arXiv:2301.06037v2 [cs.LG] 5 Feb 2023

基于 Copula熵传递熵识别动力系统中的时间滞后

Jian MA*

日立中国研究实验室

摘要

变量之间的时间滞后是不同领域动力系统的一项关键特征，而识别此类时间滞后是复杂系统中的一个重要问题，具有广泛的应用。近年来，传递熵（TE）被提出作为时滞识别的工具。然而，估计传递熵一直是一个众所周知的难题。Copula熵（CE）是衡量统计独立性的指标，且已有证明表明，传递熵可以仅用Copula熵表示。因此，近期有研究根据这一表示提出了一种基于Copula熵的传递熵非参数估计器。本文提出使用这种基于Copula熵的传递熵估计器来识别动力系统中的时间滞后。实验采用模拟数据和真实数据验证所提方法的有效性。实验结果表明，该方法能够识别四个模拟系统中的时间滞后。基于得土安城电力消耗数据的真实数据实验也表明，我们的方法能够通过天气因素到城市电力消耗的估计传递熵，识别时间滞后的模式。

关键词： 时间滞后；传递熵；Copula熵；系统辨识

1 引言

世界的运行伴随着时间滞后。在各种物理、社会或生物系统中，由于质量、能量或信息的传递，要观察到一个因果因素对其他因素的影响需要时间。系统变量之间的时间滞后是不同领域动力系统的关键特征之一。在复杂系统中，识别此类时间滞后是一项重要问题 [1]，有助于我们理解底层系统中的因果关系。

时滞识别的一种传统直观方法，是基于时间变量 [2, 3, 4] 的时间序列之间的时间相关性。然而，相关系数只能在高斯假设下使用，这对于具有非高斯性的非线性系统而言并不现实。时间相关性的替代方法是时延互信息（MI），它能够处理

*电子邮箱：majian@hitachi.cn

时间非线性相关性 [5, 6, 7, 8]。互信息（MI）是信息论中的一个概念，可用于衡量随机变量之间的非线性统计依赖关系，也可视为相关性的非线性推广 [9]。然而，相关性和互信息都是对称的，旨在衡量无方向的关系，因此不适合识别动力系统中的时间滞后，因为与时间滞后相关的因果关系具有方向性。

传递熵（TE）是另一种信息论概念，用于衡量时间变量之间的因果关系 [10]。它被定义为两个条件熵之间的 Kullback-Leibler散度，其中一个以因果因素为条件，另一个则不以其为条件。从本质上说，TE 是条件互信息，可视为格兰杰因果关系（GC）的非线性推广 [11, 12]。它能够测量动力系统中的方向性信息，近期也被提出作为时滞识别的工具 [13, 14, 15, 16]。

传递熵估计对于其实际应用至关重要，曾一度被认为是一个极其困难的问题。估计传递熵的一种基本方法是依据传递熵的定义，例如 [17, 18]。这类估计量通常会产生不一致且有偏的估计结果，因为这些估计量中从有限数据估计密度这一关键步骤并不稳定。

Copula熵（CE）是一种用于衡量统计独立性的数学概念，且已被证明等价于互信息 [19]。Ma 证明，传递熵可以仅用 Copula熵表示，并据此提出了一种传递熵的非参数估计器 [20]。由于该估计器经过严谨且可靠的证明，其性能一致且可靠。它已应用于许多实际案例，例如故障诊断 [21], 情感分析 [22], 电力分配分析 [23]。

本文提出采用基于 CE 的方法估计传递熵，并据此识别动力系统的时间滞后。由于基于 CE 的传递熵估计量是非参数的，因此无需任何假设即可应用于任意系统。我们将通过仿真数据和真实数据验证所提方法的有效性。在仿真实验中，我们将从具有不同时间滞后的四种不同随机动力系统中模拟生成一组时间序列数据。我们还将在得土安的能源消耗预测实验中使用真实数据，以展示本方法识别出的天气因素与能量之间的时间滞后。

本文的组织结构如下：第2节介绍相关工作；第3节阐述 CE 理论，以及估计 CE 和通过 CE 估计传递熵的方法；第4节提出基于 CE 的传递熵识别时间滞后的方法；第5节和第6节分别介绍使用仿真数据和真实数据开展的实验；第7节总结全文。

2 相关工作

近期有人提出将传递熵作为识别不同系统中时间滞后的工具。Lahiri等人 [13] 利用基于传递熵的时滞识别研究了细胞过程之间的方向性影响。该研究中的传递熵通过解析方法和数值方法推导得出。Rosinbery等人 [14] 将传递熵应用于具有延迟耦合的平稳线性随机过程，其中

采用 Wiener-Hoef因子分解计算关于时间滞后的传递熵。
Amornbunchornvej等人 [24] 提出了一种利用格兰杰因果关系和传递熵推断时变延迟的方法。他们通过实际应用展示了该方法。

传递熵还有若干相关的实际应用。利用传递熵进行时滞识别已应用于交通分析，以回答有关交通拥堵成因及传播的基本问题 [16], ，其中采用有效传递熵方法结合分箱来估计传递熵。Marzbanrad等人 [15] 使用传递熵研究不同时间滞后下母体和胎儿心率之间的生物相互作用，并通过基于分箱的定义计算传递熵。他们研究了母亲与胎儿之间信息流的滞后。Wing等人[25]运用包括互信息和传递熵在内的信息论工具，研究太阳风与地球辐射带耦合。他们发现，传递熵可用于检测这种耦合的变化，且太阳风到地球同步轨道兆电子伏特电子通量峰值的时间滞后约为2天。在他们的研究中，传递熵也通过分箱技术进行估计，而分箱数量需要谨慎选择。

此前也有研究根据传递熵的原始定义进行估计。Faes等人 [17]提出，将传递熵估计为其定义中两个条件熵之差。Vicente等人 [18] 提出，将传递熵估计为由传递熵原始定义展开得到的四个熵之和。Staniek和Lehnertz [26] 建议采用一种符号化技术来估计传递熵：先将原始变量重新排列为符号，再使用基于符号的传递熵定义作为估计结果。

Copula也被用于估计动力系统中的时延。Wang等人 [27] 提出使用基于Copula的依赖度量来识别线性系统的时延。他们论文中使用的度量定义为经验Copula与乘积Copula之间的距离。

3 方法

3.1 CE理论

Copula理论统一了多元依赖关系的表示 [28, 29]。根据Sklar定理 [30], ，多元联合密度函数可以表示为其边缘分布与Copula密度的乘积，其中Copula密度刻画了随机变量之间的依赖结构。符号说明请参阅 [19] 。

利用Copula密度，Ma和Sun [19] 定义了一种新的数学概念，称为Copula熵，具体如下：

定义1 (Copula熵)。 设$\mathbf{X}$ 为边缘分布为$\mathbf{u}$且*Copula*密度为 $c(\mathbf{u})$的随机变量。$\mathbf{X}$ 的*CE*定义为

$$H_c(\mathbf{x}) = -\int_u c(\mathbf{u}) \log c(\mathbf{u}) d\mathbf{u}. \tag{1}$$

在信息论中，互信息是一个不同于熵的基本概念[9]。Ma和Sun [19]证明，互信息本质上等于负的Copula熵，如下所示：

定理1。 随机变量的互信息等价于负的*Copula*熵：

$$I(\mathbf{x}) = -H_c(\mathbf{x}), \tag{2}$$

其中 I 表示互信息。

定理1有一个简洁的证明 [19]，并立即推得一个推论，说明联合密度函数、边缘分布与Copula密度所包含的信息之间的关系。

推论1。 随机变量的联合熵等于各变量的熵之和与随机变量的*Copula*熵之和：

$$H(\mathbf{x}) = \sum_i H_i(x_i) + H_c(\mathbf{x}), \tag{3}$$

其中 H 表示熵。

上述言简意赅的结果揭示了互信息与Copula之间的关系，因此在信息论与Copula理论之间架起了一座桥梁。

3.2 估计Copula熵

互信息是信息论中的一个基本概念，在各个领域都有广泛应用。然而，互信息的估计一直颇具挑战性。得益于定理1，Ma和Sun [19] 提出了一种简洁优雅的非参数方法，用于从数据中估计Copula熵（互信息）；该方法分为两个步骤：

1. 估计经验Copula密度（ECD）；
2. 根据估计得到的ECD估计Copula熵。

给定由随机变量生成的样本 $\{X_1, \ldots, X_T\}$ i.i.d. $\mathbf{X}$ $=\{X_1, \ldots, X_N\}$，可以利用经验函数轻松推导出经验Copula密度：

$$F_i(x_i) = \frac{1}{T} \sum_{t=1}^{T} I(X_t^i \leq x_t^i), \tag{4}$$

其中 $i = 1, \ldots, N$和 I表示示性函数。估计出经验Copula密度后，Copula熵的估计就转化为熵估计问题，可以通过许多现有方法解决。文献 [19]提出了使用k近邻法 [31]进行此类熵估计。由此可得一种估计Copula熵的非参数方法。

3.3 通过Copula熵估计传递熵

传递熵是一种信息论概念，用于衡量时间序列之间的因果关系。它基于Wiener原理定义，该原理指出，如果随机变量 X 能够改善对 Y 的预测，那么 X 就是 Y的原因。格兰杰因果关系是基于这一原理、采用线性向量自回归模型定义的一种度量。传递熵可视为格兰杰因果关系的非线性推广。传递熵是一种无模型度量，不作任何假设。

给定两个时间序列 X_i，Y_i，$i = 1$，.....，T，从 X到 Y的传递熵定义为

$$T_{X \to Y} = \sum p(y_{i+1}, y^i, x_i) \log \frac{p(y_{i+1}|y^i, x_i)}{p(y_{i+1}|y^i)}, \tag{5}$$

其中 p 表示概率密度函数， $y^i =$ （y_1，……， y_i）。

Ma [20] 证明，传递熵（TE）仅用Copula熵（CE）即可表示，如下所示：

$$T_{X\to Y} = -H_c(y_{i+1}, y^i, x_i) + H_c(y_{i+1}, y^i) + H_c(y_{i+1}, x_i) - H_c(y^i). \tag{6}$$

基于这一表示，Ma [20] 提出了一种传递熵（TE）的非参数估计器，由两个简单步骤组成：

1. 使用非参数Copula熵（CE）估计器估计式（6）中的Copula熵项；
2. 根据估计得到的Copula熵项计算传递熵（TE）。

4 提出的方法

本节提出使用我们的传递熵（TE）估计器识别动力系统中的时间滞后。我们将利用传递熵，根据变量的时间序列观测量化系统中变量之间的因果关系。首先，通过改变两个变量之间的时间滞后，利用时间序列数据估计变量间的一组传递熵；然后，将传递熵值最大的滞后识别为动力系统中变量之间的滞后。传递熵将使用第3.3节中的非参数估计器进行估计。该估计器无模型，因此不对底层系统作任何假设。识别出的时间滞后将被视为系统动力学的特征，而相应的传递熵值将被视为两个变量之间的因果强度。

5 仿真

5.1 实验

我们进行了仿真实验，以验证所提方法的有效性。在仿真中，我们设计了四个动力系统，用于模拟具有时间滞后的不同动态过程。这些系统均考虑了状态滞后和输出滞后。

系统 1 第一个系统的状态变量服从高斯分布，输出变量则由状态变量按照预设时间滞后平移得到。该系统的数学模型如下：

$$\begin{aligned} X_i &= \xi_1, \\ Y_{i+l} &= X_i + \xi_2, \\ \xi_1 &\sim N(\mu_1, \delta_1), \\ \xi_2 &\sim N(\mu_2, \delta_2), \end{aligned} \tag{7}$$

其中， l 表示从状态变量 X 到输出变量 Y的时间滞后， $N(\mu_i, \delta_i)$表示均值为 μ_i、方差为 δ_i的正态分布。

系统 2 第二个系统与第一个系统相似，但状态变量中增加了一个正弦动力学项。第二个系统的数学模型如下：

$$\begin{aligned} X_i &= \sin(2\pi i/m) + \xi_1, \\ Y_{i+l} &= X_i + \xi_2, \\ \xi_1 &\sim N(\mu_1, \delta_1), \\ \xi_2 &\sim N(\mu_2, \delta_2), \end{aligned} \tag{8}$$

其中，l 表示状态变量 X 到输出变量 Y的时间滞后，$N(\mu_i, \delta_i)$表示均值为 μ_i、方差为 δ_i的正态分布，m 表示模拟时间序列的长度。由于正弦项的存在，第二个系统的动力学特性与第一个系统不同。

系统 3 第三个系统是一阶系统，包含一个状态变量和一个关于时间滞后的输出变量。第三个系统的数学模型如下：

$$\begin{aligned} X_i &= X_{i-1} + \xi_1, \\ Y_{i+l} &= X_i + \xi_2, \\ \xi_1 &\sim N(\mu_1, \delta_1), \\ \xi_2 &\sim N(\mu_2, \delta_2), \end{aligned} \tag{9}$$

其中，l 表示状态变量 X 到输出变量 Y的时间滞后，$N(\mu_i, \delta_i)$表示均值为 μ_i、方差为 δ_i的正态分布。这实际上是维纳过程，其中状态变量 X引入了自回归。

系统 4 在前述三个系统中，模拟的是输出延迟。这里的第四个系统是一阶延迟系统，用于模拟状态变量中的延迟。其数学模型如下：

$$\begin{aligned} X_i &= \alpha X_{i-1} + \beta X_{i-l} + \xi_1, \\ \xi_1 &\sim N(\mu_1, \delta_1), \end{aligned} \tag{10}$$

其中，l 表示系统状态的时间滞后，α，β 是系统的参数，ξ_1 是一个服从均值为 μ_1 、方差为 δ_1的正态分布的随机变量。在这种设定下，系统的当前状态不仅与上一状态有关，还与 l 个时间滞后之前的某个状态有关。

在这些仿真实验中，我们将时间滞后 l的值从 1 改为 4，为每个系统生成 4 条状态变量或输出变量的轨迹。我们将所提方法应用于生成的轨迹数据，以估计：1）前三个系统中从状态变量到输出变量的传递熵（TE）；2）第四个系统中状态变量内部与不同时间滞后对应的传递熵（TE），然后通过估计的传递熵值的最大值识别系统的时间滞后。我们估计了时间滞后从 1 到 8 时的传递熵，该范围大于模拟系统的时间滞后范围。在所有仿真中，$\mu_i = 0$、$\delta_i = 0.001$，且模拟时间序列数据的长度为 500。第四个系统中的参数 α，β 分别为 0.2 和 0.8。

5.2 结果

仿真实验结果如图1–5所示。第一个系统的一条模拟时间序列数据如图 1(a) 所示。该系统在四种时间滞后下，从状态变量到输出变量的估计传递熵值如图2所示。可以看出，在这四种情况下，传递熵值的最大值都对应于系统的时间滞后。

第二个系统的一条模拟时间序列数据如图 1(b) 所示，由此可以很容易看出，系统的输出按照预设时间滞后跟随系统的状态。图3展示了该系统在四个时间滞后下从状态变量到输出变量的估计传递熵值。从图中可以看出，在所有四种情况下，传递熵值的最大值都对应于系统的时间滞后。第一个系统与第二个系统的结果差异在于，随着时间滞后的变化，第二个系统的传递熵值不像第一个系统那样回归到 0。这是由于第二个系统的正弦项引入了动力学。

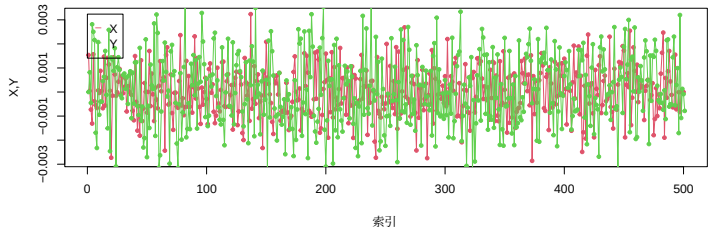
仿真实验3的结果如图 1(c) 和图4所示。图 1(c) 展示了模拟数据，反映了仿真系统的随机动力学。由于该系统是一阶系统，其动力学比前两个系统复杂得多。图4展示了该系统在四个时间滞后下从状态变量到输出变量的估计传递熵值。从图中可以看出，在所有四种情况下，传递熵值的最大值都能正确识别出这四个时间滞后。四种情况下的估计传递熵值均随着时间滞后的增加而逐渐增大至最大值，随后逐渐衰减。

仿真数据和仿真实验4的估计结果分别如图 1(d)和图5所示。从图5可以看出，我们的方法能够通过估计的传递熵值的最大值，正确识别模拟系统中的四个时间滞后。

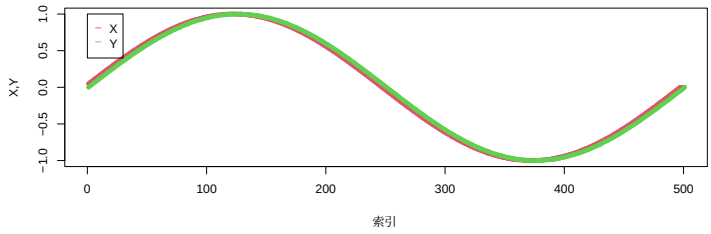
6 真实数据

我们还使用 UCI 机器学习资料库 [32]中得土安城的电力消耗真实数据验证了我们的方法。该数据集包含2017年[33]摩洛哥北部得土安城三个不同配电网络（夸兹、斯米尔和Boussafou）的电力消耗历史数据。该数据集旨在利用天气因素预测电力消耗。由于季节和社会因素，三个网络的数据有所不同。除了三个网络的电力消耗时间序列数据外，数据集还包含影响电力消耗的五种天气因素：温度、湿度、风速、一般漫射流和漫射流。数据的采样率为每10分钟一个样本。

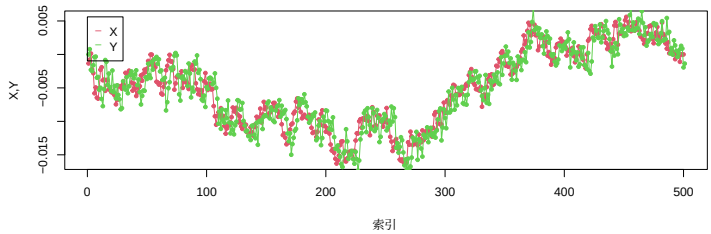
众所周知，天气变化会在一段时间后影响电力消耗。因此，我们希望从时间滞后的角度研究天气对电力消耗的影响。我们将把该方法应用于数据，以研究天气对电力消耗的时间滞后。在实验中，我们使用了五个工作日（2017年-10-23∼2017-10-27）的数据，因为工作日与节假日有所不同。我们估计了来自五个



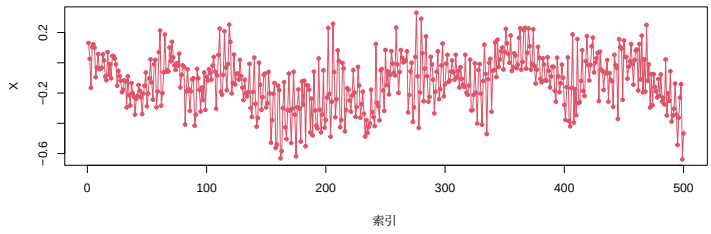
(a) 系统 1



(b) 系统 2



(c) 系统 3



(d) 系统 4

图 1：仿真实验中的模拟轨迹。

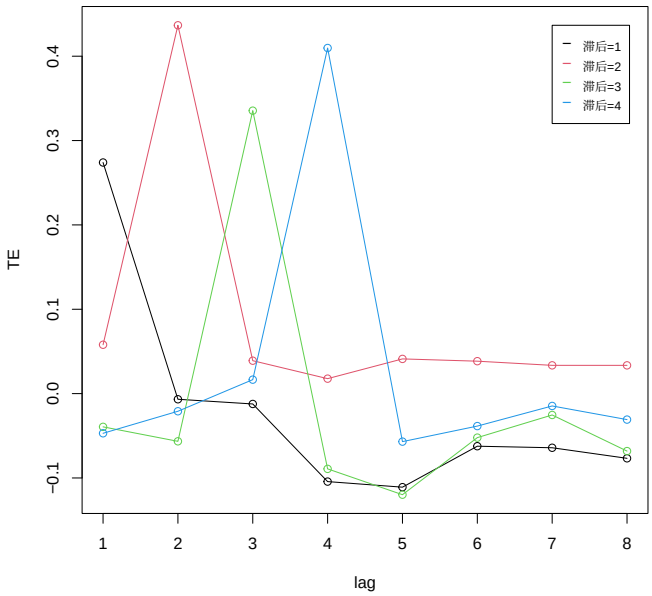


图2：模拟实验1的结果。

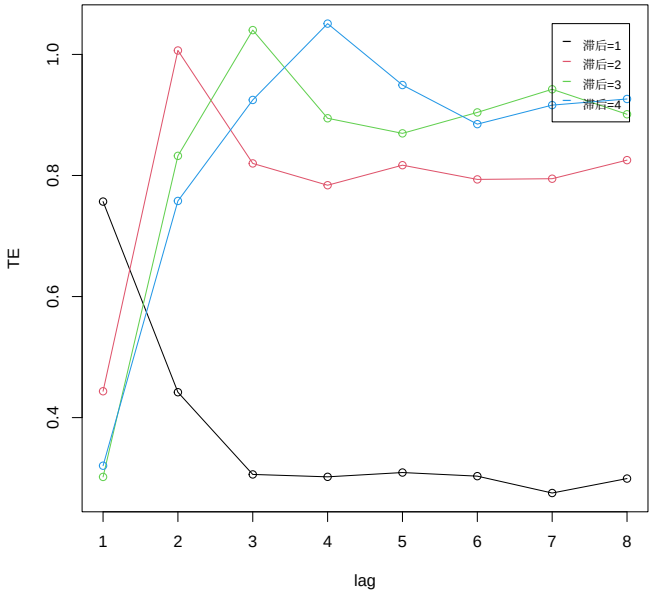


图3：模拟实验2的结果。

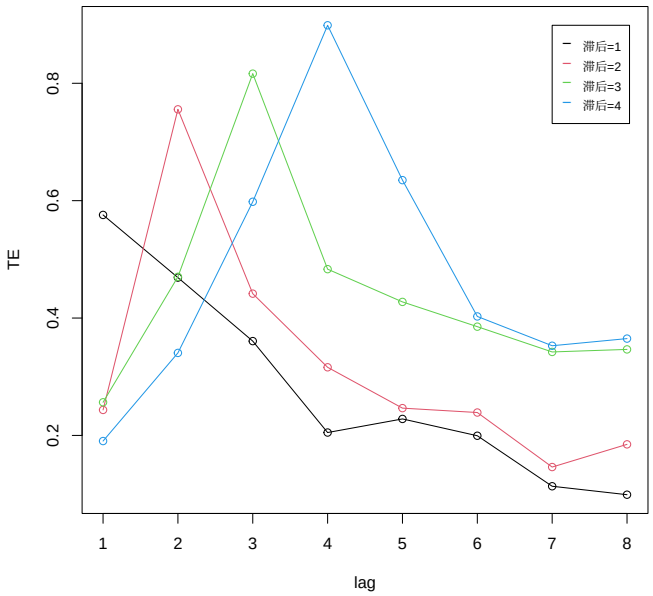


图4：仿真实验3的结果。

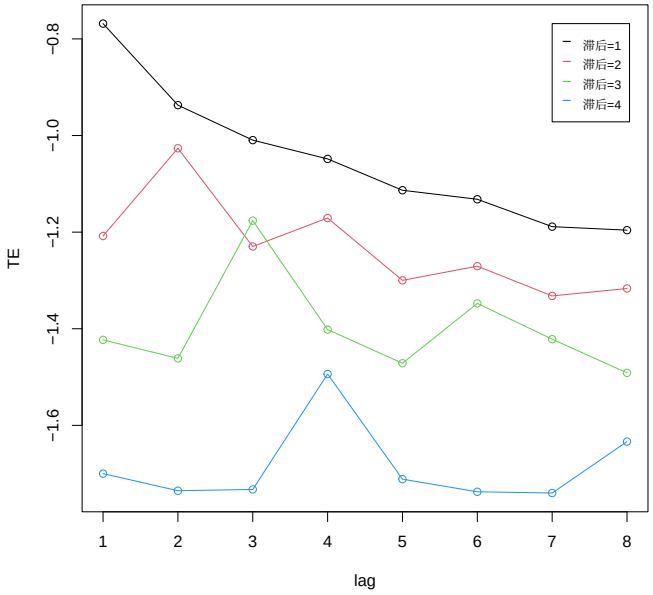


图5：仿真实验4的结果。

天气因素对三个网络电力消耗的影响。时间滞后范围为1至24小时，因为我们关注的是估计的传递熵（即天气影响电力消耗的因果关系）在整整一天中的变化。

实验的估计结果如图6所示，其中每行对应一个天气因素，每列对应一个网络。每幅图展示了传递熵随时间滞后从1小时变化至24小时的情况。从中可以看出：1）天气因素对三个网络电力消耗的几乎所有估计传递熵都存在时间滞后（约4小时），在此期间因果效应迅速增强；2）当因果效应达到最大值时，温度对电力消耗的时间滞后要长得多，变化也更缓慢；3）其他四种因素对三个网络电力消耗的估计传递熵可视为每半天有两个峰值，这意味着天气因素的日变化模式导致了两个时间滞后。

7 结论

本文提出使用基于CE的传递熵估计器识别动力系统中的时间滞后。我们通过模拟数据和真实数据，在实验中检验了该方法。我们设计了四种随机动力系统，以模拟具有时间滞后的时间序列数据。实验结果表明，该方法能够从模拟数据中识别这些系统中的时间滞后。我们使用得土安城的电力消耗真实数据，展示了该方法的有效性。估计结果表明，该方法能够识别天气因素对得土安城电力消耗影响的时间滞后模式。

A 代码

代码可在 https://github.com/majianthu/timelag 获取。

参考文献

[1] S. Bjorklund 和 L. Ljung。时延估计技术综述。载于第*42*届 *IEEE*决策与控制国际会议（*IEEE Cat. No.03CH37475*），第3卷，第2502–2507页 Vol.3，2003年12月。[2] C. Knapp 和 G. Carter。用于估计时延的广义相关法。*IEEE*声学、语音与信号处理汇刊，24(4):320–327，1976年8月。[3] Guojun Gu 和 Robert F. Adler。年际时间尺度上的降水和温度变化：评估厄尔尼诺—南方涛动和火山喷发的影响。《气候杂志》，24(9):2258–2270，2011年。[4] Daniel J Kiviet、Philippe Nghe、Noreen Walker、Sarah Boulineau、Vanda Sunderlikova 和 Sander J Tans。单细胞水平的代谢随机性与生长。《自然》，514(7522):376–379，2014年。

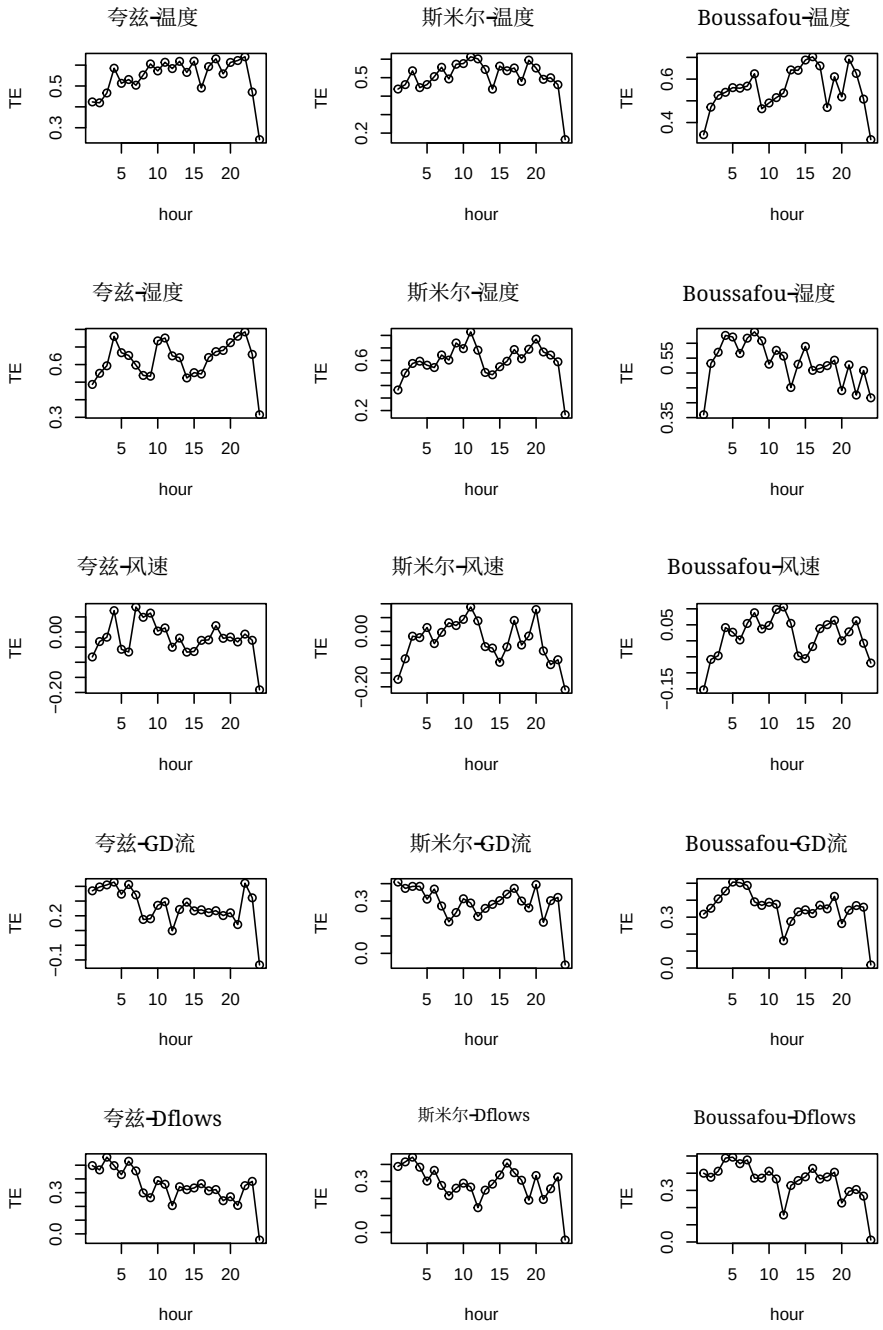


图6：使用得土安城真实数据进行实验的结果。

[5] N .J.I. Mars an d G.W. van Arragon. Time delay es timation in non- li
利用互信息平均量分析邻近系统。《Signal Processing》，4(2):139–153，1982年。[6] Fei Wen 和 Qun Wan。基于互信息估计的时延估计。载于《2009年第二届国际图像与信号处理大会》，第1–5页，2009年10月。[7] D. J. Albers 和 George Hripcsak。利用时延互信息发现并解释复杂群体中的时间相关结构。《Chaos: An Interdisciplinary Journal of Nonlinear Science》，22(1):013111，2012年。[8] Andreas Wilmer、Marc de Lussanet 和 Markus Lappe。将相位的时延互信息用作功能连接的测度。《PLoS ONE》，7(9):e44633，2012年。[9] Thomas M Cover 和 Joy A Thomas。《信息论基础》。John Wiley & Sons，2012年。[10] Thomas Schreiber。测量信息传递。《Physical Review Letters》，85:461–464，2000年7月。[11] Clive WJ Granger。通过计量经济模型和互谱方法研究因果关系。《Econometrica: Journal of the Econometric Society》，第424–438页，1969年。[12] Clive WJ Granger。因果关系检验：个人观点。《Journal of Economic Dynamics and Control》，2:329–352，1980年。[13] Sourabh Lahiri、Philippe Nghe、Sander J Tans、Martin Luc Rosinberg 和 David Lacoste。细胞过程中方向性影响的信息论分析。《PLoS ONE》，12(11):e0187431，2017年。[14] M. L. Rosinberg、G. Tarjus 和 T. Munakata。时延对耦合线性随机系统间信息交换的影响。《Physical Review E》，98:032130，2018年9月。[15] Faezeh Marzbanrad、Yoshitaka Kimura、Marimuthu Palaniswami 和 Ahsan H Khandoker。通过传递熵量化母体和胎儿心率之间的相互作用。《PloS ONE》，10(12):e0145672，2015年。[16] YongKyung Oh、JiIn Kwak 和 Sungil Kim。基于统计因果性的事故导致交通拥堵传播时延估计。《Electronic Research Archive》，31(2):691–707，2023年。[17] Luca Faes、Giandomenico Nollo 和 Alberto Porta。补偿传递熵：可靠估计生理时间序列中信息传递的工具。《Entropy》，15(1):198–219，2013年。[18] Raul Vicente、Michael Wibral、Michael Lindner 和 Gordon Pipa。传递熵——一种面向神经科学的无模型有效连接测度。《Journal of Computational Neuroscience》，30(1):45–67，2011年。

[19]Jian Ma 和 Zengqi Sun。互信息即 Copula熵。《清华科技》，16(1):51–54，2011年。另见 arXiv 预印本 arXiv:0808.0845，2008年。[20] Jian Ma。基于 Copula熵估计传递熵。《arXiv 预印本》，arXiv:1910.04375，2019年。[21] Hao Liu、Dechang Pi、Shuyuan Qiu、Xixuan Wang 和 Chang Guo。面向关联故障传播路径的数据驱动识别模型。《测量》，188:110628，2022年。[22] Muye Han 和 Jinsheng Zhou。基于小波、传递熵和网络分析的投资者情绪传播多尺度特征。《Entropy》，24(12)，2022年。[23] Wei Hu、Qiuting Guo、Wei Wang、Weiheng Wang 和 Shuhong Song。基于传递熵的高损耗配电区域用户损失贡献计算研究。载于《2022年中国国际供电会议（CICED）》，第499–502页，2022年。[24] Chainarong Amornbunchornvej、Elena Zheleva 和 Tanya Berger-Wolf。用于时间序列分析的变时滞格兰杰因果关系和传递熵。《《ACM 数据知识发现汇刊》》，15(4):1–30，2021年6月。[25] Simon Wing、Jay R. Johnson、Enrico Camporeale 和 Geoffrey D. Reeves。利用信息论方法发现外辐射带的太阳风驱动因素。《《地球物理研究杂志：空间物理》》，121(10):9378–9399，2016年。[26] Matth¨aus Staniek 和 Klaus Lehnertz。符号传递熵。《Physical Review Letters》，100:158101，2008年4月。[27] Xuguang Wang、Wei Dong 和 Yan Zhang。线性一阶系统的基于 Copula 的时延估计。载于《2017年应用数学、建模与统计应用国际会议（AMMSA 2017）论文集》，第414–419页。Atlantis Press，2017年。[28] Roger B Nelsen。《Copula 导论》。Springer 科学与商业媒体，2007年。[29] Harry Joe。《基于 Copula 的依赖建模》。CRC 出版社，2014年。[30] M Sklar。多维分布函数及其边缘分布。《巴黎大学统计研究所出版物》，8:229–231，1959年。[31] Alexander Kraskov、Harald St¨ogbauer 和 Peter Grassberger。互信息估计。《Physical Review E》，69(6):066138，2004年。[32] Arthur Asuncion 和 David Newman。UCI 机器学习资料库，2007年。[33] Abdulwahed Salam 和 Abdelaaziz El Hibaoui。功耗预测机器学习算法比较：得土安城案例研究。载于《2018年第六届国际可再生与可持续能源会议（IRSEC）》，第1–5页，2018年。