilon_{X_{t-2}} + \varepsilon_{X_t}\sigma^2$

Table 3 Parameters in spurious regression detection

表 3 伪回归检测中设置的参数值

模型	变量	a_1	a_2	c	u_1	u_1	σ^2
1	X_t	1					1
	Y_t	1.01					1
2	X_t	0.98		5			1
	Y_t	1		3			9
3	X_t	1					1
	Y_t	1					1
4	X_t	1		2			25
	Y_t	1		1			25
5	X_t	1			0.5		1
	Y_t	1			0.5		1
6	X_t	1.05			0.2		1
	Y_t	1.01			0.2		1
7	X_t	0.5	0.5		1	1	1
	Y_t	0.3	0.7		0.5	2	1

首先采用常规相关性判别方法分析上述 7 组序列的相关性,其结果为表 4 和表 5.

Table 4 Correlation coefficients of series with different models

表 4 各模型序列的相关系数

模型	Pearson 线性相关系数	Spearman 秩相关系数	Kendall 秩相关系数
1	–0.877	–0.811	–0.598
2	0.861	0.997	0.961
3	0.060	0.000	0.012
4	0.922	0.950	0.821
5	0.046	–0.014	0.005
6	0.893	0.951	0.849
7	0.828	0.987	0.940

Table 5 Granger causality test of series with different models

表 5 各模型序列的 Granger 因果检验

模型	单整阶数		Granger 因果检验(*Lag*=0)			
	X	Y	因→果	F 统计量	显著性	是否 Granger 原因
1	1	1	$Y\to X$	0.006	0.993	否
			$X\to Y$	0.287	0.750	否
2	0	1	$Y\to X$	1.185	0.307	否
			$X\to Y$	0.019	0.980	否
3	1	1	$Y\to X$	0.412	0.662	否
			$X\to Y$	0.129	0.878	否
4	1	1	$Y\to X$	0.412	0.662	否
			$X\to Y$	0.129	0.878	否
5	1	1	$Y\to X$	0.513	0.599	否
			$X\to Y$	0.164	0.848	否
6	4	1	$Y\to X$	6.463	0.002	是
			$X\to Y$	1.799	0.168	否
7	4	1	$Y\to X$	4.035	0.019	是
			$X\to Y$	16.653	0.000	是

由表 4 可见:模型 3 和模型 5 的相关系数比较小,不会出现伪回归现象.其他各组自相关序列的 3 种相关系数均比较高,表明常规相关系数对于自相关序列的伪回归现象不具有鉴别能力.由表 5 可见:Granger 因果检验能够识别多数序列的相关性,但仍有 3 处识别错误.其中,模型 7 为二阶自相关序列,双向 Granger 因果检验均错误.因此对于简单模型,Granger 因果检验能识别伪回归;当模型较复杂时,不能识别伪回归.

由于上述 3 种相关系数和 Granger 因果检验均不能较好地判别自相关序列的相关性,后续实验中不再使用这些方法判别相关性.

图 2 为 7 种模型的时移序列相关系数变化图.由图 2 可见:模型 3 和模型 5 的时移序列相关系数绝对值较小,不会造成伪回归.其他 5 个模型的相关系数较高,但时移序列的相关系数随 m 变化不大,采用本文方法可以快速准确地识别伪回归.虽然本文给出一阶自相关序列的伪回归判别方法,但由模型 7 的结果可见:对于二阶自相关序列,同样能够得到可靠的相关性判定结果.

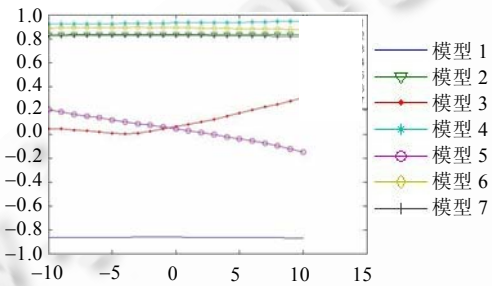
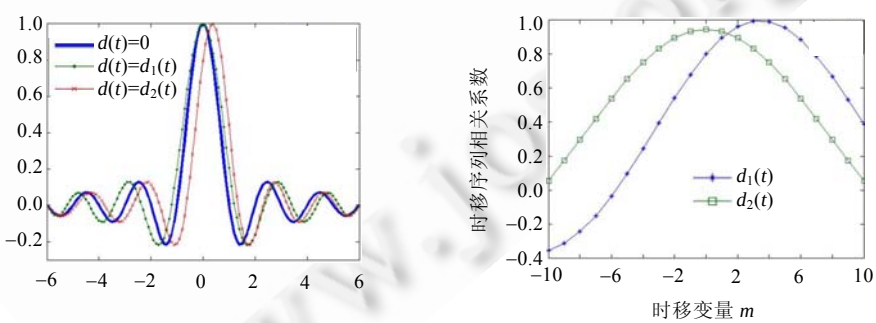


Fig.2 Correlation coefficients in time-lag series

图 2 时移序列相关系数变化图

(2) 时间弯曲序列相关性判定

选择具有波动性的 Sin*c* 函数(Sin*c*(x)=sin$\pi x/\pi x$,$x\in[-6,6]$)作为模拟数据相关性研究对象,并且做如下两种时差函数:$d_1(t)=0.01t^2-0.36$,$d_2(t)=0.005t(t-6)(t+6)$.两种情况下,与标准 Sin*c* 函数及时移序列相关系数变化趋势如图 3 所示.



(a) Sin*c* 函数与两种时差下的函数 (b) 两种时差下的时移序列相关系数变化情况

Fig.3 Sin*c* functions and correlation coefficients in their time-lag series

图 3 Sin*c* 函数和时移序列相关系数变化图

从图 3(b)中可以看出:两条时移序列相关系数曲线都有明显的上凸现象,且两个相关系数的界分别为[0.991,0.996]和[0.914,0.962].由此可判定两组序列存在相关性,且时差函数分别为 $d_1(t)$和 $d_2(t)$的序列与标准 Sin*c* 函数序列的平均滞后量分别为 0 和 3.

2.1.2 真实数据

(1) GDP 与城镇就业人数、城镇居民家庭人均可支配收入

本文收集了 1980 年~2011 年国家统计局公布的年度数据[24]:国内生产总值(GDP/亿元)、城镇就业人数(万人)、城镇居民家庭人均可支配收入(简称可支配收入/元).两两之间时移序列的相关性如图 4 所示.

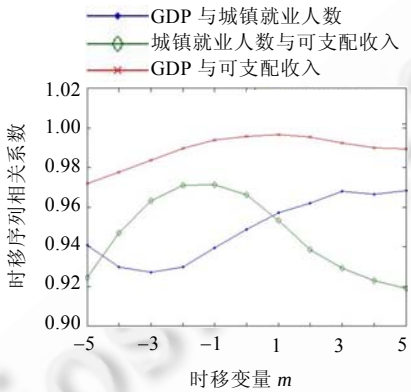


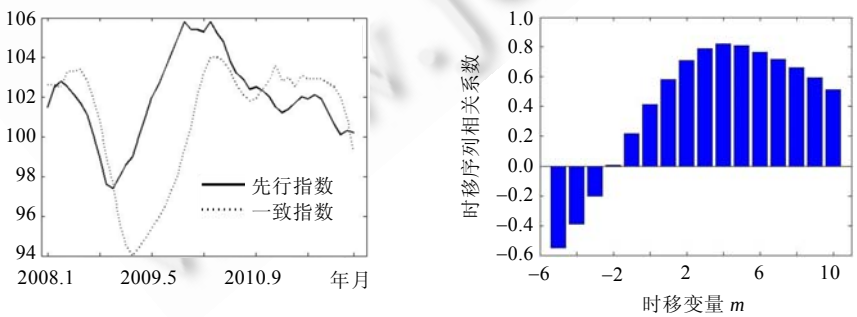
Fig.4 Correlation coefficients in time-lag series

图 4 时移序列相关系数变化图

由图 4 可见:GDP 与城镇就业人数、GDP 与可支配收入之间的相关系数很高,但其时移序列相关系数变化不大,故 GDP 与城镇就业人数及可支配收入无相关性;而城镇就业人数与可支配收入相关系数很高,在 m=−1 处达到最大值 0.97,并且估计得到相关系数的界为[0.85,0.98],符合上凸条件且相关性较大,故判定两者存在相关性,且城镇就业人数平均比可支配收入滞后一期(一年).这一结果与大多数“就业水平与人均收入存在动态均衡关系”的结论是一致的,并且说明收入的增加能带动就业人数的增长.

(2) 先行指数与一致指数

本文选择的 2008 年~2011 年期间先行指数与一致指数的月度数据同样来自文献[24].由于先行指数和一致指数分别由不同的经济指标合成,因此两者属于异质数据.图 5 为两个指数及时移序列相关系数变化图.从图 5(b)中可以看出:时移序列相关系数曲线有明显的上凸现象,在 m=4 处达到最大值 0.82,并且估计得到相关系数的界为[0.655,0.883],故判定两组序列存在相关性,且一致指数平均比先行指数滞后 4 个月.中国经济景气监测中心指出,2004 年 3 月先行指数领先 3 个月.本文结果表明:受金融危机影响,2008 年~2011 年期间,先行指数的平均领先期数略微增大.



(a) 先行指数与一致指数 (b) 时移序列相关系数变化情况

Fig.5 Leading index, coincident index and correlation coefficients in their time-lag series

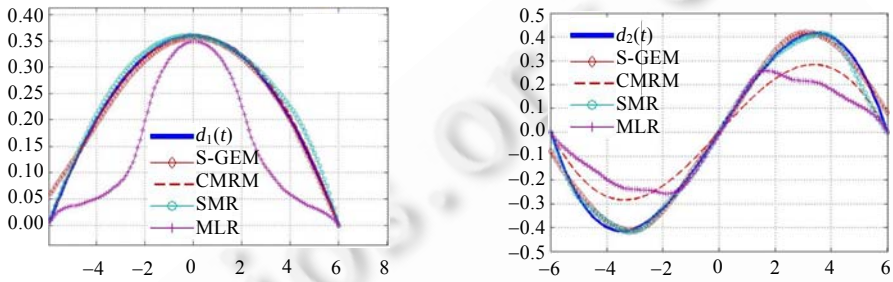
图 5 先行指数与一致指数及时移序列相关系数变化图

2.2 曲线排齐

本节主要对 S-GEM 算法的性能进行测试,并与经典的 CMRM 算法[8]、极大似然排齐(maximum likelihood registration,简称 MLR)[15]和自模型排齐(self-modeling registration,简称 SMR)[25]进行比较.为公平起见,CMRM 算法和 MLR 的结果为 5 次运行的平均结果.实验的机器配置为:Intel 四核 CPU(主频为 2.83GHz),3G 内存.

2.2.1 模拟数据

将时差函数分别为 $d_1(t)=0.01t^2-0.36$ 和 $d_2(t)=0.005t(t-6)(t+6)$的含噪声的 Sinc 函数与标准 Sinc 函数做曲线排齐.在两种时差函数下,4 种排齐方法经过调参后的排齐效果如图 6 所示.



(a) 时差函数 $d_1(t)$ (b) 时差函数 $d_2(t)$

Fig.6 Registration results of 4 methods

图 6 4 种方法的排齐结果

由图 6 可见:时差函数为 $d_1(t)$时,MLR 排齐效果较差,其他 3 种比较接近实际值;时差函数为 $d_2(t)$时,MLR 和 CMRM 排齐效果较差,SMR 和 S-GEM 非常接近时差函数.

表 6~表 9 为 Sinc 函数在两种时差函数下分别采用 CMRM,MLR,SMR,S-GEM 的排齐结果.

从表 9 中可以看出:总体上,S-GEM 算法的精度主要与采样点数密切相关,而且时差函数越复杂,需要的采样点数越多;S-GEM 算法的运行时间主要随采样点数的增加而变长.

对比表 6~表 9 的结果可见:

- 时差函数为 $d_1(t)$时,CMRM 精度最高,但效率最差;SMR 和 S-GEM 的精度和效率都差不多,但 SMR 结果不如 S-GEM 稳定;MLR 的精度最差,效率不高;
- 时差函数取 $d_2(t)$时,SMR 和 S-GEM 精度较高,但 SMR 效率远不及 S-GEM,而且不如 S-GEM 稳定.

以上实验表明:对于简单的时差函数,CMRM,SMR 和 S-GEM 的精度都比较高,但 CMRM 效率较差;当时差函数较复杂时,SMR 和 S-GEM 排齐结果较好,但在效率和稳定性上 SMR 不如 S-GEM.

Table 6 Registration results of Sinc function by CMRM (5 times on average)

表 6 Sinc 函数 CMRM 排齐结果(5 次平均)

时差函数	运行时间(s)	RMSE
$d_1(t)$	20.32	**0.003 8**
$d_2(t)$	30.74	**0.089 1**

Table 7 Registration results of Sinc function by MLR (5 times on average)

表 7 Sinc 函数 MLR 排齐结果(5 次平均)

时差函数	参数		运行时间(s)	RMSE
	尺度参数	最大迭代次数		
$d_1(t)$	0.1	50	**14.60**	**0.108 6**
	0.1	100	22.24	0.146 2
$d_2(t)$	0.1	50	14.51	0.150 4
	0.1	100	**21.83**	**0.133 7**

Table 8 Registration results of Sinc function by SMR

表 8 Sinc 函数 SMR 排齐结果

时差函数	实验序号	参数		运行时间(s)	RMSE
		成分个数	基函数个数		
$d_1(t)$	1	4	7	5.49	0.060 6
	2	3	8	**6.87**	**0.011 8**
	3	3	8	6.05	0.013 3
	4	4	7	5.65	0.063 2
	5	3	7	6.27	0.051 2
	平均	3.4	7.4	6.06	0.040 0
$d_2(t)$	1	4	12	**21.99**	**0.020 8**
	2	4	11	25.10	0.047 4
	3	5	13	24.56	0.067 4
	4	4	12	23.82	0.061 1
	5	4	15	23.00	0.088 8
	平均	4.2	12.6	23.69	0.057 1

Table 9 Registration results of Sinc function by S-GEM

表 9 Sinc 函数 S-GEM 算法的排齐结果

时差函数	迭代误差 *eps*	采样点数 *n*	运行时间(s)	RMSE
$d_1(t)$	0.01	10	0.15	0.261 8
		30	4.29	0.029 3
		50	**7.76**	**0.011 9**
		80	16.52	0.013 5
		100	25.47	0.016 9
	0.05	10	1.13	0.083 5
		30	3.2	0.018 0
		50	**7.96**	**0.012 0**
		80	17.15	0.023 9
		100	24.87	0.024 5
	0.1	10	0.15	0.261 8
		30	3.13	0.030 1
		50	**7.33**	**0.019 1**
		80	15.19	0.023 2
		100	19.72	0.029 8
$d_2(t)$	0.01	10	0.15	0.296 9
		30	4.74	0.035 6
		50	8.04	0.029 7
		80	**16.61**	**0.026 9**
		100	27.82	0.034 0
	0.05	10	0.15	0.296 9
		30	4.09	0.033 7
		50	**8.15**	**0.022 4**
		80	16.78	0.026 1
		100	22.75	0.028 8
	0.1	10	0.15	0.296 9
		30	3.31	0.037 9
		50	7.09	0.032 4
		80	**14.73**	**0.026 5**
		100	20.60	0.037 4

2.2.2 真实数据

表 10 为先行指数对照一致指数做曲线排齐的结果.由于真实的时差函数未知,对于真实数据,采用相关系数比较两种方法的效果.由表 10 可见,S-GEM 算法主要对采样点数敏感.即采样点过少,效果变差,但相关系数仍比其余 3 种方法要大;采样点过多,运算量大,但 S-GEM 算法无论从结果还是效率上都优于 CMRM 算法.SMR 和 MLR 虽然运行时间短,但相关系数较小.

Table 10 Curve registration results of leading index and coincident index

表 10 先行指数与一致指数排齐结果

方法	采样点数 *n*	迭代误差 *eps*	运行时间(s)	相关系数
S-GEM	20	0.1	6.32	0.980 5
		0.2	6.35	0.980 5
		0.3	6.40	0.980 5
		0.4	16.22	0.980 8
		0.5	18.16	0.980 9
	30	0.1	9.27	0.999 5
		0.2	9.04	0.999 5
		0.3	8.32	0.999 5
		0.4	8.01	0.999 5
		0.5	7.88	0.999 5
	40	0.1	16.96	0.999 4
		0.2	15.47	0.999 5
		0.3	15.86	0.999 5
		0.4	15.47	0.999 5
		0.5	15.33	0.999 5
CMRM (5 次平均)			32.31	0.943 5
SMR (5 次平均)			11.77	0.873 9
MLR (5 次平均)			3.78	0.856 6

图 7 为 S-GEM 算法(*n*=30,*eps*=0.3)与其他 3 种方法的排齐效果.为方便观察,图中给出了 4 个极值点处的虚线.可以观察到:在第 2 个极值点处,只有 MLR 不能对齐;在第 1、第 3 极值点处,SMR 和 MLR 稍逊于 S-GEM 和 CMRM;在第 4 个极值点处,只有 S-GEM 能够对齐.

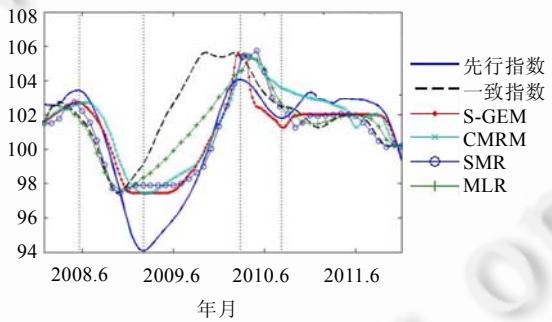


Fig.7 Comparison of registration results

图 7 排齐结果比较

总之,在先行指数对照一致指数的排齐过程中,相比于 CMRM,SMR 和 MLR,S-GEM 算法在排齐效果(相关系数和直观图形)上均优于其他方法,且运算效率高于 CMRM 和 SMR.

3 总结与展望

本文在一定显著性水平上给出总体相关系数的上下界,并用于判别相关性.伪回归问题产生的原因较多,目前尚未找到严格、准确的识别方法.本文从伪回归产生的主要原因出发,得到关于时移序列相关系数的特点,能够排除多数常见的伪回归现象;对于另外一种相关性错误,可从时移序列相关系数的特征认定其相关性.对存在时间弯曲的相关序列,建立了基于相关系数最大化的模型和改进的 S-GEM 求解算法,模型的适用范围比 AISE 准则更广.实验结果表明,本文的相关性判别方法在伪回归识别中比 Pearson 线性相关系数、Spearman 秩相关系数、Kendall 秩相关系数以及 Granger 因果检验更有效.提出的 S-GEM 算法在大多数情况下明显优于 CMRM,SMR 和 MLR.本文考虑的是双序列的线性相关问题和函数型曲线排齐方法,这些结果可为回归分析的相关性判定和时间对齐提供理论基础,并为多序列相关性分析和曲线排齐提供参考方向.

对于具有一定光滑性的高维数据,可以将其转化为函数型数据进行分类、回归、聚类等分析;另一方面,也可以采用各个基函数的系数对高维数据进行降维.本文提出的函数型相关系数和曲线排齐方法可为高维数据的回归、聚类等做好相关性度量以及降维等前期工作.

函数型数据的光滑技术便于处理数据量比较大的时序数据,在一定程度上能够克服数据含噪声、数据缺失和数据异常的问题,本文进一步将其用于加速曲线排齐迭代算法.然而,S-GEM 算法的采样点数直接影响排齐效果和排齐时间,因此,如何根据排齐问题确定最佳采样点(数量和位置),也是今后研究的一个重要方向.

References:

[1] Adelfio G, Chiodi M, D'Alessandro A, Luzio D, D'Anna G, Mangano G. Simultaneous seismic wave clustering and registration. Computers & Geosciences, 2012,44:60−69. [doi: 10.1016/j.cageo.2012.02.017]

[2] Ye L, Keogh E. Time series shapelets: A new primitive for data mining. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2009. 947−956. [doi: 10.1145%2f1557019.1557122]

[3] Zhang ZM, Salerno JJ, Yu PS. Applying data mining in investigating money laundering crimes. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2003. 747−752. [doi: 10.1145/956750.956851]

[4] Zou PC, Wang JD, Yang GQ, Zhang X, Wang LN. Distance metric learning based on side information autogeneration for time series. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2642−2655 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4464.htm [doi: 10.3724/SP.J.1001.2013.04464]

[5] Lin ZY, Jiang Y, Lai YX, Lin C. A new algorithm on lagged correlation analysis between time series: TPFP. Journal of Computer Research and Development, 2012,12:2645−2655 (in Chinese with English abstract).

[6] Kneip A, Gasser T. Statistical tools to analyze data representing a sample of curves. Annals of Statistics, 1992,20(3):1266−1305. [doi: 10.1214/aos/1176348769]

[7] Silverman BW. Incorporating parametric effects into functional principal components analysis. Journal of the Royal Statistical Society (Section B), 1995,57(4):673−689.

[8] Ramsay JO, Li X. Curve registration. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1998,60(2): 351−363. [doi: 10.1111/1467-9868.00129]

[9] Wang K, Gasser T. Alignment of curves by dynamic time warping. Annals of Statistics, 1997,25(3):1251−1276. [doi: 10.1214/aos/1069362747]

[10] Wang K, Gasser T. Asymptotic and bootstrap confidence bounds for the structural average of curves. Annals of Statistics, 1998, 26(3):972−991. [doi: 10.1214/aos/1024691084]

[11] Wang K, Gasser T. Synchronizing sample curves nonparametrically. Annals of Statistics, 1999,27(2):439−460. [doi: 10.1214/aos/1018031202]

[12] Kneip A, Li X, MacGibbon KB, Ramsay JO. Curve registration by local regression. Canadian Journal of Statistics, 2000,28(1): 19−29. [doi: 10.2307/3315251.n]

[13] Liu X, Müller HG. Functional convex averaging and synchronization for time-warped random curves. Journal of the American Statistical Association, 2004,99(467):687−699. [doi: 10.1198/016214504000000999]

[14] Rønn BB. Nonparametric maximum likelihood estimation for shifted curves. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2001,63(2):243−259. [doi: 10.1111/1467-9868.00283]

[15] Gervini D, Gasser T. Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. Biometrika, 2005, 92(4):801−820. [doi: 10.1093/biomet/92.4.801]

[16] James GM. Curve alignment by moments. The Annals of Applied Statistics, 2007,1(2):480−501. [doi: 10.1214/07-AOAS127]

[17] Liu X, Yang MCK. Simultaneous curve registration and clustering for functional data. Computational Statistics & Data Analysis, 2009,53(4):1361−1376. [doi: 10.1016/j.csda.2008.11.019]

[18] Granger CWJ, Newbold P. Spurious regressions in econometrics. Journal of Econometrics, 1974,2(2):111−120. [doi: 10.1016/0304-4076(74)90034-7]

[19] Phillips PCB. New tools for understanding spurious regressions. Econometrica, 1998,66(6):1299−1325. [doi: 10.2307/2999618]

[20] Liu HZ. The analysis of spurious regressions in stationary processes without drifts. The Journal of Quantitative & Technical Economics, 2010,(11):142−154 (in Chinese with English abstract).

[21] Liu HZ. The analysis of spurious between weak stationary processes based on autocorrelation perspective. Statistics & Information Forum, 2012,27(4):10−16 (in Chinese with English abstract).

[22] Jin H, Zhang JS, Zhang S, Yu C. The spurious regression of AR(p) infinite-variance sequence in the presence of structural breaks. Computational Statistics & Data Analysis, 2013,67:25−40. [doi: 10.1016/j.csda.2013.04.011]

[23] Zhao ZW, Liu YP, Song LX. Asymptotic normality of sample correlation coefficient of a bivariate normal distribution. Journal of Jiamusi University, 2009,27(4):607−608, 614 (in Chinese with English abstract).

[24] National bureau of statistics of the PRC. 2013-11-16/2013-12-10 (in Chinese). http://data.stats.gov.cn/

[25] Gervini D, Gasser T. Self-Modelling warping functions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2004,66(4):959−971. [doi: 10.1111/j.1467-9868.2004.B5582.x]

附中文参考文献:

[4] 邹朋成,王建东,杨国庆,张霞,王丽娜.辅助信息自动生成的时间序列距离度量学习.软件学报,2013,24(11):2642−2655. http://www.jos.org. cn/1000-9825/4464.htm [doi: 10.3724/SP.J.1001.2013.04464]

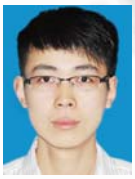
[5] 林子雨,江弋,赖永炫,林琛.一种新的时间序列延迟相关性分析算法——三点预测探查法.计算机研究与发展,2012,12: 2645−2655.

[20] 刘汉中.无漂移平稳过程下的伪回归分析——基于改进的 HAC 方法.数量经济技术经济研究,2010,(11):142−154.

[21] 刘汉中.基于自相关视角的弱平稳过程之间的伪回归分析.统计与信息论坛,2012,27(4):10−16.

[23] 赵志文,刘银萍,宋立新.二元正态总体样本相关系数的渐近正态性.佳木斯大学学报,2009,27(4):607−608,614.

[24] 中华人民共和国国家统计局统计数据.2013-11-16/2013-12-10. http://data.stats.gov.cn/



姜高霞(1987－),男,山西新绛人,博士生,主要研究领域为统计机器学习,数据挖掘.
E-mail: jianggaoxia@163.com

王文剑(1968－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,计算智能,数据挖掘.
E-mail: wjwang@sxu.edu.cn