基于最大信息系数的时延数据相关性分析方法

王 鹏[1,2] 张善从[1,2]
(1. 中国科学院空间应用工程与技术中心 北京 100094;2. 北京国科环宇空间技术有限公司 北京 100190)

摘 要: 针对无法有效检测两组时延数据间相关关系的情况,提出以最大信息系数(*MIC*)为基础的平移搜索法。根据实际应用场景,设置合适的平移搜索窗和平移步长,由搜索窗内取得最大 *MIC* 值的位置求得时延估计值。将此方法分别应用到航天器载荷安装表面温度之间的相关性分析和狭义货币供应量(M1)与居民消费价格指数(CPI)的相关性分析中,结果表明针对两组时域上不对应的相关数据,利用此方法可以有效地检测出它们的相关性和时延。

关键词: 最大信息系数;时延估计;相关关系;航天器载荷;狭义货币供应量
中图分类号: TP274.2 **文献标识码**: A **国家标准学科分类代码**: 520.60

Method for the correlation analysis of data with time delay based on maximal information coefficient

Wang Peng[1,2] Zhang Shancong[1,2]
(1. Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China;
2. Beijing UCAS Space Technology Co., Ltd, Beijing 100190, China)

Abstract: Aiming at the problem that the correlation of two sets of data with time delay can't be detected effectively, a parallel moving search method based on Maximal Information Coefficient(*MIC*) is proposed. According to the practical application of the scene, set the appropriate parallel moving search window and step, obtain the time delay estimation value by the position with the maximum *MIC* value. Apply these methods on the correlation analysis of the equipment installation surface temperature data in aerospace and M1-CPI data in economics, results show that if two sets of data have correlation but with a time delay in time domain, using this method can detect the correlation and time delay effectively.

Keywords: Maximal Information Coefficient; time delay estimation; correlation; spacecraft equipment; narrowly defined money supply

1 引 言

在信息爆炸的当今社会,在大数据集中识别两个变量间的相关关系越来越重要[1]。利用 Pearson 相关系数或者 Spearman 相关系数可以有效的度量数据的线性相关性[2]。然而由于自然规律的复杂性,现实世界中数据之间存在许多非线性相关关系而且无法用简单的数学公式表达。为了度量数据间非线性相关性的强弱,基于阀值相关[3]、互信息[4]、相位同步[5]等度量方法先后被提出。

最大信息系数(the maximal information coefficient, *MIC*)是在互信息的基础上发展起来的,*MIC* 方法能快速通过给不同类型的关联进行评估,从而发现广泛范围的关系类型[6]。然而,如果两个相关变量间存在一定的时间延迟,如不同部位的温度数据或者经济学中的 M1 与 CPI 等,*MIC* 方法并不能很好的对这种相关度进行评估。针对这种问题,可以对数据进行平移,同时计算并记录每个平移位置的 *MIC* 值。当 *MIC* 达到最大时即可表征两个变量间真实的相关关系,通过计算平移步数与步长的乘积,可以得到两变量间的时延估计值。

2 最大信息系数

2.1 最大信息系数的原理概述

在有足够的样本量时,*MIC* 的公平性和通用性可以保证它能够捕获各种各样的有趣的关联[7],而不限于特定的函数类型。也就是说如果 X 与 Y 是相关的,它们必定存在这种关系: $Y=f(X)$,并不关注 f 的具体表达式是什么(事

收稿日期:2015-02

实上在大多数情况下,真正求出复杂关系的函数公式是相当困难的),只要 X 与 Y 存在 f 这种映射关系,都有 $MIC(X,Y)=1$。MIC 是基于互信息的,实际上 $MIC(X,Y)$ 就是 Y 中能被 X 解释的信息量的百分比[8]。在理想化的情况下,如果两个变量相互独立,则它们的 MIC 应该为 0。

直观地说,MIC 是基于这样的思想:如果两个变量之间存在着一种关系,则可以在两个变量的散点图中绘制网格,这些网格可以将散点图中的数据分割,这样有些网格是空的有些则含有散点图中的点。逐步增大网格的分辨率(比如由 2 乘 2 增大到 x 乘 y),通过网格中的点数可以计算每种分辨率下可能产生的最大互信息值,然后标准化这些互信息值,以确保不同分辨率的网格之间进行公平的比较。定义矩阵 $\boldsymbol{M}=(m_{x,y})$,其中 $m_{x,y}$ 是在每种分辨率下计算得到的最大互信息标准化值,而 MIC 就是 $\boldsymbol{M}$ 中的最大值。

更正式的表达是,点对集合 $D=\{(a_1,b_1),\cdots,(a_n,b_n)\}$ 共有 n 个点,以 a 为横坐标,b 为纵坐标画散点图。令 $G_{x,y}$ 表示以 x 乘 y 分辨率分割的网格。令 $p_0,p_1,\cdots,p_x$ 表示横坐标轴上的分割点(其中 $p_0=a_1,p_x=a_n$);同理,令 $q_0,q_1\cdots q_y$ 表示纵坐标轴上的分割点(其中 $q_0=b_1,q_y=b_n$)。改变 $p_1\cdots p_{x-1}$ 和 $q_1\cdots q_{y-1}$ 的值可以得到 x 乘 y 分辨率下不同的分割方式,令 $\max\{\boldsymbol{I}_{x,y}\}$ 表示 $G_{x,y}$ 中不同分割方式散点分布产生的最大互信息值,则此分辨率分割计算得到的特征矩阵值等于[9]:

$$m_{x,y}=\frac{\max\{\boldsymbol{I}_{x,y}\}}{\log(\min\{x,y\})} \tag{1}$$

此矩阵值均落在[0,1]的范围内。$m_{\max}$即代表这两个变量的 MIC,其中 $xy<n^{0.6}$。

2.2　互信息的计算方法

为了计算 M,理论上应该对所有可能的分辨率网格下的所有分割方式进行计算,但是当 n 很大时,计算量会非常大。有一种近似计算的方式,即将每种分辨率下纵坐标轴均进行等距离分割,计算出 MIC 后将 a 和 b 交换坐标轴重新计算 MIC,取两次计算的最大值为最终的 MIC 值。

令 P 为 x 分辨率下的横坐标轴的某种分割方式,令 Q 为 y 分辨率下的纵坐标轴的等距离分割方式。令 $E_{i,*}$ 的值为散点图落在 P 中第 i 列中的点数,$\#_{*,j}$的值为散点图落在 Q 中第 j 行中的点数,$\#_{i,j}$的值为散点图落在第 i 列第 j 行的格子中的点数。于是有:

$$H(P)=\sum_{i=1}^{x}\frac{\#i*}{n}\log\frac{n}{\#i*} \tag{2}$$

$$H(Q)=\sum_{j=1}^{y}\frac{\#*j}{n}\log\frac{n}{\#*j} \tag{3}$$

$$H(P;Q)=\sum_{i=1}^{x}\sum_{j=1}^{y}\frac{\#ij}{n}\log\frac{n}{\#i*j} \tag{4}$$

而互信息 $I(P;Q)=H(P)+H(Q)-H(P;Q)$。

2.3　最大信息系数的算法

在明确了网格分割方式和互信息的计算方法后,求矩阵 $\boldsymbol{M}$ 的核心算法如表 1 所示。

表 1　求矩阵 $\boldsymbol{M}$ 算法伪代码

Algorithm 矩阵 $\boldsymbol{M}(D)_{x,y}$的算法
输入:集合 $D=\{(a_1,b_1),\cdots,(a_n,b_n)\}$,整数 B(大于 3)
输出:特征矩阵 $\boldsymbol{M}_{x,y}$
1. 令 $D^{\perp}\leftarrow D$ 中 a 与 b 的位置互换
2. 对所有的 $y\in\{2,\cdots,[B/2]\}$,令 $x\leftarrow[B/y]$
3. 令 a 为横坐标,b 为纵坐标,等划分纵坐标轴
4. 按 a 值升序排列 D 中点对,优化横坐标轴分割,在每种分割下得到最大互信息:$\{\boldsymbol{I}_{2,y},\cdots,\boldsymbol{I}_{x,y}\}$
5. 对 $D^{\perp}$重复以上过程得到 $I^{\perp}$,取 I 和 $I^{\perp}$中较大的值
6. 标准化 $\boldsymbol{I}_{x,y}$,令 $\boldsymbol{M}_{x,y}\leftarrow\boldsymbol{I}_{xy}/\min\{\log x,\log y\}$

3　平移搜索法原理

3.1　应用场景

现实生活中许多变量之间的相互影响并不是立即发生的,因变量和自变量之间的关系可能会存在一定的滞后性,如被动探测系统多个接收点接收到的目标信号[10]或者钢筋两端震动幅度等。

假设存在两个相关变量 A、B,A 的变化会影响 B 的变化且不一定是线性影响,同时这个影响具有滞后性。由对 MIC 的阐述可知,无论是对线性还是非线性相关关系 MIC 都是非常有效的。然而 MIC 只能计算时域上实时对应的数据,如果 AB 之间的相关关系具有时间延迟,MIC 的计算结果就不再确定和有效。利用基于 MIC 的平移搜索法可以有效的捕获到此种类型的相关关系。

3.2　计算方法

针对数据的特点和情景,可以设置合适的平移搜索窗和平移步长对 A 或 B 进行向前或者向后平移,每平移一次计算一次 MIC。在搜索窗的大小内计算得到最大的 MIC 记为 MIC_T,取得 MIC_T 的步数乘以步长即为 A 与 B 的时间延迟。

对于数据量不大和延迟较小的数据,此方法既简单又快捷;然而在两个变量的数据量 n 较大,同时它们之间的延迟也较大的情况下,计算量会非常大。通过第一节 MIC 的计算过程可知,利用并行计算将会是一个不错的选择。同时,根据数据的周期性和应用情景可以预测并选择合适的平移窗,加大平移步长来减小计算量和进行时延估计。

4　实际应用

本文以航天领域中的某航天器载荷安装支架表面温度数据和经济领域中的经济指数数据作为应用对象,有力地证明了平移搜索法的有效性。

4.1　某航天器载荷安装表面温度之间的相关性分析

随着航天技术的发展,航天器上配置了越来越多的载荷,空间应用任务越来越复杂,载荷系统不同模块之间的潜在影响关系也越来越难以明确。为了能够确定同一载荷或者不同载荷的电气、温度等要素之间的影响关系,对大量的

载荷遥测数据进行分析是十分必要的。掌握了这些关系，在航天器系统测试或者实时监控中如果发现设备故障，就可以通过关联列表将人的兴趣度缩小到一个较小的范围内，迅速定位可能引起故障的因素从而大大减少排故时间。此外，还可以通过数据间的相关关系变化来进行故障诊断。在航天器的实际运行中，每天都有大量的遥测数据下传，通过传统的人工分析的方式已经不再适用。

在航天器中，不同的载荷因为安装位置、功能、导热性等因素必定会有温度上的相关关系，同时这种关系必定具有时间延迟，以“器件 A 安装支架表面测温”和“器件 B 安装支架表面测温”遥测数据为例，经过移动平滑[11]去噪处理后两个设备的温度曲线如图 1 所示。

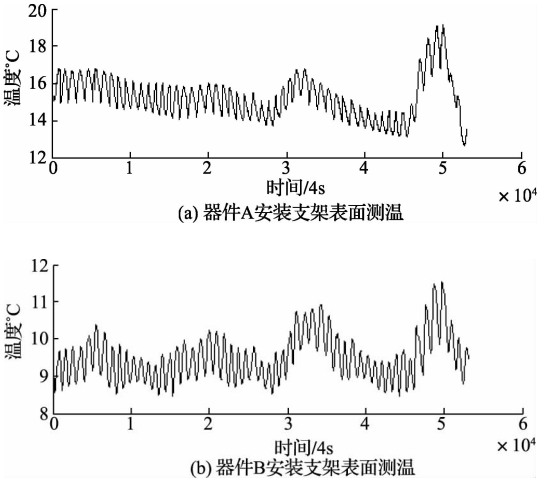

图 1　两个设备安装面的温度曲线

可以看出两个设备安装面的温度曲线非常相似，二者必定存在某种相关关系，并且可以看出器件 B 安装支架表面的温度变化要比器件 A 安装支架表面的温度超前。可利用平移搜索法寻找 *MIC* 最大时的平移量。根据电气特性与温度传导特性，将平移搜索窗设置为 10 min，步长为 4 s。

计算过程如下：

1)平移前两组数据的 *MIC*：0.5900；

2)将时间“落后”的数据逐步向前平移，每平移一次计算一次 *MIC* 的大小，在 *MIC* 达到最大值时，平移的大小即为延迟的大小；

3)平移 46 个点时得到 *MIC* 最大值 0.7650。每两个点的间隔为 4 s，所以可得二者的时延为 4s×46，也就是大约 3 min。

通过实验结果可知，器件 B 安装支架表面对器件 A 安装支架表面的温度有一定的影响或者二者受同一因素的影响但是器件 B 安装支架表面因为安装位置或者热传导性等因素先受到影响。

4.2　狭义货币供应量(*M*1)与居民消费价格指数(CPI)的相关性分析

图 2 所示的是 1999 年 12 月到 2013 年 12 月的 M1 和 CPI 月度同比增长率曲线。直接计算 M1 和 CPI 的 *MIC* 为 0.244，从这个结果来看似乎二者的相关关系是较弱的。然而实际上，由于货币供给波动是常态，本期货币波动效果在第 6 个月的时候会表现出来[12]。M1 与 CPI 的关系是非常密切的，M1 对 CPI 的影响具有滞后性。

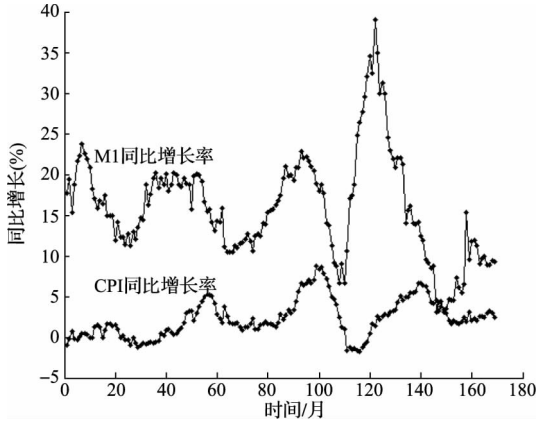

图 2　M1 与 CPI

利用平移搜索法估计二者的时间延迟。设置平移搜索窗为 12 个月、步长为 1 个月逐步向前平移 CPI 数据的过程中计算搜索得到一年内 *MIC* 最大值 0.501 8，发生在平移第六步时。计算结果如表 2 所示。

表 2　12 个月内 *MIC* 平移计算结果

步数	*MIC*
0	0.243 903
1	0.277 309
2	0.297 428
3	0.364 842
4	0.379 229
5	0.398 324
6	0.501 778
7	0.39 717
8	0.431 434
9	0.417 139
10	0.431 389
11	0.425 004
12	0.394 765

由以上结果可得知 M1 对 CPI 是有一定影响的，并且影响的滞后性为 6 个月。这个结果与詹帆[12]的研究结论是一致的。由 *MIC* 的定义可知，*MIC* 越接近 1 说明两个变量的相关性越强。M1 与 CPI 的最大 *MIC* 为 0.5018，只能算作中等强度相关。这是因为 CPI 除了受 M1 的影响外还受到其它繁多的复杂因素影响，如 PPI、人民币外汇占款、全国农产品批发价格指数、煤电油价格指数等[13]。

5 结　论

本文针对时延数据间相关关系的检测问题提出了基于 *MIC* 的平移搜索法。在航天和经济两个完全不同领域中的两类数据上进行了验证,实验结果令人满意,证明此方法也可以应用到其他各个学科的时延数据相关性的检测中。由 *MIC* 的计算过程和特性可以看出,计算不同分辨率下不同分割方式的最大互信息之间是互不影响的,从而可知这个算法是非常适合做并行化的。如果能利用云计算或者 GPU 实现 *MIC* 算法,提高几个数量级的计算效率,同时根据应用场景设置合适的搜索窗和平移步长,即使在大数据量、多变量的情况下也可以检测出有效的时延和非时延的相关关系。

参考文献

[1] DAVID N R, YAKIR A R, HILARY K F. Detecting novel associations in large datasets[J]. Science, 2011, 334(6062): 1518-1524.

[2] 魏瑜,马开平. 相关分析的误用表现与解决方案[J]. 统计与决策,2015(2):86-88.

[3] MOKHTAR B A. On a robust correlation coefficient [J]. The Statistician, 1990, 39(4): 455-460.

[4] 詹曙,李敏,徐甲甲,等. 局域化互信息度量的 ACM 下医学图像的分割[J]. 电子测量与仪器学报,2013, 27(4):340-346.

[5] ERNESTO P, RODRIGO Q Q, JOYDEEP B. Nonlinear multivariate analysis of neurophysiological signals[J]. Progress in Neurobiology, 2005, 77(1): 1-37.

[6] 战泉茹. 基于最大信息系数的人脸特征选择[D]. 长春:东北师范大学,2013.

[7] 高凤. 互信息的最大信息系数法在股市关联度上的应用[J]. 新西部,2014(29):56-59.

[8] 蒋杭进. 最大信息系数及其在脑网络分析中的应用[D]. 武汉:中国科学院大学,2013:17-23.

[9] 魏中强,徐宏喆,李文,等. 基于最大信息系数的贝叶斯网络结构学习算法[J]. 计算机应用研究,2014, 31(11):3261-3265.

[10] 王锋,刘美全,孙钦蕾. 被动探测中相关时延估计研究[J]. 国外电子测量技术,2014,33(2):61-64.

[11] 唐玉发,张合,徐国泰,等. 基于加权移动平均的姿态角测量技术实现[J]. 仪器仪表学报,2012,33(8): 24-28.

[12] 詹帆. 浅析 M1,M2 与 CPI 的时滞关系[J]. 江苏科技信息,2013(6):24-25.

[13] 张海涛. 影响我国当前 CPI 的因素分析及动态预测[J]. 北方经济,2012(9):10-11.

作者简介

王鹏(通讯作者),硕士研究生,现就读于中国科学院空间应用工程与技术中心。主要研究方向为空间信息传输与处理。

张善从,博士,中国科学院空间应用工程与技术中心研究员,中国载人航天工程应用系统副总设计师。曾参与了载人航天应用系统、空间硬 X 射线天文卫星和对地观测小卫星平台等系统的设计与研究工作。荣获国家级科技进步二等奖一项,省部级科学技术二等奖一项。目前所承担的课题任务主要包括星载计算机、操作系统、应用软件和相关地面仿真测试系统等。