

# 小样本概率关联度模型研究

宁小磊<sup>1</sup>, 吴颖霞<sup>1</sup>, 赵新<sup>1</sup>, 赵军民<sup>2</sup>, 张凯<sup>1</sup>, 郝跳锋<sup>1</sup>,  
吕梅柏<sup>3</sup>, 胡亚峰<sup>1</sup>, 陈韵<sup>2</sup>

(1.中国华阴兵器试验中心, 陕西 华阴 714200; 2.西安现代控制技术研究所, 陕西 西安 710065;  
3.西北工业大学 航天学院, 陕西 西安 710072)

**摘 要:**针对小样本条件下多元时间序列之间的关联分析问题,提出了一种小样本概率关联度模型。使用样本函数或样本统计量代替样本构造累积经验分布函数并计算概率关联系数。当比较序列样本容量  $n<5$ ,使用样本顺序统计量的经验分布函数计算概率关联系数;当比较序列样本容量  $n\geq 5$ ,使用 Bootstrap 方法重抽样扩充样本后再构造经验分布函数计算概率关联系数,改进了概率关联度模型中关联系数的计算方法,解决了小样本条件下这一类多元数据关联分析问题。证明了小样本概率关联度模型满足的基本性质,给出了小样本概率关联分析的基本步骤,仿真案例和实际应用验证了文中模型的正确性和有效性。

**关 键 词:**数据关联;概率关联度;小样本;多元时间序列;样本统计量  
**中图分类号:**TJ761.12      **文献标志码:**A      **文章编号:**1000-2758(2022)05-1164-08

关联分析是发现、查询数据之间关联性或相关性的一种实用分析技术,描述了事物中某些属性同时出现的规律和模式<sup>[1-2]</sup>。关联分析在经济学、军事学、互联网、航空航天及人工智能等领域有着广泛应用<sup>[3-4]</sup>,其中时间序列分析是数据关联分析领域的热点研究内容。

对不同工程背景及应用需求,目前已经提出了多种数据关联分析方法<sup>[5-6]</sup>。文献[1]提出了一种信息不确定条件下的时间序列关联分析法,解决了时间序列中含有不确定信息时的数据关联分析问题。文献[7]提出了一种灰色关联分析方法,解决了时间序列曲线相似性关联分析问题。常见的时间序列关联分析方法还有 TIC 法、相关系数法、误差分析法、EARTH 方法等<sup>[7]</sup>,这些方法操作简单,应用广泛,但主要适用于单样本时间序列之间的分析,不适合工程中经常遇到的多样本(多元)时间序列情况<sup>[4]</sup>。针对多样本关联分析,文献[4]提出了一种概率关联度模型和概率关联分析方法,实现了由序

列曲线关联向立体面板关联的方法拓展,但计算时需要由比较序列样本构建累积分布函数。实际应用中由于试验抽样的昂贵性、复杂性和困难性等,比较序列在实际情况中常常是小样本数据容量,构建的累积分布函数与真实分布差别一般较大,若直接应用概率关联度模型,关联分析效果较差,严重影响了概率关联度模型的工程应用范围。

针对上述问题,本文提出一种小样本概率关联度模型。使用样本函数或样本统计量代替样本构造累积经验分布函数并计算概率关联系数,发展了概率关联度模型的理论框架。当比较序列样本容量  $n<5$ 时,使用样本函数或样本统计量构建累积分布函数计算概率关联系数;当比较序列样本容量  $n\geq 5$ 时,使用 Bootstrap 方法重抽样扩展样本后构建累积分布函数计算概率关联系数,改进了概率关联度模型中关联系数的计算方法,从而提高了概率关联分析的效率。通过仿真案例和实际应用验证了本文方法的正确性和有效性。

# 1 问题描述与分析

假设有参考序列  $\mathbf{x}$  为

$$\mathbf{x} = [x(1), x(2), \dots, x(k), \dots, x(T)] \quad (1)$$

比较序列  $\mathbf{y}$  为

$$\mathbf{y} = \begin{bmatrix} y_1(1), y_1(2), \dots, y_1(k), \dots, y_1(T) \\ y_2(1), y_2(2), \dots, y_2(k), \dots, y_2(T) \\ \vdots \\ y_n(1), y_n(2), \dots, y_n(k), \dots, y_n(T) \end{bmatrix} \quad (2)$$

式中:  $k$  为时刻;  $T$  为序列长度。

参考序列  $\mathbf{x}$  和比较序列  $\mathbf{y}$  之间的数据关联度为  $C$ , 一般地,  $C$  可由如(3)式计算

$$C(\mathbf{x}, \mathbf{y}) = \text{consistency}(\mathbf{x}, \mathbf{y}) \quad (3)$$

式中:  $\text{consistency}(\cdot)$  为一致性分析函数, 也称一致性检验算子, 一般有  $C(\mathbf{x}, \mathbf{y}) \in [0, 1]$ 。文献[4]总结了目前常见的一致性分析函数  $\text{consistency}(\cdot)$ , 并提出了一种适合多样本时间序列的概率关联算子和概率关联分析方法。

参考序列  $\mathbf{x}$  和比较序列  $\mathbf{y}$  之间的概率关联度计算步骤如下<sup>[4]</sup>:

**步骤1** 计算  $k$  时刻概率关联系数  $p_{\text{operator}}(k)$

$$p_{\text{operator}}(k) = F_{y(k)}(x(k)) \quad (4)$$

**步骤2** 计算概率关联度  $p$

$$p = c(p_{\text{operator}}) \quad (5)$$

式中:  $F_{y(k)}(\cdot)$  为由比较序列样本  $y_{1:n}(k)$  确定的累积分布函数;  $F_{y(k)}(x(k))$  为  $x(k)$  在累积分布函数  $F_{y(k)}$  中的累积分布函数值;  $c(\cdot)$  为概率关联度模型中的综合概率度计算公式, 可取均匀检验结果。

从(4)~(5)式可见, 步骤1中需要由比较序列样本  $y_{1:n}(k)$  构建累积分布函数  $F_{y(k)}$ , 当比较序列  $y_{1:n}(k)$  样本容量较大时(样本量  $n$  较大时), 可以构建比较精确的累积分布函数  $F_{y(k)}$ , 综合计算得到的概率关联度  $p$  比较符合实际。当比较序列  $y_{1:n}(k)$  样本容量较小时, 由比较序列  $y_{1:n}(k)$  构建的累积分布函数  $F_{y(k)}$  因抽样误差与实际分布函数差距较大, 此时, 若应用综合计算概率关联度  $p$  进行分析决策风险较大。工程实际中, 由于试验抽样的昂贵性、复杂性和困难性, 经常会遇到小样本情况的比较序列数据, 限制了概率关联度模型的实际应用效果和应用范围。

# 2 小样本概率关联度模型

## 2.1 改进思路

综上可知, 小样本条件下概率关联度模型应用遇到的主要问题是: 比较序列  $y_{1:n}(k)$  样本容量  $n$  有限, 基于比较序列样本  $y_{1:n}(k)$  不容易构建精确的累积分布函数  $F_{y(k)}$ , 导致概率关联度模型使用效果较差。解决该问题的有效方法是使用小样本比较序列  $y_{1:n}(k)$  构建满足要求的累积分布函数  $F_{y(k)}$ 。

Bootstrap 方法是 Efron 于 1979 年提出的一种逼近复杂系统统计量估计值分布的统计方法<sup>[8]</sup>, 是目前被广泛采用的小样本数据处理方法, 可以用来解决概率关联度模型面临的小样本问题, 即使用 Bootstrap 方法重抽样扩展样本后构建累积分布函数计算概率关联系数。但通常 Bootstrap 方法样本容量以  $n \geq 5$  较合适<sup>[9-10]</sup>, 当  $n < 5$  时, 其计算结果的任意性很大, 构造的经验分布函数和统计效果较差。针对样本容量  $n < 5$  的情况, 文献[3, 9]使用了一种样本顺序比率统计量  $K$ , 在样本容量  $n < 5$  的条件下应用效果较好<sup>[3]</sup>, 认为相对于样本总体分布函数或经验分布函数, 使用样本顺序比率统计量  $K$  分布函数相容性检验效率更高, 可见小子样本构建样本顺序比率统计量  $K$  经验分布函数比小子样本构建样本经验分布函数精度要好。因此, 可以使用样本顺序比率统计量  $K$  的经验分布函数代替概率关联度模型中的经验分布函数。

## 2.2 变量函数概率关联度模型改进

文献[4]提出的概率关联度模型, 使用比较序列样本  $y_{1:n}(k)$  直接构建经验分布函数, 但样本顺序比率统计量  $K$  是比较序列样本的函数, 不能直接使用。为了解决该问题, 拓展概率关联度模型的使用范围, 提出了变量函数概率关联度改进模型, 使其能同时应用于样本或样本函数构建经验分布函数。

为便于描述, 首先引入 2 个引理。

**引理1** 设  $\mathbf{X}$  是一连续随机变量, 其分布函数为  $F(\mathbf{X})$ , 则  $F(\mathbf{X})$  服从  $[0, 1]$  上的均匀分布。

**引理2** 设  $\mathbf{Y} = f(\mathbf{X})$  是一连续随机变量, 其分布函数为  $F(f(\mathbf{X}))$ , 则  $F(f(\mathbf{X}))$  服从  $[0, 1]$  上的均匀分布。

从上述 2 个引理可见, 变量或变量函数(变量统计量)均可以代人概率关联度模型中计算概率关联系数, 因此对基本概率关联度模型改进如下:

**步骤1** 计算  $k$  时刻概率关联系数  $p_{\text{operator}}(k)$

$$p_{\text{operator}}(k) = F_{y(k)}(f(x(k))) \quad (6)$$

**步骤 2** 计算概率关联度  $p$ , 算子如下

$$p = c(p_{\text{operator}}) \quad (7)$$

式中:  $f(\cdot)$  为样本函数或样本统计量, 一般为线性或非线性函数关系式, 根据问题背景灵活选择。

变量函数概率关联度改进模型, 可以采用样本或样本函数构建经验分布函数, 来满足小样本概率关联度模型改进使用需求。

注 1: 通过上述改进, 概率关联度模型适应性更广, 但需要注意的是, 并不是每一种样本/变量函数  $f(\cdot)$  都可以参与计算概率关联系数, 这是因为函数计算会引入误差, 或者说抽样误差会通过函数计算放大, 所以选择合适的变量函数或样本统计量很关键。(虽然很难, 但仍可以寻找到一些有用统计量, 即样本函数  $f(\cdot)$ , 在小样本条件下应用效果较好, 比如文献[9] 找到的样本顺序比率统计量  $K$ 。)

## 2.3 经验分布函数构造

### 2.3.1 经验分布函数

设  $x(1), x(2), x(3), \dots, x(n)$  为来自分布函数  $F$  的随机样本, 其经验分布函数  $F_n(x)$  定义为<sup>[11]</sup>

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{[x(i) \leq x]} = \frac{\#\{x(i) \leq x, i=1, 2, \dots, n\}}{n} \quad (8)$$

式中:  $I_{[\cdot]}$  为示性函数;  $\#A$  为集合  $A$  中元素的个数。 $F_n(x)$  为  $x$  的右函数, 共有  $n$  个跳跃点, 跳跃度为  $1/n$ , 即  $F_n(x_i) - F_n(x_{i-1}) = 1/n, i=1, \dots, n$ , 且有  $F_n(-\infty) = 0, F_n(+\infty) = 1$ 。

### 2.3.2 样本容量 $n < 5$ 经验分布函数构造

设样本  $\mathbf{x} = [x(1), x(2), x(3), \dots, x(n+1)]$  (其中 1 个样本模拟参考序列样本, 其他样本模拟比较序列样本), 对样本按自小至大顺序排列, 得到样本顺序统计量  $\mathbf{x}' = [x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n+1)}]$ , 则样本顺序比率统计量  $K_{ijk} = (x_{(j)} - x_{(i)}) / (x_{(k)} - x_{(i)}), 1 \leq i < j < k \leq n+1, 0 \leq K_{ijk} \leq 1$ 。

当  $n = 2$  时, 有 1 个统计量  $K_{123} = (x_{(2)} - x_{(1)}) / (x_{(3)} - x_{(1)})$ , 且  $0 \leq K_{123} \leq 1$ , 对于样本  $\mathbf{x}$  来自总体  $\mathbf{X} \sim N(\mu, \sigma^2)$ ,  $K_{123}$  的累积分布函数为<sup>[3,9]</sup>

$$f(K) = \frac{3}{\pi} \left( \arctan \left( \frac{2K-1}{\sqrt{3}} \right) + \frac{\pi}{6} \right) \quad (9)$$

对于样本  $\mathbf{x}$  来自其他形式的情况,  $K_{123}$  的累积分布函数见文献[9,12]。当  $n = 3$  时, 有  $C_4^3 = 4$  个统计量, 分别为  $K_{123}, K_{124}, K_{134}, K_{234}$ ; 当  $n = 4$  时, 有  $C_5^3 =$

10 个统计量, 分别为  $K_{123}, K_{124}, K_{125}, K_{134}, K_{135}, K_{145}, K_{234}, K_{235}, K_{245}, K_{345}$ ; 当  $n = 5$  时, 有  $C_6^3 = 20$  个统计量, 分别为  $K_{123}, K_{124}, K_{125}, K_{126}, K_{134}, K_{135}, K_{136}, K_{145}, K_{146}, K_{156}, K_{234}, K_{235}, K_{236}, K_{245}, K_{246}, K_{256}, K_{345}, K_{346}, K_{356}, K_{456}$ 。对于样本量  $n = 3, 4$  等情况或  $n$  更大时, 无法给出  $K_{ijk}$  的具体理论分布解析式, 但可以采用数值模拟方法<sup>[3]</sup> 得到  $K_{ijk}$  的累积分布函数, 然后计算参考样本在其中的函数值, 得到概率关联系数。

### 2.3.3 样本容量 $n \geq 5$ 经验分布函数构造

#### 2.3.3.1 经典 Bootstrap 方法步骤

**步骤 1** 观测样本  $\mathbf{x} = [x(1), x(2), \dots, x(n)]$  的值为有限的总体样本, 把  $n$  个样本按值自小至大顺序编号排列, 可得到样本的顺序统计量  $\mathbf{x}' = [x_{(1)}, x_{(2)}, \dots, x_{(n)}]$ , 式中  $x_{(1)} = \min(x(i)), x_{(2)} = \text{second smallest}(x(i)), \dots, x_{(n)} = \max(x(i))$ , 其中,  $i = 1, 2, \dots, n$ 。用最简洁算法得  $i$  处的累积概率分布函数值为  $F_n(i) = \frac{i}{n}$ , 由此可构造原始样本的累积经验分布函数为

$$F_n(\mathbf{x}) = \begin{cases} 0, & x \leq x_{(1)} \\ \frac{i}{n}, & x_{(i)} < x < x_{(i+1)} \\ 1, & x \geq x_{(n)} \end{cases} \quad (10)$$

**步骤 2** 根据原始样本的累积经验分布函数, 得到  $N$  组随机抽样集合  $\mathbf{x}^k = (x_1^k, x_2^k, \dots, x_n^k)$ , 其中,  $k = 1, 2, \dots, N, \mathbf{x}^k$  称为 Bootstrap 样本, 或自助样本。

**步骤 3** 设参数  $\theta$  是总体的一个未知参数,  $F_n(\mathbf{x})$  是由样本得到的累积经验分布函数(样本量为  $n$ ), 直接可由  $F_n(\mathbf{x})$  得到未知参数  $\theta$  的估计  $\hat{\theta}(F_n(\mathbf{x}))$ , 这种直接估计的估计误差  $T_n$  可表示为

$$T_n = \hat{\theta}(F_n(\mathbf{x})) - \theta(F) \quad (11)$$

根据 Bootstrap 样本集合也可得到参数  $\theta$  的估计  $\hat{\theta}(F_n^k(\mathbf{x}))$ , 偏差可表示为

$$T_n^k = \hat{\theta}(F_n^k(\mathbf{x})) - \theta(F) \quad (12)$$

式中, 由于真实的  $\theta(F)$  未知, 可由  $\hat{\theta}(F_n(\mathbf{x}))$  替代  $\theta(F)$ ,  $T_n^k$  可近似表达为

$$T_n^k = \hat{\theta}(F_n^k(\mathbf{x})) - \hat{\theta}(F_n(\mathbf{x})) \quad (13)$$

根据统计理论, 用  $T_n^k$  分布逼近  $T_n$  分布, 当得到大量  $\theta(F)$  样本后, 运用这些样本即可统计求出参数  $\theta$  估计值的分布及其特征值。

#### 2.3.3.2 Bootstrap 方法改进

Bootstrap 方法通过大量再生子样进行统计推

断,缓解了小样本问题,但经典的 Bootstrap 方法的采样方式具有一定局限性,主要是:①样本的累积经验分布函数将样本的取值范围限制在  $[x_{(1)} \ x_{(n)}]$  中,且样本的取值是离散的,对于连续取值的变量无法获取样本点之外的信息。②从公式(10)可见,当  $i = n$  或  $x = x_{(n)}$  时,有  $p_n = 1$ ,但理论上应是当  $n \rightarrow \infty$ ,才有  $p_n = 1$ ;同理  $x = x_{(1)}$ ,有  $p_n = 0$ ,但理论上应是当  $x \rightarrow -\infty$ ,有  $p_n = 0$ 。图 1 给出不同样本构造的经验分布函数及真实分布函数对比。为了普适性应用概率关联度模型,对基本 Bootstrap 方法进行修正,改进的主要思路是:

1) 使用样条函数代替原经验分布函数构造使用的阶跃连接,从而解决了对于连续取值的变量无法获取样本点之外的信息问题。

2) 根据 DKW (Dvoretzky-Kiefer-Wolfowitz) 不等式,对经验分布函数的尾概率进行了刻画,由 DKW 不等式  $\Pr(\sup_x |F_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}$  得到

$\varepsilon_n = \sqrt{\frac{1}{2n} \ln(\frac{2}{\alpha})}$ ,使用  $\min(F_n(x_{(1)}) + \varepsilon_n, F_n(x_{(2)}))$  左平滑计算  $F_n(x)$ ,解决  $x = x_{(1)}$ ,有  $p_n = 0$  的问题;使用  $\max(F_n(x_{(n)}) - \varepsilon_n, F_n(x_{(n-1)}))$  右平滑计算  $F_n(x)$ ,解决  $x = x_{(n)}$ ,有  $p_n = 1$  的问题。

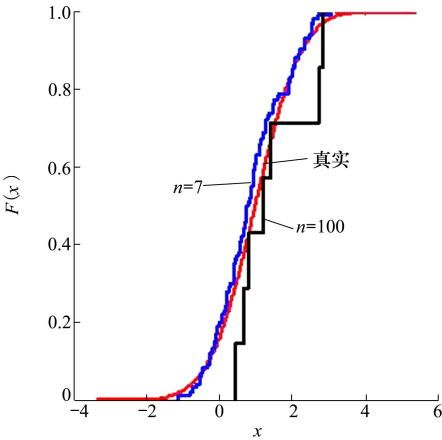


图 1 经验分布函数

2.4 小样本概率关联度模型

按照以下步骤计算参考序列  $x$  和比较序列  $y$  之间的小样本概率关联度。

步骤 1 计算  $k$  时刻概率关联系数  $p_{operator}(k)$

$$p_{operator}(k) = \begin{cases} F_{y(k)}(x(k)), & \text{样本} \\ F_{y(k)}(f(x(k))), & \text{样本函数} \end{cases} \quad (14)$$

步骤 2 计算概率关联度  $p$

$$p = c(p_{operator}) \quad (15)$$

2.5 小样本概率关联度模型的基本性质

性质 1 小样本概率关联度具有以下基本性质。

- 1) 规范性,即  $0 \leq p(x,y) \leq 1$ ;
- 2) 整体性,对于不同的相关因素序列  $x_i, x_j$ , 一般有  $p(x_i, x_j) \neq p(x_j, x_i)$ ,  $i \neq j$ ;
- 3) 可比性和唯一性;
- 4) 干扰因素独立性。

性质 2 概率关联度不满足偶对称性,即  $x = \{x,y\}$ ,有  $p(x,y) \neq p(y,x)$ 。

性质 3 概率关联度模型不满足数乘变换一致性和平移变换一致性。

3 适用于小样本问题的概率关联分析步骤

步骤 1 在相同初始条件下,分别得到参考序列  $x$  和比较序列  $y$ 。

$$x = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(k) & \cdots & x_1(T) \\ x_2(1) & x_2(2) & \cdots & x_2(k) & \cdots & x_2(T) \\ \vdots & \vdots & & \vdots & & \vdots \\ x_m(1) & x_m(2) & \cdots & x_m(k) & \cdots & x_m(T) \end{bmatrix} \quad (16)$$

$$y = \begin{bmatrix} y_1(1) & y_1(2) & \cdots & y_1(k) & \cdots & y_1(T) \\ y_2(1) & y_2(2) & \cdots & y_2(k) & \cdots & y_2(T) \\ \vdots & \vdots & & \vdots & & \vdots \\ y_n(1) & y_n(2) & \cdots & y_n(k) & \cdots & y_n(T) \end{bmatrix} \quad (17)$$

式中:  $m$  为参考序列  $x_{1:m}(k)$  样本容量,  $n$  为比较序列  $y_{1:n}(k)$  样本容量,此处可以使用矩阵型概率关联度模型<sup>[4]</sup>。

步骤 2 对参考序列  $x$  和比较序列  $y$  进行预处理,使其满足等步长、等长度的数据序列要求。

步骤 3 计算  $k$  时刻概率关联系数  $p_{operator}(k)$

- 1) 当  $y_{1:n}(k)$  样本容量  $n < 5$  时
  - ① 构建样本顺序比率统计量  $K_{ijk} = (x_{(j)} - x_{(i)}) / (x_{(k)} - x_{(i)})$ ,  $1 \leq i < j < k \leq n + 1$ ,  $0 \leq K_{ijk} \leq 1$ ,并基于样本估计  $K$  的经验分布函数  $F_{y_{1:n}}(K)$ 。
  - ② 将参考样本  $x(k)$  带入经验分布函数  $F_{y_{1:n}}(K)$  计算累积分布函数值,得到关联关系

数  $p_{operator}(k)$ 。

2) 当  $y_{1:n}(k)$  样本容量  $n \geq 5$  时

① 对比较样本  $y_{1:n}(k)$  使用 Bootstrap 方法进行重抽样得到重抽样样本  $y'_{1:n}(k)$ , 根据重抽样样本  $y'_{1:n}(k)$  估计  $y$  的经验分布函数  $F_{y_{1:n}}(y)$ 。

② 将参考样本  $x(k)$  带入经验分布函数  $F_{y_{1:n}}(K)$  计算累积分布函数值, 得到关联系数  $p_{operator}(k)$ 。

**步骤 4** 决策。检验  $p_{operator}$  在一定置信水平  $\alpha$  下是否服从  $[0,1]$  上的均匀分布。若通过检验, 说明通过关联分析。否则, 未通过概率关联分析。

4 仿真测试与分析

通过几个仿真测试案例验证本文改进模型的正确性和有效性, 仿真案例分别测试样本容量  $n = 2(K_{ijk}$  有具体理论分布解析式)、 $n = 3(K_{ijk}$  没有具体理论分布解析式) 和样本容量  $n \geq 5$  时(Bootstrap 方法改进) 的应用场景。

**仿真 1** 参考时间序列  $X$  和比较时间序列

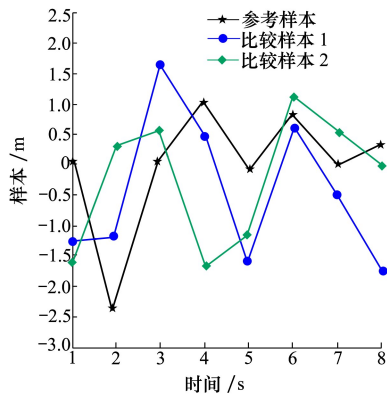
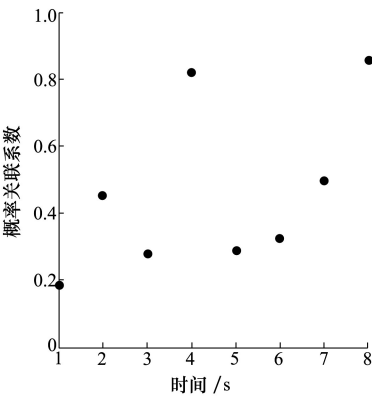
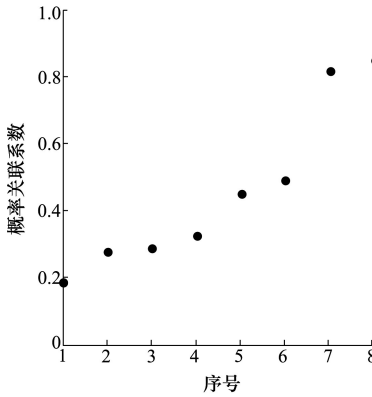


图 2 样本序列



a) 按照时间序列排序后的概率关联系数



b) 按照大小排序后的概率关联系数

图 3 概率关联系数

**仿真 2** 参考时间序列  $X$  和比较时间序列  $Y$  如下:

$X = [0.503 \quad 1.231 \quad 1.103 \quad 0.078 \quad 0.447 - 0.280$   
 $- 0.419 - 0.384 - 1.851 - 0.406];$

$Y = [0.607 \quad 0.337 \quad 0.119 \quad 0.311 \quad 0.623 - 0.824$   
 $1.801 - 2.039 \quad 0.029 \quad 0.745; 0.333 - 0.352$   
 $0.356 - 0.914 \quad 1.588 \quad 1.765 \quad 1.429 - 1.196$   
 $- 0.479 \quad 1.567; 0.985 \quad 0.968 \quad 1.506 \quad 1.341$   
 $1.526 \quad 0.403 - 1.318 - 0.110 \quad 0.567 \quad 0.213]$

$Y$  为:

$X = [0.053 \quad 0.059 - 1.912 \quad 0.402 \quad 1.837 - 0.618$   
 $- 0.043 \quad 1.125];$

$Y = [-0.437 - 2.843 \quad 0.846 \quad 0.196 \quad 0.870 - 0.524$   
 $1.573 - 0.191; -0.014 - 0.465 - 0.950 - 1.404$   
 $- 1.257 \quad 0.447 - 1.328 - 1.241]$

检验参考时间序列  $X$  和比较时间序列  $Y$  的一致性。

比较时间序列  $Y$  样本容量  $n = 2$ , 有 1 个统计量  $K_{123} = (x_{(2)} - x_{(1)}) / (x_{(3)} - x_{(1)})$ , 且  $0 \leq K_{123} \leq 1$ , 累积分布函数为

$$f(K) = \frac{3}{\pi} \left( \arctan \left( \frac{2K - 1}{\sqrt{3}} \right) + \frac{\pi}{6} \right) \quad (18)$$

图 2 给出了参考时间序列和比较时间序列样本; 将参考时间序列  $X$  带入(18) 式所示累积分布函数, 得到各个时刻的概率关联系数, 如图 3 所示; 图 3b) 给出了排序后的概率关联系数。最后计算得到的概率关联度  $P = 0.608 \ 3$ , 结果:  $H = 0$ 。其中,  $H$  表示仿真模型验证结果(kstest 函数的计算返回值), 当  $H = 0$  时, 表示两组数据一致; 当  $H = 1$  时, 表示两组数据不一致。

检验参考时间序列  $X$  和比较时间序列  $Y$  的一致性。

由于比较时间序列  $Y$  样本容量  $n = 3$  时, 样本顺序比率统计量  $K$  没有显式表达式, 首先通过数值模拟的方法给出了累积概率密度分布如图 4 所示。得到累积分布函数后, 计算参考样本在其中的函数值便可得到概率关联系数。图 5 给出了各个时刻的概率关联系数。最后计算得到的概率关联度  $P = 0.716 \ 1$ , 结果:  $H = 0$ 。

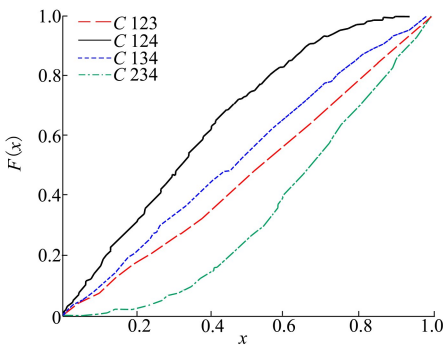


图 4  $n = 3$  时的累积概率密度分布

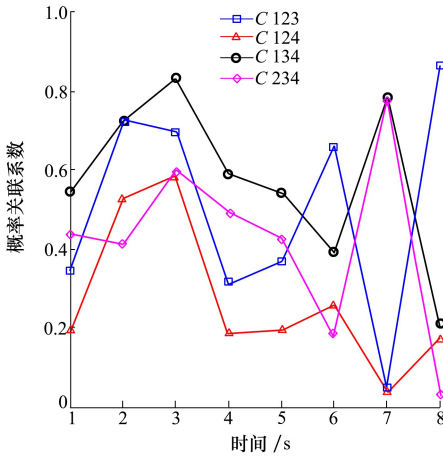


图 5 概率关联系数

**仿真 3** 参考时间序列  $X$  和比较时间序列  $Y$  如下:

$X = [-1.769 \ 2.155 \ 2.174 \ -0.493 \ -0.268 \ -0.070$   
 $\quad -0.803 \ -0.992 \ -1.147 \ -1.917];$

$Y = [-1.241 \ 1.583 \ 2.739 \ 0.508 \ 2.029 \ 1.004$   
 $\quad -0.348 \ -0.546 \ 2.404 \ 0.709; -1.874 \ 2.420$   
 $\quad 2.878 \ 0.228 \ 0.493 \ 0.434 \ -1.546 \ -0.184$   
 $\quad -0.560 \ 0.118; -1.233 \ 2.319 \ 2.485 \ 0.025$   
 $\quad -0.559 \ -0.757 \ -0.210 \ -0.699 \ -1.800 \ -1.646;$   
 $\quad -1.027 \ 2.307 \ 2.529 \ -1.065 \ 1.811 \ 0.859$   
 $\quad 1.098 \ 0.295 \ -0.840 \ -0.931; -1.042 \ 1.667$   
 $\quad 2.707 \ 0.148 \ -0.364 \ 0.194 \ -0.355 \ -1.012$   
 $\quad 1.490 \ -0.653; -1.619 \ 2.085 \ 2.432$   
 $\quad 1.055 \ -0.943 \ -0.559 \ -1.386 \ 0.343 \ -0.475$   
 $\quad 0.568; -1.616 \ 2.210 \ 2.019 \ 1.827 \ 1.334$   
 $\quad 1.828 \ -0.350 \ 0.440 \ 0.479 \ -0.776]$

检验参考时间序列  $X$  和比较时间序列  $Y$  的一致性。

由于  $k$  时刻比较序列  $Y$  样本容量  $n \geq 5$ , 首先使用 Bootstrap 方法重抽样, 然后基于重抽样样本构建

累积分布函数, 部分结果如图 6 所示。由图 6 可见, 改进的自助样本构建的累积概率分布函数明显比由原始样本构建的累积概率分布函数光滑。各个时刻的概率关联系数计算结果为  $[0.162 \ 0.455 \ 0.164 \ 0.152 \ 0.911 \ 0.295 \ 0.286 \ 0.115 \ 0.116 \ 0.015]$ , 最后计算得到的概率关联度  $P = 0.073$ , 结果:  $H = 1$ 。

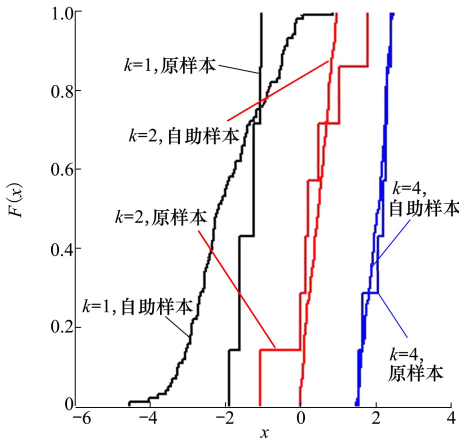


图 6 部分时刻累积分布函数

**仿真 4** 参考时间序列  $X_{1 \times 10}$  和比较时间序列  $Y_{2 \times 10}$  均来自正态分布  $N(0,1)$ , 检验参考时间序列  $X_{1 \times 10}$  和比较时间序列  $Y_{2 \times 10}$  的一致性。分别采用概率关联分析<sup>[4]</sup>和本文改进的小样本概率关联分析进行检验, 验证本文所提方法的有效性, 共进行 10 000 次蒙特卡洛实验。很显然, 由于  $X_{1 \times 10}$  和  $Y_{2 \times 10}$  均来自正态分布  $N(0,1)$ ,  $X_{1 \times 10}$  和  $Y_{2 \times 10}$  是一致的, 检

验结果应为:  $H = 0$ 。定义误检率  $\tau = \frac{\sum_{i=1}^T H}{T}$  为检验方法的评价指标,  $T$  为实验总次数。运行 10 000 次蒙特卡洛实验后的误检率分别为: 本文改进的小样本概率关联分析方法的误检率  $\tau_1 = 0.1998$ 、概率关联分析方法的误检率  $\tau_2 = 0.4725$ , 可见本文方法的误检率明显低于概率关联分析方法, 说明本文方法的有效性。

## 5 弹道一致性检验应用

炮射导弹是由坦克炮发射的一种精确制导武器, 提高了坦克炮的远距离精确打击能力。研究不同状态下炮射导弹的弹道一致性有利于部队训练使用。由于炮射导弹价格的昂贵性, 现场试验组织的

复杂性,现场飞行试验样本容量一般为小子样。假设高原、平原 2 种状态下试验数据如表 3 所示,根据炮射导弹的弹道特征<sup>[13]</sup>,选择了 3 个典型弹道特征点(第一波谷、第一波峰、平稳点)进行一致性检验。

表 1 试验数据

| 状态 | 特征点 1    | 特征点 2   | 特征点 3   |
|----|----------|---------|---------|
| S1 | -1.695 4 | 1.802 8 | 0.378 4 |
|    | -1.810 3 | 2.041 7 | 0.860 0 |
|    | -1.806 6 | 1.650 9 | 2.197 9 |
| S2 | -1.317 8 | 0.853 7 | 0.593 6 |

表 2 检验结果

| K                | 特征点 1   | 特征点 2   | 特征点 3   | 结果            |
|------------------|---------|---------|---------|---------------|
| K <sub>123</sub> | 0.032 9 | 0.388 7 | 0.452 8 | H=0,P=0.221 8 |
| K <sub>124</sub> | 0.007 7 | 0.277 7 | 0.446 8 | H=0,P=0.211 4 |
| K <sub>134</sub> | 0.233 4 | 0.714 4 | 0.986 8 | H=0,P=0.660 9 |
| K <sub>234</sub> | 0.227 5 | 0.604 6 | 0.976 1 | H=0,P=0.873 6 |

检验结果见表 2,有理由认为这 2 种状态满足弹道一致性,这与实际现场试验结果相符。

6 结 论

针对小样本条件下的数据关联分析需求,研究提出了一种小样本概率关联度模型。使用样本函数或样本统计量的累积经验分布函数代替由样本构造的累积经验分布函数,发展拓展了概率关联度模型。当比较序列样本容量  $n<5$ ,使用样本样本顺序统计量的经验分布函数计算概率关联系数;当比较序列样本容量  $n\geq 5$ ,使用 Bootstrap 方法重抽样构造经验分布函数计算概率关联系数,解决了小样本比较序列构造的累积概率分布函数误差较大导致概率关联系数计算不准确的问题。仿真案例和实际应用验证了本文方法的合理性和有效性。

参考文献:

[1] 高亮,孙卫,朱荣昌.信息不确定条件下时间序列的关联分析法[J].西安交通大学学报,2010,44(6):67-71  
GAO Liang, SUN Wei, ZHU Rongchang. A relational analysis approach of time series with uncertain information[J]. Journal of Xi'an Jiaotong University, 2010, 44(6): 67-71 (in Chinese)

[2] 张培忠.制导炸弹仿真试验技术[M].北京:国防工业出版社,2019  
ZHANG Peizhong. Guided bomb simulation test technology[M]. Beijing: National Defense Industry Press, 2019 (in Chinese)

[3] 张领科,赵峰,余永刚.一种小子样情况下的弹道一致性评定方法[J].兵工学报,2014,35(12):2124-2128  
ZHANG Lingke, ZHAO Feng, YU Yonggang. An evaluation method of ballistic consistency under small sample circumstance [J]. Acta Armamentarii, 2014, 35(12): 2124-2128 (in Chinese)

[4] 宁小磊,赵新,吴颖霞,等.基于概率关联分析的仿真模型验证方法研究[J].西北工业大学学报,2021:39(5):1158-1167  
NING Xiaolei, ZHAO Xin, WU Yingxia, et al. Research on simulation model validation based on probability relational analysis [J]. Northwestern Polytechnical University, 2021: 39(5): 1158-1167 (in Chinese)

[5] 林圣琳,李伟,杨明,等.考虑相关性的多元输出仿真模型验证方法[J].自动化学报,2019,45(9):1666-1678  
LIN Shenglin, LI Wei, YANG Ming, et al. Multivariate validation method under correlation for simulation model[J]. Acta Automatica Sinica, 2019, 45(9): 1666-1678 (in Chinese)

[6] 李伟,焦松,陆凌云,等.基于特征差异的仿真模型验证及选择方法[J].自动化学报,2014,40(10):2134-2144  
LI Wei, JIAO Song, LU Lingyun, et al. Validation and selection of simulation model based on the feature differences[J]. Acta Automatica Sinica, 2014, 40(10): 2134-2144 (in Chinese)

[7] 胡玉伟,马萍,杨明,等.基于改进灰色关联分析的仿真数据综合一致性检验方法[J].北京理工大学学报,2013,33(7):711-715  
HU Yuwei, MA Ping, YANG Ming, et al. A comprehensive consistency test method based on improved grey relational analysis for simulation results[J]. Journal of Beijing Institute of Technology, 2013, 33(7): 711-715 (in Chinese)

[8] EFRON B. Bootstrap methods[J]. The Annals of Statistics,1979,7(1):1-26

[9] 唐雪梅.小样本场合下相容性检验方法[J].系统工程与电子技术,2001,23(10):66-68  
TANG Xuemei. Consistency test methods in small sample situation[J]. Systems Engineering and Electronics, 2001, 23(10): 66-68 (in Chinese)

[10] 徐颖强,陈仙亮,曹栋波. 样本量为 2 的极小样本相容性检验方法[J]. 航空学报, 2018, 39(5): 221936  
XU Yingqiang, CHEN Xianliang, CAO Dongbo. Compatibility test method in minimal samples situation with two samples[J].  
Acta Aeronauticae Astronautica Sinica, 2018, 39(5): 221936 (in Chinese)

[11] 杨振海,程维虎,张军舰. 拟合优度检验[M]. 北京: 科学出版社, 2011  
YANG Zhenhai, CHENG Weihu, ZHANG Junjian. Goodness of fit test[M]. Beijing: Science Press, 2011 (in Chinese)

[12] 唐雪梅,张金槐,邵凤昌,等. 武器装备小子样试验分析与评估[M]. 北京: 国防工业出版社, 2001  
TANG Xuemei, ZHANG Jinhuai, SHAO Fengchang, et al. Test analysis and evaluation of weapon systems in small-sample circumstance[M]. Beijing: National Defense Industry Press, 2001 (in Chinese)

[13] 邱荣剑,张永录. 驾束制导导弹仿真系统的可信度评估[J]. 装备制造技术, 2011(2): 27-28  
QIU Rongjian, ZHANG Yonglu. Trustworthiness assessment of beam ride missile simulation system[J]. Equipment Manufacturing Technology, 2011(2): 27-28 (in Chinese)

## Research on small sample probability relational degree model

NING Xiaolei<sup>1</sup>, WU Yingxia<sup>1</sup>, ZHAO Xin<sup>1</sup>, ZHAO Junmin<sup>2</sup>, ZHANG Kai<sup>1</sup>,  
HAO Tiaofeng<sup>1</sup>, LYU Meibo<sup>3</sup>, HU Yafeng<sup>1</sup>, CHEN Yun<sup>2</sup>

(1.China Huayin Ordnance Test Center, Huayin 714200, China;  
2.Xi'an Modern Control Technology Research Institute, Xi'an 710065, China;  
3.School of Astronautics, Northwestern Polytechnical University, Xi'an 710072, China)

**Abstract:** The probability degree model is suitable for correlation analysis between multivariate time series, but its application effect is poor under the small sample condition. An improved probability degree model for small sample problem is proposed in this paper. Firstly, by using sample function or sample statistics instead of samples to construct cumulative empirical distribution function and calculate probability correlation coefficient, the framework of probability correlation degree model is developed and expanded. Secondly, the calculation method of correlation coefficient in the original model of probability correlation degree is improved. For the problem of sample size  $n < 5$ , the empirical distribution function of sample order statistics is used to calculate the probability correlation coefficient; For the problem of comparison sequence sample size  $n \geq 5$ , the Bootstrap method is used to resample to construct the empirical distribution function to calculate the probability correlation coefficient. Thus, the problem of dynamic multivariate data association analysis under the condition of small samples is solved. It is proved that the small sample probability correlation degree model satisfies the correlation theorem and its basic properties. The basic steps of small sample probabilistic correlation analysis are given. At last simulation cases and practical application verify the rationality and effectiveness of the model and method.

**Keywords:** data association; probability correlation degree; small samples; multivariate time series; sample statistics

**引用格式:** 宁小磊, 吴颖霞, 赵新, 等. 小样本概率关联度模型研究[J]. 西北工业大学学报, 2022, 40(5): 1164-1171  
NING Xiaolei, WU Yingxia, ZHAO Xin, et al. Research on small sample probability relational degree model[J]. Journal of Northwestern Polytechnical University, 2022, 40(5): 1164-1171 (in Chinese)