控制与决策
Control and Decision

基于因子模型和动态规划的多元时间序列分段方法

王玲， 徐培培， 彭开香

引用本文:
王玲, 徐培培, 彭开香. 基于因子模型和动态规划的多元时间序列分段方法[J]. 控制与决策, 2020, 35(1): 35−44.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.0535

文章编号: 1001-0920(2020)01-0035-10 DOI: 10.13195/j.kzyjc.2018.0535

基于因子模型和动态规划的多元时间序列分段方法

王 玲†, 徐培培, 彭开香

(1. 北京科技大学 自动化学院，北京 100083；2. 北京科技大学
工业过程知识自动化教育部重点实验室，北京 100083)

摘 要: 针对经典动态规划分段算法只适用于低维时间序列的问题,提出一种基于因子模型和动态规划的多元时间序列分段方法. 首先利用增量聚类自动对变化趋势相似的变量序列进行聚类,然后引入动态因子模型使降维后的低维多元时间序列能够最大限度反映原始多元时间序列的整体变化趋势,最后利用动态规划在低维多元时间序列的架构上实现高维多元时间序列的分段. 实验结果表明,所提方法对变量个数较多的多元时间序列数据具有良好的分段效果.

关键词: 多元时间序列分段；因子分析；动态规划；增量聚类

中图分类号: TP273 **文献标志码:** A

Segmentation of multivariate time series with factor model and dynamic programming

WANG Ling†, XU Pei-pei, PENG Kai-xiang

(1. School of Automation and Electrical Engineering，University of Science and Technology Beijing，Beijing 100083，China；2. Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education，University of Science and Technology Beijing，Beijing 100083，China)

Abstract: The classical dynamic programming based segmentation algorithm is only suitable for low dimensional time series. To solve this problem, a segmentation method of multivariate time series with factor model and dynamic programming is proposed. Firstly, incremental clustering is used to automatically cluster variable sequences with similar trend. Then, a dynamic factor model is introduced to make the low-dimension multivariate time series obtained after dimension reduction reflect the overall trend of the original multivariate time series. Finally, the segmentation of high-dimension multivariate time series in the framework of low-dimension time series is realized by using dynamic programming. The experimental studies show that the proposed method has a good segmentation effect on multivariate time series data with a large number of variables.

Keywords: multivariate time series segmentation；factor analysis；dynamic programming；incremental clustering

0 引 言

时间序列是由一系列按照时间顺序排列观测值的有序集合,广泛存在于金融、气象、医学、工业过程等领域. 时间序列分段是将数据集的整个时间范围划分到多个同质且不重叠的区间[1]. 作为时间序列分析的一个分支,时间序列分段问题长久以来吸引着众多研究者的注意[2-6]. 早期的研究主要是针对一元时间序列,采用的方法有等宽度划分[7]、基于云模型[8]、基于重要点[9]、基于动态规划[10]、基于贝叶斯准则[11]等. 然而,在现实生活中,实际采集到的时间序列往往都是包含多个变量,考虑如何对这些变量同时进行分割,就需要将原有一元变量分段算法进行扩展或创造新的分段算法,以适应新的挑战.

因子模型是一种用少量潜在的、不可观测的因子来描述许多变量间相关关系的模型[12]. 经典因子模型主要用于处理截面数据. 为了将因子分析应用于时间序列, Geweke[13]最先提出了动态因子模型的概念,它是经典因子模型在时态数据上的扩展. 动态因子模型不仅允许时间序列在任意时刻的观测值可以受到动态因子滞后项的影响,也允许动态因子自身服从某种自回归过程. 文献[14]引入动态因子模型对多元时间序列提取共同因子,并将共同因子的分段结

收稿日期: 2018-04-25；**修回日期:** 2018-06-04.

基金项目: 国家自然科学基金项目(61572073)；北京科技大学中央高校基本科研业务费专项资金项目(FRF-BD-17-002A)；北京市重点学科共建项目(XK100080537).

†通讯作者. E-mail: lingwang@ustb.edu.cn.

果视为原始多元时间序列的分段结果,但是由于只提取一条共同因子序列,此方法只能用于变量个数较少且彼此之间相关性较大的多元时间序列;文献[15]考虑不同子段进行因子分析后得到的模型可能不同,结合小波变换与累积和统计对高维时间序列进行变点检测,但是却将时间序列视为截面数据,没有考虑因子滞后项的影响.

然而,不论序列是一元还是多元,分段算法要实现的目标是:寻找最优分段个数以及最优分段位置.因此,在众多阐述时间序列分段技术的文献中,许多研究都引入了优化算法来解决问题,动态规划就是其中之一. 文献[10]将动态规划过程应用到水文和环境时间序列的分段中,并借助贝叶斯信息准则(BIC)评估得到最优的分段个数;文献[16]对长水文气象时间序列进行离线分割,提出了一种将动态规划与分支定界法的剩余代价[17]概念相结合的方法. 上述两种方法只针对一元序列有效. 文献[18]提出了用动态规划来分段多元时间序列(segmentation of multivariate time series with dynamic programming, SMTS-DP),借助向量自回归模型将文献[10]的方法进行扩展,推导出了分段误差的递归计算公式,且实验显示对包含了3个变量的水文气象数据集具有良好的分段效果. 但是,此方法需要对多元时间序列计算所有可能分段情况下的分段误差,当数据集中包含的变量个数过多时,算法运行效率会下降,且得到的分段结果并不十分精确.

本文所提出的基于因子模型和动态规划的多元时间序列分段方法(segmentation of multivariate time series with factor model and dynamic programming, SMTS-FD)是在SMTS-DP方法基础之上进行的改进. 与原始的SMTS-DP方法相比,SMTS-FD引入动态因子模型对多元时间序列进行降维,尤其在处理复杂数据集时,对变量个数较多的多元时间序列具有更好的分段效果,提高了算法运行效率. 其次,为了保证在低维多元时间序列的架构上用动态规划实现高维多元时间序列的分段而不降低分段精确性,SMTS-FD引入增量聚类算法自动将变化趋势相似的变量聚成一类,使得因子分析后得到的低维多元时间序列能够最大限度反映原始多元时间序列的共同变化趋势.

1 问题的定义

假定一个多元时间序列数据集为$Z(t), t = 1, 2, \cdots, T$,$t$是数据集的时间戳,$T$是序列的长度,$Z(t) = [z_1(t), z_2(t), \cdots, z_v(t), \cdots, z_k(t)]'$是$t$时刻获取到的采样点,$k$是此多元时间序列包含的变量个数,$z_v(t)$ $(v = 1, 2, \cdots, k)$是第v个变量在时刻t的采样值. 用$N(2 \leqslant N \leqslant T)$表示序列的分段个数,则序列的分段位置$t_i(i = 0, 1, \cdots, N)$满足$0 = t_0 < t_1 < \cdots < t_N = T$. 由序列的分段位置划分出的$N$个区间$[t_0+1, t_1], [t_1+1, t_2], \cdots, [t_{N-1}+1, t_N]$称为多元时间序列$Z(t)$的$N$个段.

时间序列的分段问题可以被视为一个优化问题,通过最小化某个代价函数来得到最优分段个数和相应的最优分段位置. 定义如下分段代价函数:

$$L(t) = \sum_{i=1}^{N} e_{t_{i-1}+1, t_i}, \tag{1}$$

其中$e_{t_{i-1}+1, t_i}$是段$[t_{i-1}+1, t_i]$对应的分段误差. 分段误差$e_{t_{i-1}+1, t_i}$的大小依赖于段$[t_{i-1}+1, t_i]$包含的数据子集$\{z(t_{i-1}+1), z(t_{i-1}+2), \cdots, z(t_i)\}$,用如下公式来计算:

$$e_{t_{i-1}+1, t_i} = \sum_{\tau=t_{i-1}+1}^{t_i} (Z(\tau) - \hat{Z}(\tau))'(Z(\tau) - \hat{Z}(\tau)). \tag{2}$$

其中:$Z(\tau)$是一个k维时间序列,$\hat{Z}(\tau)$是$Z(\tau)$的某种回归估计. 假定$Z(\tau)$在段$[t_{i-1}+1, t_i]$内满足p阶向量自回归模型

$$Z(\tau) = \theta_0^{(i)} + \theta_1^{(i)} Z(\tau - 1) + \cdots + \theta_p^{(i)} Z(\tau - p) + u^{(i)}(\tau). \tag{3}$$

其中:$\tau - p \geqslant 1$;$u^{(i)}(\tau)$是k维误差向量,服从均值为0、协方差矩阵为Σ的多元正态分布;$\theta_0^{(i)}$是k维列向量;$\theta_1^{(i)}, \cdots, \theta_p^{(i)}$都是$k \times k$维矩阵;上标$i$与段$[t_{i-1}+1, t_i]$对应,表示第$i$个段内的数据子集满足的回归模型,即不同段内的数据子集满足的p阶向量自回归模型可以不同.

基于如上假定,式(2)中的$\hat{Z}(\tau)$是$Z(\tau)$在段$[t_{i-1}+1, t_i]$内的p阶向量自回归估计,有

$$\hat{Z}(\tau) = \hat{\theta}_0^{(i)} + \hat{\theta}_1^{(i)} Z(\tau - 1) + \cdots + \hat{\theta}_p^{(i)} Z(\tau - p), \tag{4}$$

其中$\hat{\theta}_0^{(i)}, \hat{\theta}_1^{(i)}, \cdots, \hat{\theta}_p^{(i)}$是第$i$个段内回归系数的估计.

为了缩减分段误差的计算复杂度,文献[18]用递归的形式计算所有可能情况下的分段误差,代入式(1)即可得到相应的分段代价. 如果序列中包含N个段,将最优分段位置表示为$\hat{t} = (\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_N)$,则有

$$\hat{t} = \arg\min_{t \subseteq T_N} L(t), \tag{5}$$

其中T_N是将多元时间序列分成N段时所有可能分段位置的集合. 对于向量自回归模型的阶数p和分段

个数N的选择,借助模型选择准则进行综合考量,能够同时得到模型阶数和分段个数的最优值.

2 基于因子模型和动态规划的多元时间序列分段方法(SMTS-FD)

从现实世界中得到的多元时间序列数据集有许多都是高维的. 随着时间的变化,多元时间序列中的某些变量往往会展现出十分相似的变化趋势. 在这种情况下,直接对原始的多元时间序列进行分段通常效率较低. 如果能够从原始多元时间序列中提取出一个可以很好地反映其整体变化趋势的低维多元时间序列,通过对此低维序列进行分段进而得到原始序列的分段结果,则算法整体的执行效率将会得到很大提高.

2.1 变量聚类

为了在较低的维度上处理高维多元时间序列的分段问题,SMTS-FD方法首先对原始多元时间序列进行变量聚类. 文献[19]提出了一种增量聚类算法,通过计算任意两个变量之间的Pearson相关系数,进而得到归一化相关系数来度量变量之间的相似性,且不用提前设置聚类个数即可自适应地获得聚类结果. 为了使此聚类算法得到的每个簇中只包含变化趋势相似的变量序列,这里只用Pearson相关系数来度量变量之间的相似程度,且只考察彼此之间具有统计学上显著的正相关性的变量序列是否能聚成一类. Pearson相关系数的计算公式如下:

$$\mathrm{corr}\,(z_i(t),z_j(t))=\frac{\sum\limits_{t=1}^{T}[z_i(t)-\bar{z}_i][z_j(t)-\bar{z}_j]}{\sqrt{\sum\limits_{t=1}^{T}[z_i(t)-\bar{z}_i]^2\sum\limits_{t=1}^{T}[z_j(t)-\bar{z}_j]^2}}. \tag{6}$$

其中: $z_i(t)$和$z_j(t)(i,j=1,2,\cdots,k,t=1,2,\cdots,T)$是多元时间序列中任意两个变量序列; $\bar{z}_i$和$\bar{z}_j$分别是变量序列$z_i(t)$和$z_j(t)$的均值.

这里使用置信水平为α的t检验来确定相关性是否在统计学上显著,且按使用惯例取$\alpha=95\,\%$. 另外,在某种程度上,正相关性代表相似的变化趋势,负相关性代表相反的变化趋势,因此这里只用到了正相关性对变量进行增量聚类.

2.2 共同因子序列的提取

原始多元时间序列经过变量聚类得到的每个簇中都包含一个或多个变量序列. 对于包含单变量序列的簇,不再对其进行降维处理,直接将此单变量序列加入所要得到的低维多元时间序列中. 对于包含多变量序列的簇,由于聚类过程采用了相关系数来度量相似性,最终得到的簇中变量之间都具有较大的相关性. 假定存在一个潜在的因子驱动着此簇中多变量序列的变化,则采用动态因子模型提取出这个潜在的共同因子将是十分有效的.

假定对原始多元时间序列进行变量聚类得到了$c(1\leqslant c\leqslant k)$个簇,将第$m(m=1,2,\cdots,c)$个簇中的多变量序列数据集记为$\{Z^{(m)}(t)|t=1,2,\cdots,T\}$, $Z^{(m)}(t)=[z_1^{(m)}(t),z_2^{(m)}(t),\cdots,z_{k^{(m)}}^{(m)}(t)]'$是$t$时刻的数据点, $k^{(m)}\geqslant 2$是第m个簇中包含的变量个数. 用$\Lambda^{(m)}$表示由滞后算子多项式组成的$k^{(m)}\times 1$维载荷向量, $f^{(m)}(t)$表示一个随时间变化的共同因子, $v^{(m)}(t)$表示一个$k^{(m)}\times 1$维异质性向量, $w^{(m)}(t)$表示一个扰动项,则第m个簇中的多变量序列$Z^{(m)}(t)$可用动态因子模型建模为

$$Z^{(m)}(t)=\Lambda^{(m)}f^{(m)}(t)+v^{(m)}(t), \tag{7}$$

$$\Psi^{(m)}(B)f^{(m)}(t)=w^{(m)}(t). \tag{8}$$

其中: $\Psi^{(m)}(B)=1-\Psi_1^{(m)}B-\cdots-\Psi_{q^{(m)}}^{(m)}B^{q^{(m)}}$是一个$q^{(m)}$阶滞后算子多项式,对于$i=1,2,\cdots,q^{(m)}$, $\Psi_i^{(m)}$是此多项式的系数,且B是一个滞后算子,满足$B^if^{(m)}(t)=f^{(m)}(t-i)$.

动态因子模型的估计可采用频域方法或时域方法. 然而,频域方法并不能直接估计出共同因子$f^{(m)}(t)$,限制了动态因子模型的应用. 后来的研究中,学者们普遍采用时域方法,将动态因子模型表示为静态形式,从而得到$f^{(m)}(t)$的估计.

假定$\Lambda^{(m)}$中滞后算子多项式的阶数为$s^{(m)}$,为了得到动态因子模型(7)、(8)的静态形式,记$F^{(m)}(t)=[f^{(m)}(t),f^{(m)}(t-1),\cdots,f^{(m)}(t-s^{(m)})]'$为$(s^{(m)}+1)\times 1$维向量; $A^{(m)}=[\Lambda_0^{(m)},\Lambda_1^{(m)},\cdots,\Lambda_{s^{(m)}}^{(m)}]$为$k^{(m)}\times(s^{(m)}+1)$维矩阵,其中$\Lambda_i^{(m)}(i=0,1,\cdots,s^{(m)})$为$\Lambda^{(m)}$的第$i$阶滞后算子系数所组成的$k^{(m)}\times 1$维向量. 因此,式(7)转换为

$$Z^{(m)}(t)=A^{(m)}F^{(m)}(t)+v^{(m)}(t). \tag{9}$$

令$\varepsilon^{(m)}(t)=G^{(m)}w^{(m)}(t)$,其中$G^{(m)}$是由0和1组成的列向量;记$D^{(m)}(B)$是由0、1及$\Psi^{(m)}(B)$组成的矩阵,则式(8)可以用$F^{(m)}(t)$表示为

$$D^{(m)}(B)F^{(m)}(t)=\varepsilon^{(m)}(t). \tag{10}$$

考虑到高维数据集会带来繁琐的计算,本文采用主成分评估方法来估计出参数$A^{(m)}$及因子$F^{(m)}(t)$. 主成分评估方法仅需要动态因子模型的静

态表达式(9),而不需要对因子$F^{(m)}(t)$设定类似于式(10)的参数模型[12],使得对高维数据的分析能力大大提高. 首先,计算样本协方差矩阵

$$\hat{\Sigma}_{Z^{(m)}(t)}=\frac{1}{T}\sum_{t=1}^{T}Z^{(m)}(t)Z^{(m)}(t)'. \tag{11}$$

由主成分评估的原理,一般根据累积方差贡献率大于等于0.8,得到样本协方差矩阵的前$s^{(m)}+1$个最大特征值对应的特征向量. 参数的估计$\hat{A}^{(m)}$就可以用这些特征向量所组成的$k^{(m)}\times(s^{(m)}+1)$维矩阵来表示,从而因子的主成分估计量[20]$\hat{F}^{(m)}(t)=(\hat{A}^{(m)})'Z^{(m)}(t)/k^{(m)}$. 由于只提取一条共同因子序列,$\hat{F}^{(m)}(t)$向量的第一个元素就是共同因子$f^{(m)}(t)$的估计.

2.3 多元时间序列分段

对聚类后得到的每个多变量序列数据集都提取一条共同因子序列,结合没有与其他变量聚成一类的单变量序列,构成了一个维度为$c(1\leqslant c\leqslant k)$的新的低维多元时间序列. 也就是说,原始包含$k$个变量的多元时间序列经过聚类和因子分析已经被降维成了包含c个变量的低维多元时间序列,低维多元时间序列包含了原始多元时间序列中的整体变化趋势信息. 对低维多元时间序列进行分段,首先需要用向量自回归模型对整个低维多元时间序列进行整体拟合,以得到模型的最大自回归阶数$p_{\max}$,然后利用动态规划分段低维多元时间序列得到分段代价最小值,从而找到最优自回归阶数$p_{\rm opt}$、最优分段个数$N_{\rm opt}$以及全局最优分段位置$\hat{t}_{\rm opt}$.

对于拟合模型的可能阶数$p=0,1,\cdots,p_{\rm set}$,其中$p_{\rm set}$是根据经验设置的阶数上限,用赤池信息准则(AIC)[21]来选择出最大自回归阶数$p_{\max}(0\leqslant p_{\max}\leqslant p_{\rm set})$:

$$\mathrm{AIC}(p)=\log|\hat{\Sigma}_{u(t)}|+\frac{2}{T-p}\times c\times(c\times p+1). \tag{12}$$

其中:c是低维多元时间序列的变量个数;$c\times(c\times p+1)$是拟合模型的参数个数;$\hat{\Sigma}_{u(t)}$是拟合模型误差$u(t)$的协方差矩阵$\Sigma_{u(t)}$的估计. 用最小二乘法来估计模型时,$\hat{\Sigma}_{u(t)}$计算为

$$\hat{\Sigma}_{u(t)}=\frac{1}{T-(c+1)p-1}\sum_{t=p+1}^{T}\hat{u}(t)\hat{u}(t)', \tag{13}$$

其中$\hat{u}(t)=Z(t)-\hat{Z}(t)$是拟合模型误差的估计.

尽管在运用向量自回归模型前,往往需要对被拟合数据进行平稳性检验,但是当假定每个段内的数据都可以用一个回归模型来拟合时,如果算法在分段过程中有个段横跨了几个真实的分段(例如,将整个低维多元时间序列视为一个段时,此段就横跨了多个真实的分段),则该段数据的真实值和估计值之间相差较大,代入式(2)得到的分段误差值就会比较大,从而分段代价值也会变大,算法不会将此分段结果输出为最优分段结果. 因此,即使被拟合数据是非平稳的,用向量自回归模型对其进行拟合也不会影响数据集整体的分段结果. 另一方面,数据集的非线性将会不可避免地使自回归阶数增大,所以使式(12)的值达到最小的自回归阶数$p=p_{\max}$并不是最优的自回归阶数$p_{\rm opt}$,最优自回归阶数$p_{\rm opt}$应小于$p_{\max}$.

对于每个可能的自回归阶数$p=0,1,\cdots,p_{\max}$和可能的分段个数$N=2,\cdots,N_{\max}$,动态规划都能够高效地找到对应的最优分段位置$\hat{t}=(\hat{t}_0,\hat{t}_1,\cdots,\hat{t}_N)$. 假定降维后的低维多元时间序列为$\{Y(t)|t=1,2,\cdots,T\}$,考虑将其划分为$N$个最优分段,记最后一个段为$[t_{N-1}+1,T]$,则前$N-1$个段就构成了集合$\{Y(1),Y(2),\cdots,Y(t_{N-1})\}$的最优分段. 更特殊地,如果用$L^{(N)}(t)$表示将$\{Y(1),Y(2),\cdots,Y(t)\}$分成$N$段的最小分段代价,可以得到

$$L^{(N)}(t)=L^{(N-1)}(t_{N-1})+e_{t_{N-1}+1,t}. \tag{14}$$

其中:$L^{(N-1)}(t_{N-1})$是将$\{Y(1),Y(2),\cdots,Y(t_{N-1})\}$分成$N-1$段的最小分段代价;$e_{t_{N-1}+1,t}$是段$[t_{N-1}+1,t]$对应的分段误差. 对于$t=1,2,\cdots,T$,动态规划分别计算式(14). 最终,可得到将低维多元时间序列$\{Y(t)|t=1,2,\cdots,T\}$划分成N段的最小分段代价为

$$\min_{t\subseteq T_N}L(t)=L^{(N)}(T). \tag{15}$$

算法在计算最小分段代价$L^{(N)}(t)$的同时会存储当前找到的一个分段位置. 当得到$L^{(N)}(T)$之后,通过回溯搜索,即可得到相应的最优分段位置$\hat{t}=(\hat{t}_0,\hat{t}_1,\cdots,\hat{t}_N)$.

然而,对于$p=0,1,\cdots,p_{\max}$和$N=2,\cdots,N_{\max}$,算法会得到$(p_{\max}+1)\times(N_{\max}-1)$种最优分段位置. 为避免参数太多或分段太多,用BIC准则[22]同时确定自回归阶数和分段个数的最优值,即

$$\mathrm{BIC}(p,N)=\log\left(\frac{L^{(N)}(T)}{T-p-1}\right)+\frac{\log(T-p)}{T-p}\times N\times c\times(c\times p+1). \tag{16}$$

使式(16)达到最小值的p和N,即为最优的自回归阶数$p_{\rm opt}$和分段个数$N_{\rm opt}$,相应的分段位置即为低维多元时间序列的全局最优分段位置$\hat{t}_{\rm opt}$. 由于经过

变量聚类和共同因子提取得到的低维多元时间序列能够代表原始多元时间序列的整体变化趋势,可以将低维多元时间序列的分段结果,包括最优分段个数和最优分段位置,视为原始多元时间序列的分段结果,实验部分将会表明本文所提SMTS-FD方法是十分有效的.

2.4 SMTS-FD方法实现步骤

为了能够分段变量个数较多的多元时间序列,SMTS-FD方法首先借助增量聚类自动将原始数据集中相似的变量序列聚成一类;然后利用因子分析,提取出能够反映原始数据集整体变化趋势的低维多元时间序列;最后通过动态规划对低维多元时间序列进行分段,结合模型选择准则获得最优分段个数和分段位置,并将低维架构上的分段结果视作原始高维数据集的分段结果. SMTS-FD方法的具体实现步骤如下:

Step 1: 读入原始多元时间序列并进行z-score标准化处理. 设置自回归阶数上限$p_{\rm set}$和最大分段个数$N_{\max}$的参数值.

Step 2: 计算任意两个变量之间的Pearson相关系数,根据统计学上显著的正相关性对变量序列进行增量聚类,得到c个簇.

Step 3: 对于聚类得到的每个簇,如果某个簇中只包含一个变量,则执行Step 3.1;否则,执行Step 3.2.

Step 3.1: 将该簇中的变量序列直接加入低维多元时间序列数据集中;

Step 3.2: 利用动态因子模型从该簇中的多变量序列数据集提取出一条共同因子序列,然后将共同因子序列加入低维多元时间序列数据集.

Step 4: 用向量自回归模型对得到的维度为c的低维多元时间序列进行拟合,然后根据AIC准则从$p=0,1,\cdots,p_{\rm set}$中选择出最大自回归阶数$p_{\max}$.

Step 5: 对于每个可能的自回归阶数$p=0,1,\cdots,p_{\max}$,分别用动态规划找到$N=2,\cdots,N_{\max}$对应的最优分段位置,并计算出最小段代价.

Step 6: 将自回归阶数、分段个数及相应的最小段代价代入BIC准则计算公式,选择出最优的自回归阶数$p_{\rm opt}$和分段个数$N_{\rm opt}$.

Step 7: 经过回溯,得到低维多元时间序列的全局最优分段位置$\hat{t}_{\rm opt}$,同时其也被视为原始多元时间序列的全局最优分段位置.

2.5 计算复杂度分析

SMTS-FD方法对多元时间序列的分段主要由3个阶段组成. 对于标准化后的k维多元时间序列,SMTS-FD方法执行变量聚类过程的计算量为$O(k^2)$. 聚类后得到的第$m(m=1,2,\cdots,c)$个簇中包含$k^{(m)}$个变量,不论是否需要对其进行因子分析,处理过程所需计算量的通式均为$O(k^{(m)})$,因此得到低维多元时间序列所需的计算量为$O\Big(\sum\limits_{m=1}^{c}k^{(m)}\Big)=O(k)$. 分段过程中,用AIC准则找$p_{\max}$的计算量为$O((p_{\rm set}+1)\times c)$;利用动态规划对每种情况进行分段所需的计算量为$O((p_{\max}+1)\times(T^2+cT+N_{\max}T^2+{N_{\max}}^2))$;用BIC准则确定$p_{\rm opt}$和$N_{\rm opt}$以及回溯得到全局最优分段位置的计算量为$O((p_{\max}+1)\times(N_{\max}-1))$. 综上所述,SMTS-FD方法总的计算复杂度经简化后为$O(k^2+(p_{\rm set}+1)\times c+(p_{\max}+1)\times((N_{\max}+1)T^2+cT+{N_{\max}}^2))$. 可以看出,该方法的计算复杂度与数据集规模、降维幅度以及参数设置均密切相关.

3 实验结果

本节中同时利用合成多元时间序列数据集和真实多元时间序列数据集来测试SMTS-FD算法的运行效果,并与SMTS-DP方法进行对比实验,以验证所提方法的有效性. 分段实验中均设置自回归阶数上限$p_{\rm set}=5$,最大分段个数$N_{\max}=5$. 另外,为了分析参数设置对SMTS-FD分段结果的影响,3.3节给出了不同数据集在不同参数设置下的分段结果对比. 所有实验均是在2.40 GHz处理器和4.00 GB内存下运行的Python 3.5环境中执行.

3.1 合成数据集的实验

采用人工合成数据集的优点是时间序列的正确分段是已知的,因此便于直观地对分段算法的运行结果进行评估. 这里利用1阶向量自回归模型生成了一个10维时间序列数据集,长度$T=100$,所用的模型如下所示:

$$Z(t)=\begin{cases}\theta_0^{(1)}+\theta_1^{(1)}Z(t-1)+u^{(1)}(t),\ 0<t\leqslant 40;\\ \theta_0^{(2)}+\theta_1^{(2)}Z(t-1)+u^{(2)}(t),\ 40<t\leqslant 100.\end{cases}\tag{17}$$

其中: $Z(t)=[z_1(t),z_2(t),\cdots,z_{10}(t)]',t=1,2,\cdots,100$; $u^{(1)}(t)$和$u^{(2)}(t)$都服从均值为0、协方差矩阵$\Sigma=I$的多元正态分布. 模型的参数设置如下:

$$\theta_0^{(1)}=[3,3,3,3,-1,-1,-1,-1,-1,-1]',\tag{18}$$

$$\theta_1^{(1)}=\begin{bmatrix}-0.9 & 0 & \cdots & 0\\ 0 & -0.9 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & -0.9\end{bmatrix},\tag{19}$$

$$\theta_0^{(2)} = [-1,-1,-1,-1,5,5,5,5,5,5]', \tag{20}$$

$$\theta_1^{(2)} = \begin{bmatrix} 0.6 & 0 & \cdots & 0 \\ 0 & 0.6 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0.6 \end{bmatrix}. \tag{21}$$

由模型可知,生成的10维时间序列应当可以分成2个子段,且真实分段位置为40、100,其中100是时间序列的最后一个时间点. 数据集经过z-score标准化处理后的时序如图1所示,并用虚线标示出了真实分段位置. 对此10维时间序列用SMTS-FD方法进行分段,结果将变量序列$\{z_1, z_2, z_3, z_4\}$聚成第1个类簇,而将变量序列$\{z_5, z_6, z_7, z_8, z_9, z_{10}\}$聚成了第2个类簇. 对得到的两个簇中的多变量序列分别建立动态因子模型,提取出的两条共同因子序列如图2所示,其中因子1和因子2分别提取自第1个簇和第2个簇.

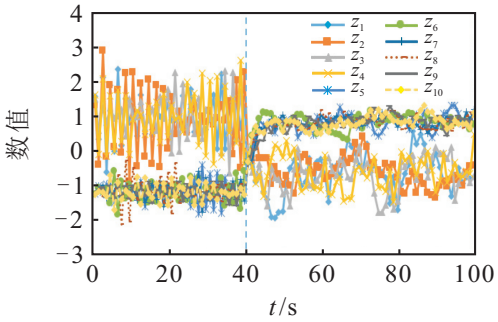

图1 人工合成的10维时间序列及其真实分段位置

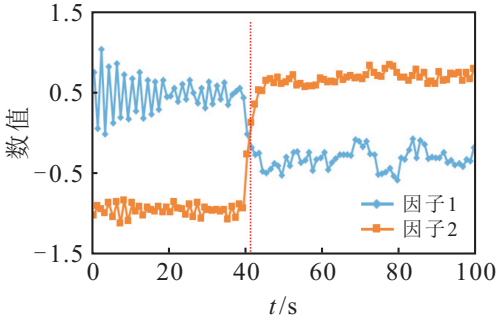

图2 提取自人工合成时间序列的共同因子及其分段结果

对于自回归阶数$p = 0, 1, \cdots, 5$,分别对图2的2维共同因子序列进行向量自回归拟合,用式(12)计算出的AIC值如表1所示. 由表1可知,当$p = 3$时AIC值达到最小,因此最大自回归阶数$p_{\max} = 3$. 为了找出最优的自回归阶数、分段个数和分段位置,对每个$p = 0, 1, 2, 3$和$N = 2, 3, 4, 5$分别借助动态规划对2维共同因子序列进行分段,在此过程中,计算式(16)得到的各种情况下的BIC值如表2所示. 可以看出,当$p = 1, N = 2$时,BIC值最小. 也就是说,BIC准则选出的2维共同因子序列的最优自回归阶数$p_{\text{opt}} = 1$,最优分段个数$N_{\text{opt}} = 2$. 经过回溯,得到对应的全局最优分段位置$\hat{t}_{\text{opt}}$为41、100,这是2维共同因子序列的最优分段位置,同时也被视为是原始10维时间序列的最优分段位置. SMTS-FD方法对此合成的10维时间序列执行分段共用了10.96 s,且最终得到的分段结果与数据集真实的分段结果十分接近.

表1 SMTS-FD方法应用于合成数据集得到的AIC值

p	0	1	2	3	4	5
AIC(p)	−1.64	−5.72	−6.20	−6.44	−6.41	−6.33

表2 SMTS-FD方法应用于合成数据集得到的BIC值

p	N			
	2	3	4	5
0	−1.46	−1.41	−1.34	−1.28
1	−2.02	−1.82	−1.95	−1.75
2	−1.65	−1.63	−1.32	−0.91
3	−1.26	−1.05	−0.54	0.06

为了对比出改进后的SMTS-FD方法对变量个数较多的多元时间序列进行分段的优越性,这里接着用改进前的SMTS-DP方法对此合成的10维时间序列进行分段. SMTS-DP方法用向量自回归模型对原始10维时间序列进行拟合,借助AIC准则对自回归阶数进行评估的结果如表3所示. 可以看出,最大自回归阶数$p_{\max} = 1$时,AIC值最小. 然后,用动态规划对$p = 0, 1$和$N = 2, 3, 4, 5$下的每种情况进行分段,再借助BIC准则对结果进行评估,计算得到的结果如表4所示,可知SMTS-DP方法选择出的原始10维时间序列的最优自回归阶数$p_{\text{opt}} = 0$,最优分段个数$N_{\text{opt}} = 2$,且经过回溯可得全局最优分段位置$\hat{t}_{\text{opt}}$为42、100. 该方法对此数据集总的执行时间为12.54 s,比SMTS-FD方法略长. 然而,由前文设置的模型(17)可知,原始10维时间序列是用自回归阶数为1的向量自回归模型合成的,因此SMTS-DP方法对此合成数据集进行分段过程中得到的模型最优自回归阶数是不准确的.

表3 SMTS-DP方法应用于合成数据集得到的AIC值

p	0	1	2	3	4	5
AIC(p)	−17.70	−22.24	−22.12	−21.34	−20.34	−20.15

表4 SMTS-DP方法应用于合成数据集得到的BIC值

p	N			
	2	3	4	5
0	1.67	2.04	2.46	2.89
1	9.82	14.70	19.63	24.32

为了更加直观地对分段结果进行观察,图3展示了原始10维时间序列的真实分段位置,并同时给出运用SMTS-FD方法和SMTS-DP方法找到的最优分段位置. 可以看出,两种方法找到的最优分段位置与数据集的真实分段位置都很接近,但SMTS-FD方法得出的结果更加精确. 另外,SMTS-FD方法对原始多元时间序列先执行变量聚类和因子分析,相当于对数据集进行了降维,然后再对得到的低维多元时间序列进行分段,因此总体上执行效率比SMTS-DP方法更高.

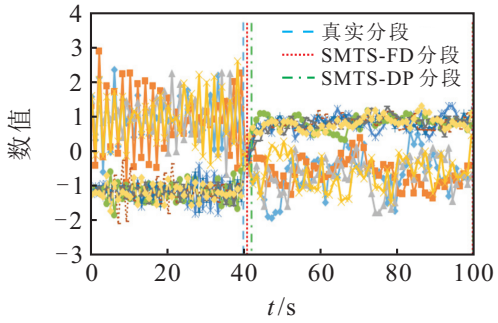

图 3　合成数据集的真实分段、SMTS-FD分段及SMTS-DP分段

3.2　真实数据集的实验

这里用两个真实多元时间序列数据集进行实验,包括Hydrometeorological(HY)数据集[18]和Human Activity Recognition (HAR)数据集[23].

3.2.1　HY数据集

HY数据集包含了windspeed、dir、gusts三个变量,记录的是美国阿雷西沃地区的水文气象数据,每个变量都对应着一条时间序列. 实验中用到的是其中的一个子集,包含了2013年10月1日为期一天的数据,长度$T = 241$. 原始数据集经z-score标准化处理后得到图4所示的时序图.

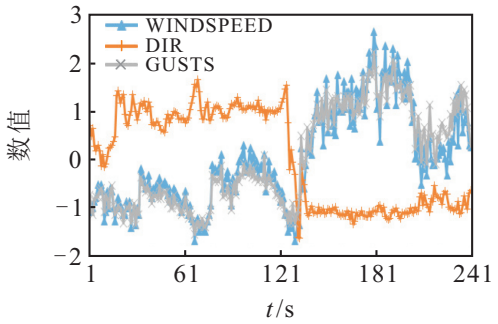

图 4　HY数据集的时序图

首先,用本文提出的SMTS-FD方法对此数据集进行分段,变量聚类和因子分析过程将windspeed和gusts聚成一类并提取出一条共同因子序列,变量dir自成一类,输出为一个单变量序列. 降维后得到的2维多元时间序列如图5所示. 用p阶向量自回归模型对2维多元时间序列进行拟合,AIC准则评估出当$p = 3$时计算出的AIC值最小,因此拟合模型的最大自回归阶数$p_{\max} = 3$. 对于$p = 0,1,2,3$和$N = 2,3,4,5$,用动态规划分别找出每种情况下对应的最优分段位置,结合BIC准则进行评估,最终得到最优自回归阶数$p_{\text{opt}} = 1$、最优分段个数$N_{\text{opt}} = 2$,且2维多元时间序列的全局最优分段位置$\hat{t}_{\text{opt}}$为134、241,已用虚线标示在图5中.

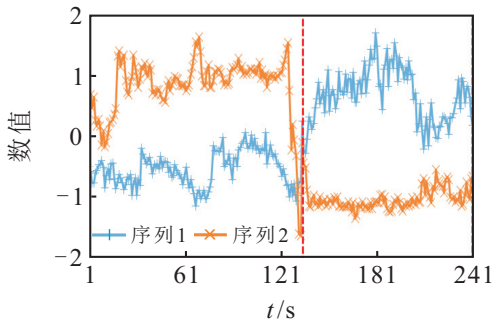

图 5　提取自HY数据集的2维时间序列及其分段结果

其次,用SMTS-DP方法对原始的数据子集进行分段,文献[18]中已描述了这部分实验内容,并得到全局最优的分段位置为134、241,与本文所提SMTS-FD方法的分段结果完全一致. 两种方法对HY数据集的分段结果对比如图6所示. 可以看出,它们均找到了此段多元时间序列合适的分段点.

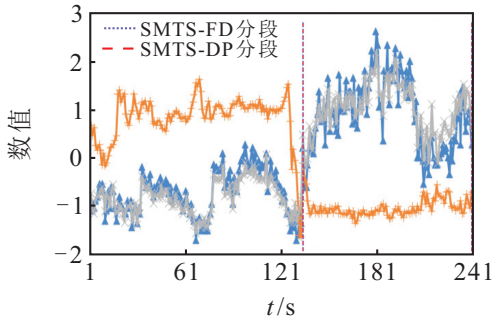

图 6　HY数据集的SMTS-FD分段及SMTS-DP分段

此外,实验中发现SMTS-FD方法对此数据集的执行时间为40.38 s,而SMTS-DP方法的执行时间为36.98 s. 这是因为改进后的SMTS-FD方法中加入了变量聚类和因子分析进行降维,然而在数据集中变量个数较少的情况下,提取出的低维多元时间序列的维度与原始多元时间序列的维度相差不大,因此包括降维在内的整个处理过程耗费的时间稍长. 尽管如此,改进后的方法依然具有很好的分段效果,虽然更适合变量个数较多的多元时间序列,但是对于变量个数较少的多元时间序列同样可以实现精确分段.

3.2.2 HAR数据集

HAR数据集记录了30个受试者的活动数据,共有561个变量,包括加速度计和陀螺仪在时域记录的数据、经过各种函数处理后得到的数据以及将时域变换到频域后得到的数据,所有数据都已经标准化到[−1,1]区间. 原始数据集中的变量个数太多,为了方便展示,实验中选择了具有代表性的17个时域变量,并随机截取了200条记录. 所选数据集的时序如图7所示,$z_1, z_2, \cdots, z_{17}$表示17个变量序列,数据集的长度$T = 200$. 由原始数据集中的活动类型标注可知,图7中时间区间[1, 18]、[117, 188]中的数据来自静止类型的活动,而时间区间[19, 116]、[189, 200]中的数据来自行走类型的活动,因此此17维多元时间序列的真实分段位置应为18、116、188、200,已经用竖虚线标示在图7中.

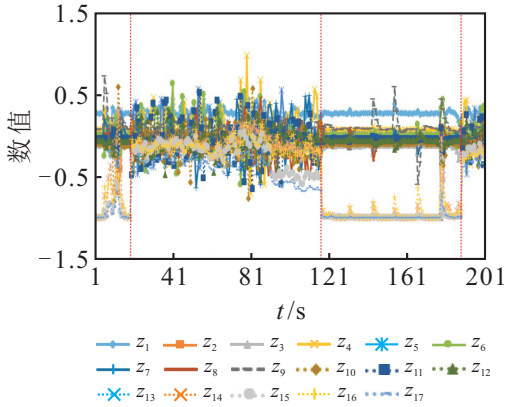

图7 HAR数据集的时序图

用本文所提SMTS-FD方法对数据集进行分段,变量聚类过程得到了7个类簇,且经过因子分析后获得的7维多元时间序列如图8所示. 为了更好地展示提取出的共同因子序列的有效性,这里给出一个例子,将其中一个包含了变量z_{12}、z_{13}、z_{14}、z_{15}、z_{16}的类簇以及从中提取出的共同因子序列7展示在图9中. 可以看出,序列7很好地反映出了变量z_{12}、z_{13}、z_{14}、z_{15}、z_{16}的整体变化趋势.

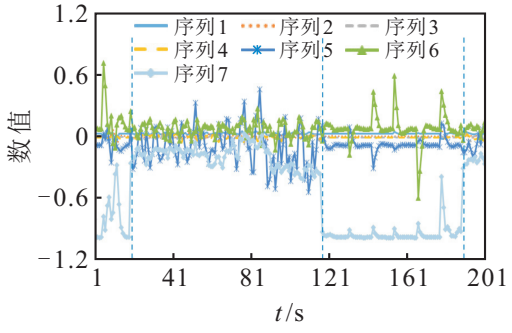

图8 提取自HAR数据集的7维时间序列及其分段结果

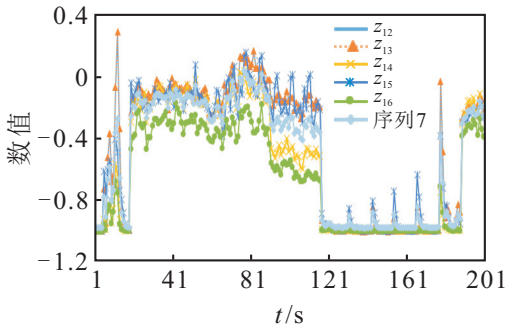

图9 共同因子序列提取示例

对降维后的7维多元时间序列进行分段,SMTS-FD方法得到的最大自回归阶数$p_{\max} = 2$. 对$p = 0, 1, 2$和$N = 2, 3, 4, 5$的各种情况下进行分段,尝试得到的BIC值如表5所示,可知最优自回归阶数$p_{\text{opt}} = 0$,最优分段个数$N_{\text{opt}} = 4$. 动态规划得到相应的全局最优分段位置为19、117、189、200,已标示在图8中,且方法总的执行时间为21.53 s. 此外,用改进前的SMTS-DP方法对原17维多元时间序列进行分段,得到最大自回归阶数$p_{\max} = 0$. 表6展示了$p = 0$时,分段个数N分别取2、3、4、5计算得到的BIC值,可以看出使BIC值达到最小的最优分段个数$N_{\text{opt}} = 2$. 经过回溯,得到全局最优分段位置为117、200,且执行时间为24.07 s. 然而,与真实分段位置对比可知,SMTS-DP方法只近似找到了原始多元时间序列中的两个分段位置,但SMTS-FD方法却能够近似捕捉到全部分段位置. 总体而言,本文所提改进后的SMTS-FD方法对于变量个数较多的多元时间序列分段精度和执行效率均更高,具有更好的适应性.

表5 SMTS-FD方法应用于HAR数据集得到的BIC值

p	N			
	2	3	4	5
0	−1.91	−2.01	−2.20	−2.13
1	−0.41	1.01	2.39	3.87
2	2.21	4.97	7.69	10.41

表6 SMTS-DP方法应用于HAR数据集得到的BIC值

p	N			
	2	3	4	5
0	0.46	0.69	0.85	1.22

3.3 参数对分段结果的影响

以上实验中设置了自回归阶数上限$p_{\text{set}} = 5$、最大分段个数$N_{\max} = 5$,且在3个多元时间序列数据集上均得到了较好的分段结果. 然而,实际应用中并不总是能够根据经验事先设置好合适的参数值,为此,

这里对SMTS-FD方法的参数设置进行评估.

对于自回归阶数$p = 0, 1, \cdots, p_{\text{set}}$,使AIC值达到最小的$p$即为降维后的低维多元时间序列的最大向量自回归阶数. 图10给出了$p_{\text{set}} = 20$时的一个例子,展示的是AIC准则在合成数据集、HY、HAR数据集上各自的计算结果,且最小值点处对应的AIC值已标注在图10中. 可以看出,p从0变化到20的过程中,每个数据集对应的AIC值曲线在达到最小后的变化趋势都是近似线性增长的,表明自回归阶数上限p_{set}的设置并不会对AIC准则找出的最大自回归阶数$p_{\max}$产生影响.

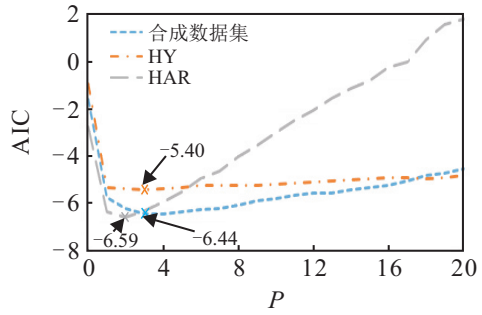

图 10　自回归阶数对分段结果的影响

由AIC准则得到最优的$p_{\max}$之后,对于$p = 0, 1, \cdots, p_{\max}$和$N = 2, 3, \cdots, N_{\max}$,利用动态规划来分段低维多元时间序列,将得到的每种情况下的最小分段代价及相应的p、N代入BIC计算公式,即可通过最小化BIC值得到最优的自回归阶数p_{opt}和分段个数N_{opt}. 以合成数据集为例,序列的长度$T = 100$意味着$N_{\max}$最大可设定为100,即每个数据点都是一个分段点. 但实际分段过程几乎不会出现一个数据点自成一段的情况,因此这里假定每个段至少包含2个数据点,设置$N_{\max} = 50$,得到如图11所示的BIC值变化曲线. 可以观察到,当分段个数逐渐增大时,不论拟合模型的自回归阶数是几,计算出的BIC值均是在不断增大的,且$p = 1$、$N = 2$时的BIC值仍然是最小的,表明最大分段个数$N_{\max}$的设置不会影响用BIC准则找到最优分段个数N_{opt}和自回归阶数p_{opt}.

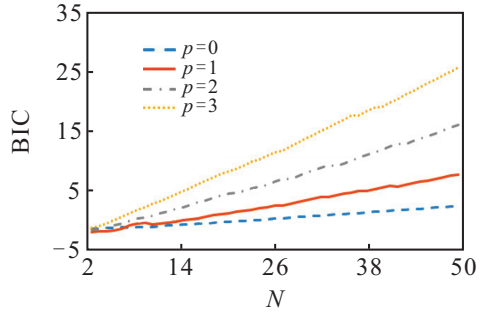

图 11　合成数据集分段个数对分段结果的影响

综上所述,为了保证能够找到最优的分段结果,参数p_{set}和$N_{\max}$的初始值可以设置得大一些,对最终分段结果没有影响.

4　结　论

本文在已有的动态规划分段方法基础上进行改进,提出了一种基于因子分析和动态规划的多元时间序列分段方法SMTS-FD. 考虑到动态因子模型具有反映多元时间序列变化共性的优势,SMTS-FD方法从整体变化趋势相似的变量序列中提取共同因子序列构成原始多元时间序列的低维表达,有效地降低了序列的冗余情况. 具有相似变化趋势的变量序列簇由增量聚类过程自动获得,使得因子分析后得到的低维多元时间序列能够最大限度反映原始多元时间序列的共同变化趋势,保证了后续在低维架构上实现高维序列分段的精确性. 实验结果表明,SMTS-FD方法对于不同的参数设置具有良好的适应性,与原始的动态规划分段方法相比,更适合对包含变量个数较多的多元时间序列进行分段,且分段精度和执行效率均得到改善. 然而,本文所提SMTS-FD分段方法只适用于静态的多元时间序列,未来将研究如何用此方法分段动态多元时间序列,并应用到序列的离散化等领域.

参考文献(References)

[1] Sadri A, Ren Y, Salim F D. Information gain-based metric for recognizing transitions in human activities[J]. Pervasive and Mobile Computing, 2017, 38: 92-109.

[2] Jamali S, Jönsson P, Eklundh L, et al. Detecting changes in vegetation trends using time series segmentation[J]. Remote Sensing of Environment, 2015, 156: 182-195.

[3] Dobos L, Abonyi J. On-line detection of homogeneous operation ranges by dynamic principal component analysis based time-series segmentation[J]. Chemical Engineering Science, 2012, 75: 96-105.

[4] Abonyi J, Feil B, Nemeth S, et al. Modified Gath-Geva clustering for fuzzy segmentation of multivariate time-series[J]. Fuzzy Sets and Systems, 2005, 149(1): 39-56.

[5] Chamroukhi F, Mohammed S, Trabelsi D, et al. Joint segmentation of multivariate time series with hidden process regression for human activity recognition[J]. Neurocomputing, 2013, 120: 633-644.

[6] Cho H. Change-point detection in panel data via double CUSUM statistic[J]. Electronic Journal of Statistics, 2016, 10(2): 2000-2038.

[7] Lin J, Keogh E, Lonardi S, et al. A symbolic representation of time series, with implications for streaming algorithms[C]. Proceedings of the 8th ACM

SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego: ACM, 2003: 2-11.

[8] 李海林, 郭崇慧. 基于云模型的时间序列分段聚合近似方法[J]. 控制与决策, 2011, 26(10): 1525-1529.
(Li H L, Guo C H. Piecewise aggregate approximation method based on cloud model for time series[J]. Control and Decision, 2011, 26(10): 1525-1529.)

[9] 周大镯, 李敏强. 基于序列重要点的时间序列分割[J]. 计算机工程, 2008, 34(23): 14-16.
(Zhou D Z, Li M Q. Time series segmentation based on series importance point[J]. Computer Engineering, 2008, 34(23): 14-16.)

[10] Kehagias A, Nidelkou E, Petridis V. A dynamic programming segmentation procedure for hydrological and environmental time series[J]. Stochastic Environmental Research and Risk Assessment, 2006, 20(1/2): 77-94.

[11] Adams R P, MacKay D J C. Bayesian online changepoint detection[R]. Cambridge: University of Cambridge, 2007.

[12] 高华川, 张晓峒. 动态因子模型及其应用研究综述[J]. 统计研究, 2015, 32(12): 101-109.
(Gao H C, Zhang X D. A survey of dynamic factor model and its applications[J]. Statistical Research, 2015, 32(12): 101-109.)

[13] Geweke J. The dynamic factor analysis of economic time series[J]. Latent Variables in Socio-economic Models, 1977: 365-383.

[14] Sun Z, Liu X, Wang L. A hybrid segmentation method for multivariate time series based on the dynamic factor model[J]. Stochastic Environmental Research and Risk Assessment, 2017, 31(6): 1291-1304.

[15] Barigozzi M, Cho H, Fryzlewicz P. Simultaneous multiple change-point and factor analysis for high-dimensional time series[J]. Journal of Econometrics, 2018, 206(1): 187-225.

[16] Gedikli A, Aksoy H, Unal N E, et al. Modified dynamic programming approach for offline segmentation of long hydrometeorological time series[J]. Stochastic Environmental Research and Risk Assessment, 2010, 24(5): 547-557.

[17] Gedikli A, Aksoy H, Unal N E. Segmentation algorithm for long time series analysis[J]. Stochastic Environmental Research and Risk Assessment, 2008, 22(3): 291-302.

[18] Guo H, Liu X, Song L. Dynamic programming approach for segmentation of multivariate time series[J]. Stochastic Environmental Research and Risk Assessment, 2015, 29(1): 265-273.

[19] Widiputra H, Pears R, Kasabov N. Dynamic learning of multiple time series in a nonstationary environment[M]. New York: Springer, 2012: 303-347.

[20] Bai J, Ng S. Large dimensional factor analysis[J]. Foundations and Trends in Econometrics, 2008, 3(2): 89-163.

[21] Akaike H. A new look at the statistical model identification[J]. IEEE Transactions on Automatic Control, 1974, 19(6): 716-723.

[22] Schwarz G. Estimating the dimension of a model[J]. The Annals of Statistics, 1978, 6(2): 461-464.

[23] Anguita D, Ghio A, Oneto L, et al. A public domain dataset for human activity recognition using smartphones[C]. The 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, 2013: 437-442.

作者简介

王玲 (1974−), 女, 副教授, 博士, 从事数据挖掘、机器学习的研究, E-mail: lingwang@ustb.edu.cn;

徐培培 (1994−), 女, 硕士生, 从事数据挖掘的研究, E-mail: xupeipei_ustb@163.com;

彭开香 (1971−), 男, 教授, 博士生导师, 从事复杂工业系统故障诊断与容错控制等研究, E-mail: kaixiang@ustb.edu.cn.

(责任编辑：齐 霁)