Copula 熵：理论和应用

马健

日立中国研究院
majian@hitachi.cn

摘要

统计独立性是统计学和机器学习领域的基础性概念，如何表示和度量统计独立性是该领域的基本问题。Copula 理论提供了统计相关性表示的理论工具，而 Copula 熵理论则给出了度量统计独立性的概念工具。本文综述了 Copula 熵的理论和应用，概述了其基本概念定义、定理和性质，以及估计方法。介绍了 Copula 熵研究的最新进展，包括其在统计学四个基本问题（结构学习、关联发现、变量选择和时序因果发现等）上的理论应用。讨论了四个理论应用之间的关系，以及其对应的深层次的相关性和因果性概念之间的联系，并将 Copula 熵的（条件）独立性度量框架与基于核函数和距离的相关性度量框架进行了对比。简述了 Copula 熵在理论物理学、化学信息学、水文学、环境气象学、生态学、认知神经学、运动神经学、计算神经学、系统生物学、生物信息学、临床诊断学、老年医学、公共卫生学、经济政策学、社会学、政治学，以及能源工程、制造工程、可靠性工程、航空航天、通信工程、测绘工程和金融工程等领域的实际应用。

关键词: Copula 熵，传递熵，统计独立性，条件独立性，相关性，因果性，结构学习，关联发现，变量选择，因果发现，交叉学科应用

1 引言

统计独立性是统计学和机器学习领域的基础性概念，如何表示和度量统计独立性是统计学的基本问题。在统计学早期的 19 世纪，就有 Pearson [1] 提出了相关系数的概念来度量统计独立性，并应用于优生学的研究。上个世纪，在对相关性的研究中 Copula 函数理论被提出，提供一种统一表示随机变量之间统计关联关系的理论工具 [2, 3]。根据 Sklar 定理 [4]，通俗地讲，任何一个多变量之间的关联关系都对应着一个用于表示这种关系的函数，称为 Copula 函数。Copula 函数表示了多变量之间全部的关联关系，且与单个变量的性质是无关的。

2008 年，马健与孙增圻提出了 Copula 熵（Copula Entropy: CE）的概念 [5]。CE 的概念由 Copula 密度函数定义而来，本质上是一种香农熵的形式。我

们也证明了它与信息论 [6] 中的互信息概念是等价的。事实上，CE 的提出是受到了这样的启发，Copula 函数被认为包含了全部的关联关系，而互信息一直被认为度量了全部的关联关系的信息，那么我们认为这二者之间必然有某种联系。对这种必然联系的研究的结果，就是提出了 CE 的理论。

CE 是一种多变量之间关联关系度量的理论，与关联关系表示理论——Copula 函数理论相对应。Copula 函数表示关联关系，而由之得到的 CE 度量了关系中的信息量。CE 是一个理想的统计独立性度量的概念，具有很多优美的属性，包括对称性、非正性、单调变换不变性、以及在高斯变量时与相关系数等价等。

CE 是一种理想的统计相关性度量工具，同时它又可以用来表示和度量另一个重要的统计学概念——条件独立性（Conditional Independence：CI）。这样，我们就得到了一个基于 CE 的（条件）独立性度量理论框架，将相关性和因果性这两个基本概念统一起来。

CE 是一个基础性的统计工具，可以用来解决多个统计学的基本问题。我们在 2008 年就将其应用到结构学习问题上 [7]，用来学习统计变量之间的关联关系结构。最近，我们又将其应用到关联发现 [8]、变量选择 [9] 和时序因果发现 [10] 三个问题上，都取得了良好的应用效果。

作为一种基础性的数据分析工具，CE 被提出以来，在多个不同学科得到了实际的应用，包括理论物理学 [11]、化学信息学 [12]、水文学 [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]、环境气象学 [10]、生态学 [25]、认知神经学 [26, 27, 28]、运动神经学 [29, 30, 31]、计算神经学 [32, 33]、系统生物学 [34, 35]、生物信息学 [36, 37]、临床诊断学 [9, 38]、老年医学 [39, 40, 41]、公共卫生学 [38]、经济政策学 [42, 43]、社会学 [44]、政治学 [45]，以及能源工程 [46]、制造工程 [47]、可靠性工程 [48]、航空航天 [49, 50, 51]、通信工程 [52]、测绘工程 [53] 和金融工程 [54, 55, 56] 等。在这些应用中，CE 被用来分析和度量多学科数据中的统计关联性或因果性，用以增加对数据中变量间统计关系的理解，或者用于建立和评价模型。CE 工具不仅带来了建立理论模型时的便利性，同时也改进了计算的可靠性和效率。

本文第 2 部分介绍 Copula 熵的理论和估计方法，第 3 部分介绍 CE 在统计学中的理论应用，用于解决统计学的四个基本问题，第 4 部分讨论三个相关的问题，第 5 部分简要叙述 CE 在多个不同学科中的实际应用，第 6 部分对论文进行总结。

2 Copula 熵

2.1 理论

Copula 理论是关于多随机变量之间相互依赖关系表示的理论 [2, 3]。此理论定义一类函数，成为 Copula 函数，定义如下：

定义 1 (Copula 函数) *给定 N 维随机向量 $\mathbf{X} = (X_1, \ldots, X_N) \in R^N$. 令 $\mathbf{u}$ 表示 $\mathbf{X}$ 的边缘分布函数 $u_i = F_i(x_i), i = 1, \ldots, N$. 则 $\mathbf{X}$ 对应 N 维 Copula 函数 $C : I^N \to I, I = [0, 1]$ 需要满足如下性质:*

1. *C 的下确界为 0 且在单位立方体内的任意子立方体内单调递增;*
2. *$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$.*

直观的理解, Copula 函数就是在单位 N 立方体上的分布函数, 边缘分布为均匀分布, 下确界为 0, 且在任意向上方向上单调增加。从 Copula 函数出发, 对各变量求导, 可以很容易地定义与之相对应的 *Copula* 密度函数 $c(\mathbf{u})$。

Copula 理论的核心结论是 Sklar 定理, 给出了如何利用 Copula 函数表示随机变量依赖关系的结论, 如下:

定理 1 (Sklar 定理) *[4] 给定任意 N 维随机变量 $\mathbf{X}$ 的联合分布函数 $F(\mathbf{X})$、边缘分布函数 $F_i(X_i)$ 和 Copula 函数 $C(\mathbf{u})$, 则联合分布函数可以表示为输入为边缘分布函数的 Copula 函数的形式, 如下:*

$$F(\mathbf{x}) = C(F_1(x_1), \ldots, F_N(x_N)). \tag{1}$$

Copula 函数的表示将多变量的联合分布与单个变量的联合分布分离开来, 将依赖关系表示为一个 Copula 函数。因此, 依赖关系与单个变量的属性是没有关系的, Copula 函数中包含了全部的依赖关系信息。对式 (1)两边求导, 就得到相应的 Sklar 定理的密度函数版本:

$$p(\mathbf{x}) = c(\mathbf{u}) \prod_i p(x_i). \tag{2}$$

其中, $p(\cdot)$ 表示概率密度函数。

利用 Copula 密度函数的表示, 我们就可以定义 Copula 熵, 如下:

定义 2 (Copula 熵) *[5] 给定多随机变量 $\mathbf{X}$, 及其边缘分布 $\mathbf{u}$ 和 Copula 密度函数 $c(\mathbf{u})$, 则 Copula 熵定义为:*

$$H_c(\mathbf{x}) = -\int_{\mathbf{u}} c(\mathbf{u}) \log c(\mathbf{u}) d\mathbf{u}. \tag{3}$$

在信息论中, 互信息 (Mutual Information: MI) 和熵是两个定义不同的概念 [6]。在文献 [5] 中, 我们证明了二者本质上是相同的, 也即是, MI 等价于负的 CE, 也可以表示成熵的形式。定理如下:

定理 2 *多随机变量的 MI 等价于其负的 CE.*

$$I(\mathbf{x}) = -H_c(\mathbf{x}). \tag{4}$$

定理的证明很简单。由定理可以立即得到一条关于联合熵、边缘熵和 CE 之间关系的推论, 如下:

推论 1 多随机变量的联合熵等于边缘熵和 *CE* 的和。

$$H(\mathbf{x}) = \sum_i H(x_i) + H_c(\mathbf{x}). \tag{5}$$

以上结论通过 CE 的定义，加深了我们对信息论基本概念及其之间关系的了解，也因此在 Copula 理论和信息论之间架起了一座桥梁。

2.2 CE 的性质

由 Copula 理论得到的 CE 具有很多有趣的性质。首先从定义来看，CE 是一种特殊的香农熵，定义在单位体的概率分布函数上，因此其也具有香农熵具有的连续性、对称性和可加性等特性。

由于 Copula 函数具有单调变换不变性，因此基于 Copula 函数定义的 CE 天然地继承了这一不变性特性。

由 Copula 密度函数而定义的 CE 从一个新的角度给出了对 MI 概念更深入的理解。Copula 函数被认为是包含了随机变量之间所有相关性的信息，那么 CE 作为相关性的随机性的度量，就等于给出了随机变量之间所有阶次相关性的信息量。

香农的 MI 定义针对的是二变量情况，但 CE 概念不限于二变量的情况，也适用于多变量的情况，且多变量之间具有对称性，扩展了 MI 的定义和适用范围。

上面提到，Copula 理论将联合分布分解为边缘函数和 Copula 函数两个相对独立的部分，这也对应到联合熵的分解：随机变量的联合熵也可以相应地分解为边缘熵和 CE 两个相互无关的部分。而 MI 与 CE 等价，因此 MI（CE）只与 Copula 函数有关，与边缘函数无关、联合函数无关，这与香农基于边缘函数和联合函数的 MI 定义构成了显著的理论区别。

需要指出的是，CE 本身是非正的，它表明了由于多变量之间具有相关性，使得多变量之间相互包含有其他变量的信息，因此就使得联合熵的总信息量减少，表现为联合熵小于各个变量的边缘熵之和。

相关系数是统计学传统的相关性度量，它隐含着分布高斯性的假设。可以很容易证明，在高斯性的情况下，相关系数与 CE 具有数学上的等价关系，即 CE 可以由相关系数矩阵来表示。

2.3 CE 估计方法

MI 作为信息论的基本概念，具有广泛的应用价值。但学界普遍认为 MI 的估计是十分困难的。我们根据定理 2，给出了一个简单且优雅的非参数 CE（MI）估计方法*[5]。该方法仅需如下 2 步：

*本方法已经实现为 R 和 Python 算法包 `copent`[57]，并已分别在 CRAN 和 PyPI 上发布。

1. 估计经验 Copula 密度函数；

2. 由经验 Copula 密度函数估计 CE。

给定随机变量 $\mathbf{X}$ 的一组独立同分布样本 $\{x_1, \ldots, x_T\}$, 可以很容易地通过次序统计量（rank）来估计经验 Copula 密度函数，如下

$$F_i(x_i) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(x_t^i < x_i), \tag{6}$$

其中 $\mathbf{1}(\cdot)$ 表示示性函数。

在得到经验 Copula 密度函数后，第 2 步就是一个熵估计的问题，有很多方法可以采用。我们采用了 Kraskov 等 [58] 提出的 k 近邻法来估计 CE，因为它是一个非参数方法，具有良好的估计性能。

由于在两步中都采用了非参数的方法（次序统计量和 k 近邻法），因此，我们就得到了一个非参数的 CE 估计方法。方法简单，易于实现，且计算量要求较低。此方法是一个典型的基于序数（rank）统计量的非参数估计方法，将 CE 非参数估计的本质等价于计算归一化的序数统计量的熵，内涵深刻。

3 理论应用

3.1 结构学习

从数据分析一组随机变量之间的关联结构，可以帮助我们了解系统内部的内在结构关联性，具有重要的应用价值。在统计和机器学习学习中，表示这种关联结构的主要工具方法是图（Graph），图中的顶点表示随机变量，顶点之间的边表示变量之间的关联，边的权重表示关联的强度。图又分为有向图和无向图，前者的边具有方向而后者则无方向，前者表示变量之间的因果关系而后者表示关联关系。从数据中学习这种关联图结构的问题，被称为结构学习（Structure Learning）。

结构学习的算法很多，其中比较著名的有 Chow-Liu 的图结构学习方法 [59]。该方法通过学习变量的互信息矩阵，再基于矩阵学习最小生成树（Minimal-Spanning-Tree：MST）来得到主要关联结构的骨架。

利用互信息和 CE 的等价性，我们给出了 Chow-Liu 算法的 CE 版本 [7]，包含两步：

1. 利用 CE 估计方法学习得到随机变量的关联矩阵；

2. 再利用 MST 生成算法从上述矩阵得到关联图结构。

由于我们的 CE 估计方法简单有效，相较于传统的互信息估计具有明显优势，因此也使得 Chow-Liu 算法更可靠有效。

我们将算法应用到两个经典的 UCI 机器学习数据集 [60]：鳗鱼生长数据集和波士顿房价数据集。实验结果显示，算法能够得到具有可解释性的关联结构，使我们对数据反映的鳗鱼生长特性和波士顿房价相关因素的内在关系有了更深入的理解。

3.2 关联发现

经验科学是分析数据的学问。通过分析收集的观察或经验数据，人们得出对象系统的科学结论。关联的概念是多元统计分析的基本工具之一。它度量了随机变量之间的统计性内在联系，进而被赋予科学意义。发现关联关系是科学研究的主要内容方法之一。

Pearson 相关系数 [1] 是一种统计学史上重要的相关性度量概念，教科书里都会讲到，应用也很广泛。但由于它是统计学早期提出的概念，因此具有很多局限性。从理论上来讲，它只适用于线性的情况，隐含着高斯分布的假设，使它在绝大多数实际情况中都不适用。它是一个二变量的度量，没有多变量的版本。

CE 则是一种更高级的相关性度量，相对于 Pearson 相关系数具有显著的优势。它没有线性和高斯性的假设，是一个多变量的相关性度量。实际上，CE 度量的是统计独立性，比相关性更宽泛的概念，在统计独立的情况下，其为 0。CE 还具有单调变换不变性，且在高斯分布的情况下，与相关系数等价。简单列一下 CE 的优点：

1. 无模型假设，
2. 可处理非线性关系，
3. 统计独立性度量，
4. 单调变换不变性，
5. 在高斯情况下与相关系数等价。

综合了如此多优点，CE 是一个完美的相关性度量，完全可以替代 Pearson 相关系数，适用于任何类型的相关性度量。Pearson 相关系数作为一个历史悠久的统计工具，可以进入历史了。

关于 CE 与 Pearson 相关系数的理论上的对比，可参见论文 [8]。论文还利用著名的 NHANES 医学体检数据 [61]，从实验上证明了 CE 的显著优越性。

3.3 变量选择

变量选择（Variable Selection），又称特征选择，是统计和机器学习的基本问题 [62, 63]。当人们试图从一组自变量和目标预测变量之间建立函数关系时，往往希望只选择真正与目标变量有内在联系的自变量的一个子集作为函数模型

的输入，以提高模型的科学性（或可解释性），降低模型的复杂度。这样的问题称为变量选择。在统计和机器学习中，变量选择主要用于多元分类/回归分析中建立的函数模型关系。

传统的变量选择方法很多，主要的有准则法、模型正则化方法和关联度量方法。主要的准则法有 AIC[64] 和 BIC[65] 等，通过在似然函数上加上对模型复杂度的惩罚项得到。模型正则化方法主要用于广义线性回归模型，在学习模型的过程中，通过在似然函数上加上模型参数（线性系数）的 1 范数或 2 范数或二者的组合得到，经典的方法包括 LASSO[66]、岭回归（Ridge Regression）[67] 和弹性网络（Elastic Net）[68] 等。以上两类方法都是基于似然函数加惩罚项的形式完成变量选择，都是模型有关的。关联度量的方法则是通过自变量和目标变量之间的关联强度来选择变量，是模型无关的方法。主要的关联关系度量包括传统的 Pearson 相关系数，但它只能度量线性关系，仅适用于线性模型。几个主要的非线性关联度量也都被应用到变量选择问题上，包括希尔伯特-施密特独立性准则（Hilbert-Schmidt Independence Criterion：HSIC）[69, 70] 和距离相关（Distance Correlation：DC）[71, 72] 等。

变量选择问题，推荐 CE 方法，不建议 LASSO 或者 p-value 等传统方法。本方法利用 CE 度量自变量和目标变量之间的关联强度，根据强度从大到小依次选择变量。在变量选择问题上，CE 已被真实数据实验证明优于以下主流变量选择方法：

- LASSO / Ridge Regression / Elastic Net [66, 67, 68],
- AIC / BIC [64, 65],
- Adaptive LASSO [73],
- Hilbert-Schimdt Independence Criterion (HSIC) [69, 70],
- Distance Correlation [71, 72],
- Heller-Heller-Gorfine Tests of Independence [74],
- Hoeffding's D test [75],
- Bergsma-Dassios T* sign covariance [76],
- Ball correlation [77].

实验[†]采用了著名的 UCI 心脏病数据集 [60]，将 CE 方法与以上方法进行对比。该数据集包含了来自世界 4 地的病人临床生理测量数据和诊断结果，用来从生理特征预测心脏病诊断结果。其中部分临床特征已被专家认定为是疾病相关特征，这就为验证变量选择方法提供了一个参照标准。实验结果表明，与其他方

[†]实验 R 代码见：`https://github.com/majianthu/aps2020`。

法相比，CE 方法选择出了最多的疾病相关特征，在预测性和可解释性上优势明显。部分对比结果见图1。

CE 为变量选择问题提供了统一的理论框架。它具有以下优点：

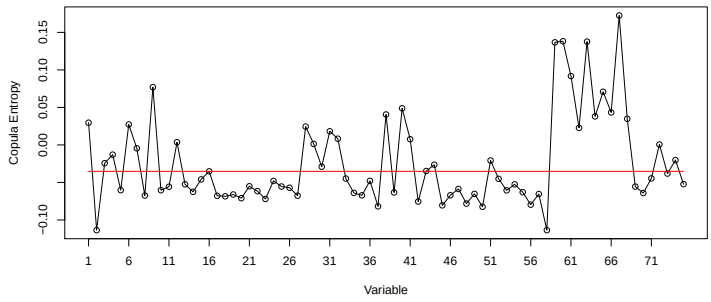
1. 模型无关，
2. 数学理论坚实，
3. 物理上可解释，
4. 具有非参数估计算法，不做理论假设，
5. 几乎不需要调参。

该方法做变量选择是模型无关的，这与基于似然函数的方法相比，无需考虑模型及其复杂度等因素，具有普适性的明显优势。作为一种关联度量工具，CE 与其他度量工具相比定义坚实，具有很多理想的独立性度量公理属性，因此也具有明显的理论优势。另外，熵是一种物理意义明确的数学概念，CE 可被认为是从自变量到目标变量的函数关系包含的信息量，因此很容易从物理上理解和解释得到的模型。在方法实现上，CE 的估计方法基于序数统计量，是非参数的，不做任何理论假设，充分发挥了其理论优势。同时，其估计方法具有良好的渐近稳定性，且几乎不需要调参，与 LASSO 等结果严重依赖超参数选择的方法形成了鲜明对比。总之，该方法具有理论和计算上的明显优势，将变量选择问题变成了一种科学，而不像 LASSO 等方法是一门艺术。

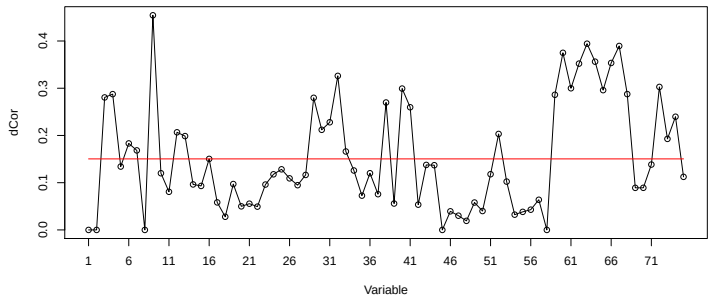
3.4 时序因果发现

因果关系普遍存在于自然界当中，发现因果关系是各门科学的主要命题之一。从一组随机变量的时序观测中发现变量之间的因果关系，被称为因果发现（Causal Discovery）问题，是统计学中时间序列分析的经典问题。时序因果关系发现方法在不同学科领域都有重要应用价值。

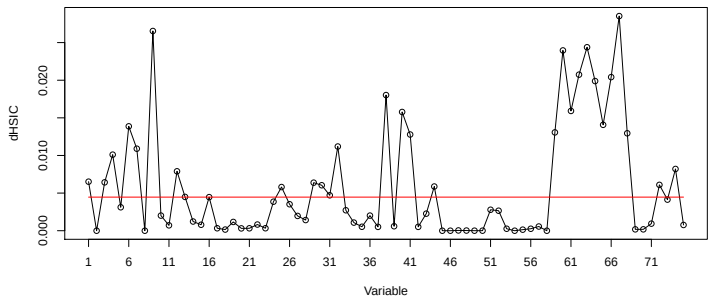
如何度量因果关系是因果发现问题解决的基础。控制论学者维纳提出了一种因果关系的哲学概念，表述为因必须有助于改善果的预测 [78]。在此理念基础上，格兰杰提出了著名的格兰杰因果关系（Granger Causality：GC）检验 [79, 80]。GC 检验是经典的因果发现工具，但它只适用于高斯的情况。Schreiber[81] 定义了用于发现稳态时序包含的因果关系的传递熵（Transfer Entropy：TE）的概念。TE 是 GC 的非线性推广，等价于信息论的条件互信息（Conditional Mutual Information：CMI），本质上是检验条件独立性（Conditional Independence），是模型无关的，因此适用于任何情况的因果关系检验。TE 作为广泛采用的因果关系度量，较之其他经验式带有模型假设的传统因果关系推断方法更科学合理，具有更广泛的普适性。



(a) CE



(b) dCor



(c) dHSIC

图 1: 三种统计独立性度量选择的变量.

CE 是统计独立性度量，而 TE 是条件独立性度量。我们证明了二者之间在数学上有着本质上的内在理论联系 [10]。通过并不复杂的数学变换，可以很容易证明，TE 可以表示为只包含 CE 的数学形式。这就为估计 TE 提供了理论基础。

命题 1 *TE 可以表示为仅包含 CE 的数学形式. 从 X 到 Y 的 TE 的 CE 表示如下：*

$$TE(Y,X) = -H_c(Y_{i+1}, Y_i, X_i) + H_c(Y_{i+1}, Y_i) + H_c(X_i, Y_i). \tag{7}$$

因为 TE 本质上是条件独立性度量，因此(7)也其实是给出了一种条件独立性的 CE 表示。

在过去的研究中，因果关系的估计往往在一定的假设前提下进行，无假设前提的因果关系估计被很多研究者认为是不可能的。我们基于以上 TE 的 CE 表示形式，利用非参数的 CE 估计算法，提出了简单优雅、易于理解和实现的非参数 TE 估计方法 [10]。这样，不带任何假设条件的因果关系发现就成为了可能。此估计方法包含简单的两步[‡]：

1. 利用非参数 CE 估计方法，估计式(7)中的 3 个 CE 子项；
2. 由 3 个 CE 估计值计算得到 TE。

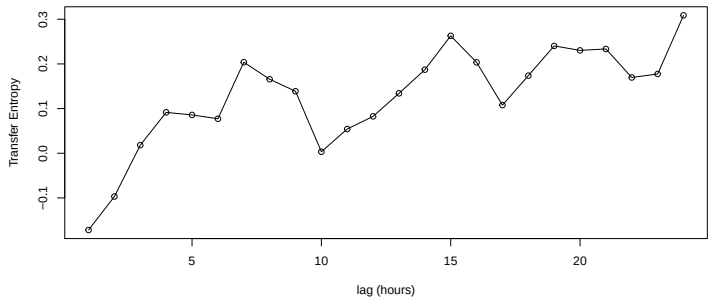
为了验证提出的非参数 TE 估计方法，我们将该方法应用于大气污染问题中的因果发现，研究了北京地区气象因素和 PM2.5 之间的因果关系[§]。实验采用了 UCI 机器学习数据集仓库中的北京 PM2.5 数据 [82]，包含了北京地区 2010 年至 2014 年之间的每小时的连续气象观测数据和 PM2.5 观测数据。我们的分析选择其中一段无缺失值的连续时间数据记录，利用上述方法很容易就可以估计出气象因素对 1 至 24 小时后 PM2.5 浓度的影响程度。利用上述估计方法并不是无条件的，我们默认假设了时序是稳态的，也假设了时间段之间的马尔科夫性，也就是不相邻的时间段之间无关。对 24 小时内滞后因果关系的分析发现，温度、湿度、压力等气象因素对 PM2.5 的形成的因果关系是一个由迅速增加到缓慢增强的过程。

同样在上述实验数据的基础上，我们将提出的 TE 估计方法与另外两种条件独立性度量进行了对比实验，估计从气象因素到 PM2.5 的因果关系 24 小时走势。这两种度量分别是基于核函数的条件独立性度量（Kernel-based Conditional Independence：KCI）[83] 和条件距离相关（Conditional Distance Correlation：CDC）[84]。论文通过将用 CE 估计 TE 与其它两种方法进行了对比，结果（见图2）显示 TE 的估计效果更好。

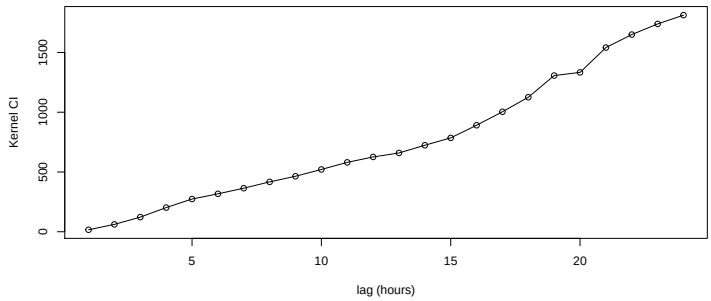
基于 CE 的条件独立性度量作为一种基本的统计学工具，可以用于解决其他的派生统计学问题，比如域迁移（Domain Adaptation：DA）问题。DA 问题

[‡]此方法已在 R 和 Python 包 `copent`[57] 中实现。

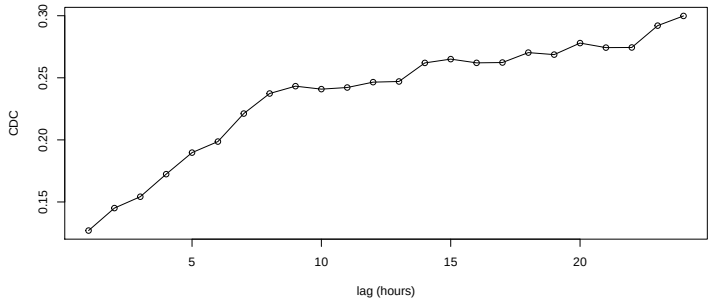
[§]实验 R 代码见：`https://github.com/majianthu/transferentropy`。



(a) TE



(b) KCI



(c) CDC

图 2: 由三种因果关系度量估计的从压力到 PM2.5 的因果关系强度变化图.

是为了处理模型在不同分布的多个域上应用的难题而产生，一般需要对问题做一定的模型假设。马健 [44] 提出了一种从因果角度解决 DA 问题的方法，利用基于 CE 的条件独立性测试发现域迁移背后的因果关系，从而解决了问题。

4 讨论

4.1 四个理论应用之间的联系

以上介绍的 CE 的四个理论应用之间有着内在的联系。从理论基础上讲，它们都是基于 CE 对统计独立和条件独立的度量的理论框架，学习某种内在的统计关系，这是共同点。区别在于四个应用研究的关系不同，以及关联结构的表示方式不同。关联发现问题主要关注成对变量之间的静态的统计相关，表示为相关矩阵的形式；结构学习则关注一组变量之间整体的关联结构，表示为图的形式；变量选择的目的是要建立一个多对一的关联结构，最终要表示为函数的形式；时序因果发现是动态系统中的因果关系，它也可以构建表示变量之间因果关系的有向图结构，也可以用来进行变量选择，构建时序预测的函数关系模型。

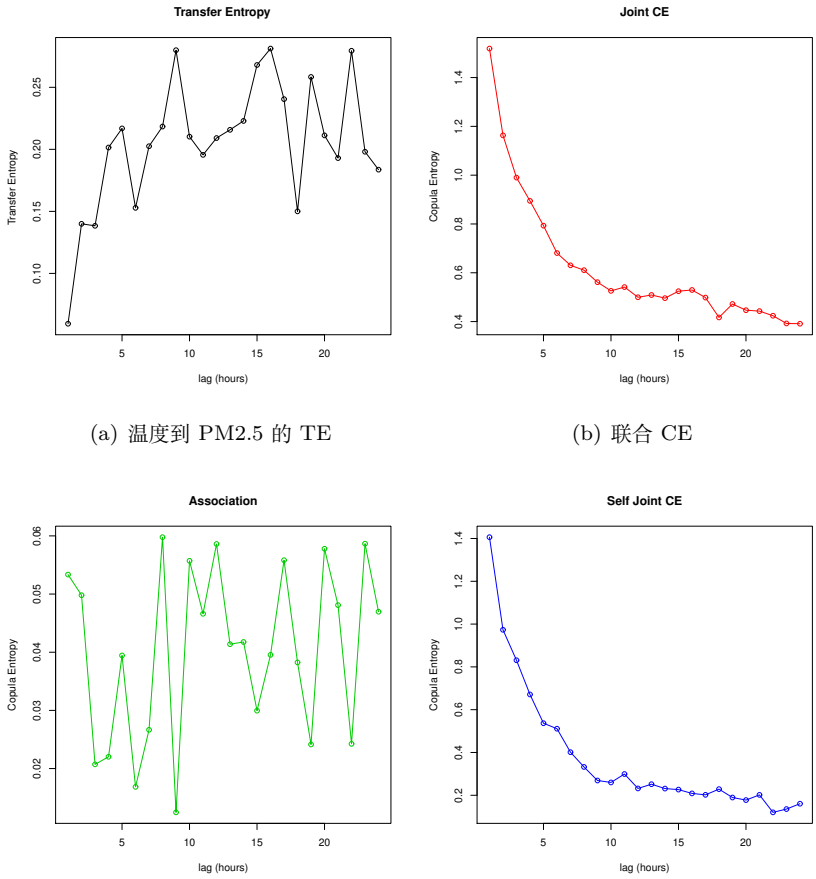
总之，利用 CE 度量统计独立和条件独立关系，可以估计随机变量之间的相关性和因果性关系强度，进而通过相关或因果关系发现表示成基本的矩阵形式，通过结构学习生成直观的无向或有向图的形式，或者通过变量选择构造具有预测能力的静态或动态时序的函数模型的形式。

4.2 相关性和因果性

相关性和因果性是统计学中的两个基础性概念，对应于概率论中的统计独立和条件独立。统计独立和条件独立是两个不同的概念，但又有着内在的联系。我们通过 CE 的概念，给出二者之间的内在联系的理论框架，以及在此理论框架基础上的估计方法。

前者可以用 CE 来衡量。CE 是一个完美的衡量统计独立性/相关性的数学概念，具有很多数学家梦寐以求的独立性度量的公理属性。它等价于信息论中的 MI 概念。后者可以用 TE 来衡量。TE 等价于条件 MI。我们证明了 TE 可以用 CE 来表示。也就是说，条件独立可以通过统计独立来表示和计算。因此二者之间具有内在的理论联系。后者可以用 TE 来衡量。TE 等价于条件 MI。因此，二者之间具有内在的理论联系。

相关性不等于因果性，二者是不同的概念，但人们有时却很容易误把二者等同起来。举一个我们做的时序因果发现的研究 [10] 作为例子加以说明。 论文给出了一种利用 CE 来估计 TE 的算法，并采用了一个环境气象的数据来验证 TE 估计算法 [10]。数据是北京的 PM2.5 观测数据，以及同时观测到的北京地区气象数据。论文实验分析了气象因素（温度、露点、气压和风速等）对 PM2.5 浓度的因果强度，用从时序观测数据中估计的 TE 来衡量，发现了二者之间的



(a) 温度到 PM2.5 的 TE

(b) 联合 CE

(c) 温度和 PM2.5 的关联

(d) PM2.5 的自关联

图 3: 对温度到 PM2.5 的 TE 变化的分解.

因果关系变化规律。

这里要强调的是论文的讨论部分。我们讨论对比了时序相关性和时序因果性，发现即使是气象因素和 PM2.5 浓度之间相关性微弱的情况下，二者之间仍然有时滞因果关系。论文以温度因素为例（图3），对此做了说明。子图 (a) 和 (c) 分别对应 TE 和 CE，也就是因果性和相关性。我们可以发现，相关性强度几乎为 0, 而因果性强度依然很高。

我们认为，这一分析结果是由时序观测的对象系统的动态性造成的，气象因素对 PM2.5 浓度变化的影响不是即时的，而是由于大气系统的内部运动过程，有一个滞后的效应所致。此时，时序变量之间没有即时的相关关系，但存在时滞的因果关系。

表 1: 三种统计独立性度量的对比.

	CE	**DC**	**HSIC**
定义	基于 Copula 函数	相关性的非线性扩展	核函数空间的相关性
多变量	是	distance multivariance	dHSIC
独立性测试	是	total distance multivariance	需要满足核函数条件
条件独立性测试	TE	CDC	KCI
不变性	单调变换不变	无	无
高斯性	与相关系数等价	未知	未知
计算复杂度	$O(n^2)$	$O(n^4)$	$O(n^4)$

4.3 三种理论框架的对比

TE 估计方法将统计独立性度量 CE 用于条件独立性的度量——TE 的表示和估计。从而，我们就提出了一个基于 CE 概念，能够度量独立性和条件独立性两个基本概念的理论框架。与此类似，核函数的方法 [70, 83] 和距离相关的方法 [72, 84] 也可以应用到这两个概念的度量问题上，也分别构成了类似的理论框架。但基于 CE 的理论框架更优越，理论上，CE 的定义更严格；计算上，基于 CE 的估计方法也更简单优雅，普遍适用，且计算量相对要小。

我们利用表1对比了三种统计独立度量概念，可以看到 CE 具有多方面的理论优势。比如，CE 天然的是一个多变量的度量，而其他二者需要通过扩展定义来满足多变量的情况；CE 具有单调变换不变性和在高斯条件下与相关系数等价等属性，而其他二者不具备。在计算成本上，CE 计算复杂度低，而其他二者则具有较高的计算复杂度。

在变量选择和因果发现两个理论应用中，我们利用真实数据对比三种框架中的相应方法。实验结果也表明了 CE 框架的（条件）独立性度量工具均优于其他两个框架中的相应的工具，能够更高效、准确地发现更多的相关或因果关系。

5 实际应用

5.1 理论物理学

热力学是一门古老的理论物理学分支，在 19 世纪由克劳修斯、波尔兹曼和吉布斯等人建立，研究物理系统的宏观状态（如温度）与其微观状态之间的理论联系。熵和热力学第二定律是其最为核心的理论内容。香农的信息论就是受热力学的熵概念启发而建立的。一直以来，热力学和信息论之间的理论联系就是相关领域的重要话题之一。CE 是从信息论领域提出的数学概念，它的物理意义和解释一直未得到研究。马健 [11] 将 CE 理论应用于平衡态相关粒子系统中熵的推导和计算，给出了 CE 的热力学解释，建立了热力学和信息论之间的又一理论联系。

5.2 化学信息学

化学信息学是化学和信息学科的交叉学科，通过表征化学结构为数据，解决诸如分子设计、化学反应模拟和规划等问题。定量构效是该领域的前沿问题，研究分子结构与分子理化性质之间的定量关系，以指导具有指定特性的分子设计，应用广泛。分子理化特性可以理解为分子结构的某种对称变换不变性，而从数据学习得到这种不变性变换是分子设计的关键目标。Wieser 等 [12] 将对称变换学习问题转化为信息瓶颈（Information Bottleneck）问题，提出了一种对称变换信息瓶颈（Symmetry-Transformation Information Bottleneck：STIB）方法。该方法将分子表征表示为由两个部分组成的隐含表示，其中一个部分对应不变性表示，基于 MI（CE）的变换不变性，设计了问题模型的学习算法。作者将算法应用于包含 13.4 万有机分子的 QM9 数据库 [85]，使用其中具有固定化学计量（$C_7O_2H_{10}$）的 6095 个分子的子集，并将其对应的带隙能量和极性作为目标不变性属性。实验结果表明，STIB 方法给出了能够学习出表征分子属性、带隙能量和极性不变性的对称变换，验证了方法的有效性。

5.3 水文学

洪水是主要自然灾害之一，洪水预报是降低洪水损失和管理洪水资源的重要手段。基于降水数据的降水量-径流量模型可以用来预报一段时间后的洪水。但是，水系统具有复杂性和非线性的特点，导致建立这样的模型时选择正确的模型输入十分困难。陈璐等 [13, 14, 15] 提出利用 CE 的方法来选择输入并建立神经网络预报模型。相比于传统的方法，基于 CE 的方法可以建立高维模型且对单个变量的边缘分布不做假设，同时由 CE 来估计降水量和径流量的数量关系的误差更小。陈璐等将方法应用于建立金沙江流域的洪水预报模型，结果显示利用 CE 选择输入的神经网络模型取得了最好的预测效果。

干旱是另一类重要的水文事件和影响重大的自然灾害之一。频发的干旱严重影响着我国的经济社会安全，特别是黄河流域的干旱威胁尤其严重，迫切需要开展流域干旱驱动和预测的研究。温云亮等 [16] 利用 CE 理论分析了河南省 1951–2014 年逐月气象数据，发现在众多驱动因子中，降水量、气温、水气压和相对湿度对干旱发生的影响最大。黄春艳 [17] 研究了黄河流域的气象、水文和干旱之间的关系，探讨了干旱的驱动机制，给出了气象干旱和水文干旱的概念，并提出利用 CE 方法探究二者之间的动态非线性响应关系，通过分析黄河流域不同区域水文站的气象和水文干旱指数，得到了水文干旱对气象干旱的滞后效应时间，为应对干旱事件提供了参考。

水文事件（如洪水、干旱、高温和风暴等）的风险分析和管理需要建立多随机变量的概率模型，研究中大量地使用 Copula 工具，特别是藤 Copula（Vine Copula），解决此类问题。藤 Copula 是一种由二变量 Copula 函数构造高维 Copula 函数的方法，构造时需要确定各个子 Copula 函数间的层累结构关系。

Ni 等 [18] 利用 MI 和 CE 之间的等价关系，提出了基于 MI 的藤 Copula 结构选择方法，并应用于黄河流域干旱识别中特征变量建模问题和多水文站流量相关结构建模问题中。

水文气象观测网络是获取水文信息的基础设施。如何设计并优化网络站点是一个综合性的科学和工程问题。一个基本的设计原则是观测站点之间尽量统计独立，这样才能最大程度的获取水文系统的信息。MI 是衡量统计独立性的主要工具，但是其计算是一个难题。Xu 等 [19] 提出了一个基于 CE 的多目标优化的水文观测网络设计方法，包括两步：1）基于 CE 的信息传输将观测站点分组；2）对每个分组选择最优的站点组合。基于 CE 的计算方法不仅能够处理水文变量的非高斯性，同时在计算性能上也更可靠、更有效率。作者将方法应用于上海降水观测网络的设计。结果显示，CE 的方法计算精度更高，且可以应用于高维的多变量估计情况。同样基于最少重叠信息的原则，Li 等 [20, 21] 提出了一个由两个子目标构成的网络优化目标，其中一个子目标基于 CE 而设计，用于衡量冗余信息量。作者将此方法分别应用于汾河径流观测网、北京市区以及太湖盆地的降水观测网的设计和优化，结果表明了方法可靠且有效。

流域分类是水文学研究的重要方法，根据水文相似性特征划分流域内相似性区域，可解决无水文观测地区的水文计算等难点问题。径流响应是重要的流域水文特征，根据流域水文站点观测之间的相似性做流域分类是一种基本的研究路径。传统的流域分类方法基于相关性评价，往往难以反映水文系统内在的复杂关系。刘磊等 [22] 提出采用基于 CE 的 R 统计量来衡量径流相似性，以对流域进行分区。他们将方法应用于鄱阳湖水系，利用该流域的水文站观测对流域进行了分区，并将方法与传统的 K 均值聚类方法进行了对比。结果表明，该方法能够有效捕捉流域内湖库对径流的调节作用，从而得到较传统方法更合理的流域分区。

多站点径流生成是随机水文学的主要问题之一，生成的流量信息对任何水资源管理都是必不可少的。在径流数据记录有限的情况下，生成多站点径流数据十分必要，需要设计相应的数据生成模型。Porto 等 [23] 提出了结合广义线性模型（GLM）和 Copula 函数的多站点年度径流生成模型，前者表示时序结构，后者为多站点的空间相关性建模。在评价模型性能时，作者采用了包括 CE 在内的多个统计描述性指标，其中 CE 用来衡量非线性的全关联。作者将该模型用于生成巴西的雅瓜里比-大都市水库系统的多站径流时序数据，结果显示模型表现出了优于当前最好水平的性能，特别是在衡量多站相关性的 CE 指标上，较其他模型更接近于历史观测数据。

南水北调工程是当今世界最大的水利工程，承担着从长江的汉江流域丹江口水库向北方地区城市调水的战略任务。准确的入库径流预报是科学合理的供水调度的前提条件，能够使工程更充分高效地利用自然界的水资源。但传统方法构建的预报模型很难满足调水预报精度的要求，原因在于传统分析方法不能处理水文系统的非线性特性，导致了构建的入库径流预报模型不合理从而预测

性能不高。黄朝君等 [24] 构建了丹江口水库的月入库径流预报模型，利用 CE 选择了一组气象水文因子作为模型的输入，得到的模型具有明显优于传统模型的预报性能。模型成功的原因在于采用 CE 选择的预报因子与中长期入库径流密切相关，印证了印度洋偶极子事件和南海副高活动与汉江流域夏季强降水之间的内在联系，符合自然界水文系统的运行规律。

5.4 环境气象学

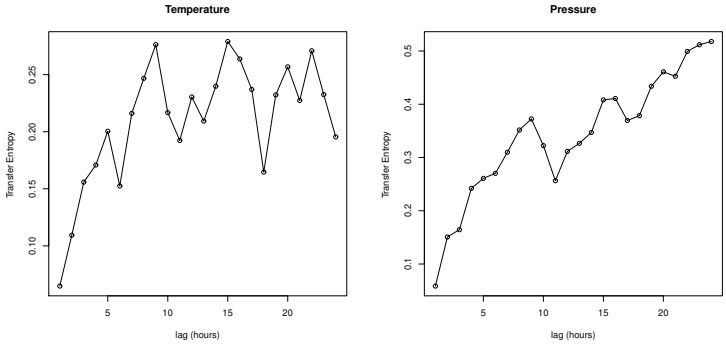
环境污染是现代社会的主要问题之一。从气象学的角度分析大气污染的成因，明晰其内在机理，有助于更好的理解污染问题，进而预测、干预和管理污染。理解大气系统中的因果关系是问题的关键。基于对气象因素和环境污染物的观测，可以利用统计学中的 TE 方法分析气象因素对环境污染的因果关系。马健 [10] 利用其提出的基于 CE 的 TE 估计方法（见3.4），分析了北京地区的气象和 PM2.5 连续观测数据 [82]，得到了四个气象因素对 PM2.5 浓度的 24 小时时滞内的因果强度变化图（见图4）。变化图显示，四种气象因素对 PM2.5 浓度的因果强度大致经历快速升高和缓慢增加两个阶段。作者还特别讨论和验证了该方法的平稳性假设和马尔科夫性假设在此中尺度数值分析问题上的适用性。论文所得到的因果变化图反映了大气系统运动的内在动态特征，增加了人们对 PM2.5 污染的气象成因的理解。同时，得到的时序因果关系也为整合气象因素，构建更优性能的污染预报模型提供了参考依据。（更多内容见3.4节）

5.5 生态学

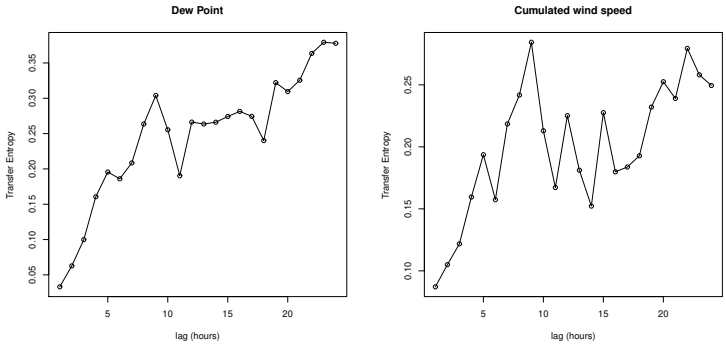
在生态学中，动物运动轨迹研究是一个重要的基本问题，可以揭示种群活动规律、种群间的竞争关系，以及种群和环境资源之间的互动等基本生态学过程。信息技术在生态领域的利用生成了大量的动物轨迹数据，对这些数据的分析需要合理的建模方法。环线数据（circular-linear data）是生态学中的一种常见的时序数据类型，描述了离散化的动物运动过程，包括运动方向和运动距离两个变量。此二变量之间通常是相关的，即直线运动时运动方向较小而运动距离较大，转向运动时运动方向较大而运动距离较小，同时运动方向变量的分布一般是对称的，因此通常采用角度对称的环线 copula 函数作为工具对此类数据进行建模，并利用基于 copula 的相关性度量来衡量二者之间的相关性。Hodel 和 Fleberg [25] 实现了环线 copula 的建模和分析的算法工具包 `Cylcop`，其中包含了基于 CE 的互信息估计算法作为相关性度量方法，用于分析动物轨迹数据。

5.6 认知神经学

认知神经学通过分析大脑活动的各种模态的观测数据，理解大脑作为信息处理器官，对外界刺激的表示、处理和通讯的机理。作为一个非线性的统计度



(a) 温度 (b) 气压



(c) 露点 (d) 累计风速

图 4: 四种气象因素到 PM2.5 浓度的 24 小时时滞内因果强度变化图.

量，MI 被认为是分析大脑信号间关联的理想统计工具。但由于 MI 的估计十分困难，使其难以得到广泛的应用。Ince 等 [27] 根据 MI 和 CE 之间的等价关系，提出了一种 MI 估计方法，称为高斯 Copula 互信息（Gaussian Copula Mutual Information：GCMI）。GCMI 方法利用了 CE 与边缘函数无关的性质，首先将每个变量的边缘函数转化为高斯函数，从而得到联合高斯分布，再根据所得高斯分布相关矩阵与 MI 的关系来计算 MI。该方法简单方便，且与分布无关。但由于从高斯分布数据计算 MI 是有偏差的，因此此方法还需要进行校正纠偏操作。Ince 等将 GCMI 与其他 MI 估计方法进行了对比，并将其应用于分析人脸检测任务的 EEG 数据 [26] 和听觉语音刺激任务的 MEG 数据 [28]。在人脸检测任务的实验中，GCMI 被用来计算图像内容与认知响应之间的关联强度，并成功选出认识响应敏感区域（图像中的眼睛部分）。在听觉刺激实验中，Ince 等研究了语音中的节奏特征对大脑听觉的节律同步的影响。通过对语音刺激的 EEG 响应数据的分析，作者发现了改变音节和词汇之间的停顿会导致听觉 delta 带同步的降低。在此实验中，GCMI 是数据分析的主要工具。

5.7 运动神经学

肌肉协同（Muscle Synergy）是运动的基础，指人完成各种动作时肌肉组合之间时空上的动作协同。人体的运动控制系统是一个具有冗余自由度的系统，一般认为神经系统通过运动基元的组合协同策略来完成一个动作。运动控制研究的一个重要基本问题是鉴别运动控制中简化的基本肌肉协同策略。通过分解运动过程的肌电（Electromyographic：EMG）信号数据理解运动控制潜在的基本协同机理是基本研究手段，但如何处理信号中的非线性是主要的难题之一，基于 CE 的 MI 估计是处理此难题的有力工具。Wu 等 [29, 30] 将多元变分模态分解与基于 CE 的 MI 相结合，构建了肌肉耦合网络模型，基于表面 EMG 数据分析了健康人伸手运动过程中上肢肌肉间的时空协同，成功刻画了肌肉耦合关系强度。Reilly 和 Delis[31] 提出利用基于 CE 的 GCMI 来度量 EMG 信号之间的时空关联关系，再利用矩阵分解的降维方法来发现 EMG 信号时空关联中的基本的肌肉协同模式。他们采集了人执行点到点动作运动的 EMG 数据，将方法应用于数据，得到了有生理学意义的肌肉协同时空模式。

5.8 计算神经学

计算神经学是利用计算理论和方法来研究和理解神经系统的功能和机理的学科，研究如何描述生物神经元对信号刺激的个体和群体响应等问题。神经可塑性（neural plasticity）是指神经元网络对外界刺激的适应性结构变化，构建可塑性理论模型是计算神经学关注的主要问题之一。Leugering 和 Pipa[32] 基于 Copula 理论提出了一个神经元群体可塑性的理论框架，构建了一种自适应网络模型，可以在未知模型输入变化的情况下保持模型输出的不变性，CE 在该框架

中用于度量神经元群的统计特性，衡量输入输出之间的信息量。神经元之间的信息传输分析是计算神经学的另一个重要问题。分析计算神经元之间的信息传输关系需要涉及多个神经元之间的 MI 的分解。部分信息分解（Partial Information Decomposition）就是将 MI 分解为协同（Synergy）、冗余（Redundancy）和独特信息（Unique Information）三个部分的理论。基于 CE 理论和方法，Pakman 等 [33] 提出了一种估计独特信息的方法，并应用于分析多个神经元模型的信息处理。

5.9 系统生物学

系统生物学的一个主要任务是通过生化运动学模型，研究调控、信号传导和代谢过程之间的交互。建立这样的模型需要选择合适的模型输入变量，MI 是变量选择的工具之一。但常用的 kNN 的 MI 估计常常是有偏差的，需要进行修正。Charzyńska 和 Gambin[34] 提出了偏差校正方法，并发现当利用 MI 和 CE 之间的关系估计 MI 时，校正效果显著。作者将方法应用于受到广泛研究的 p53 蛋白和 Mdm2 连接酶之间的负反馈环路问题模型上，结果显示此方法能够比传统的本地敏感性分析方法得出更准确地反映系统行为的模型输入输出关系的分析结果。

系统生物学对分子生物学数据分析的主要目的之一是建立复杂生物现象的网络和动态机制，以分析生命组织的功能和行为。MI 在构建基因通路网络的过程中发挥基础性作用。Farhangmehr 等 [35] 首次提出在网络构建中利用 CE 来估计 MI。他们将方法应用于酵母细胞周期数据，将分析得到的动态网络与京都基因组学百科数据库进行了对照。实验结果显示，利用 CE 来估计 MI 提高了计算效率。

5.10 生物信息学

生物信息学（Bioinformatics）是通过算法分析基因数据（包括基因表达谱数据）来研究生命和疾病机理的新兴学科。基因表达谱是利用 DNA 微阵列技术在基因分子层面观察某一生命组织动态得到的数据，从而能够在基因组水平上反映生命系统的各种现象和机理。Wieczorek 和 Roth[36] 提出了一种研究时间序列数据之间相互作用的分析方法，称为因果压缩（Causal Compression）。与传统的分析全时间序列之间的因果关系不同，该方法研究了基于定向信息（Directed Information）分解的时间序列间相互因果作用的稀疏表达，并据此给出了时序因果分割和因果二分图发现两类问题的解法。基于 CE 与 MI 之间的等价性，作者证明了该方法只与数据分布的 Copula 密度函数有关，并据此设计了求解方法。作者将该方法应用于 NCBI 数据库中的人类 C 型肝炎病毒感染数据（NCBI/GEO 查询号：GSE7123），研究了接受了聚乙二醇干扰素和利巴韦林治疗的重组丙型肝炎病毒核心蛋白基因型 1 感染的基因表达谱时序数据，关注了在干扰素信号

传导中具有重要交互角色的两个基因：转录子 STAT1 和干扰素诱导抗病毒基因 IFIT3，分别生成了二者在有效救治和无效救治病人内相互作用的不同。研究发现，根据分析结果，干扰素疗法消除了大多数有效救治病人体内两种基因之间的关联，而无效救治病人体内的关联则不受影响。同时，分析表明两种病人救治前后二者之间均存在因果交互作用，但对于有效救治病人，早期的 IFIT3 对后期的 STAT1 的影响更显著，这与已有研究结论相符合。

很多疾病的发生与基因结构变异有关。拷贝数变异（Copy Number Variations：CNVs）指长度大于 1kb 的 DNA 片段的变异，在人类基因组中大量存在。作为重要的基因变异，CNVs 包含了大量 DNA 序列、疾病点和功能单元，能为疾病研究提供线索。研究表明，多种癌症的形成和发展与不同的 CNVs 有关。因此，发现不同基因的 CNVs 与不同癌症之间的关系有助于研究癌症病因和诊断方法。从大量的 CNVs 的基因特征中选择出与癌症相关的特征是生物信息学的一个重要问题。Wu 和 Li[37] 提出了一种基因选择方法，称为相关冗余和交互分析（Correlation Redundancy and Interaction Analysis：CRIA）方法，根据 CNVs 选择与癌症有关的基因，以用于癌症分类。CRIA 方法利用了 CE 的多变量相关性特性，设计了基因特征交互强度度量，用于筛选与癌症类型相关性强的基因。他们将该方法应用于 cBioPortal 的癌症基因组数据，利用了其中的 6 种癌症数据，选择出了 200 个与癌症有关的基因。为了验证算法的有效性，他们基于亚利桑那州立大学的数据将方法与其他 8 种基因选择算法进行了对比，结果显示 CRIA 方法选择的基因能够更准确地预测癌症类型。

5.11 临床诊断学

心脏病是最常见的临床疾病之一。医生已经积累了丰富的心脏病临床诊断经验，可以通过各种生理测量结果作出诊断决策。在此经验基础上开发智能临床诊断模型是业界长期追求的目标，开发此类模型的关键在于选择一组生理测量变量来构建预测诊断模型。基于著名的 UCI 心脏病数据集 [60]，马健 [9] 提出采用 CE 作为变量选择方法，用以选择一组生理变量构建诊断模型。该数据集包含了来自世界四地真实的临床心脏病生理测量和诊断数据，其中 13 个生理测量变量被医学专家认定为是临床相关的。实验结果表明，CE 方法选择出了 13 个临床医生认定变量中的 11 个变量，是对比方法中最多的，从而得到了最好的预测准确率。同时，CE 方法还发现了认定变量以外其他与诊断相关的变量，为临床进一步检验提供了新的参考。（更多内容见3.3）

糖尿病是另一种常见临床疾病。对糖尿病人的病情管理与临床诊治结果（发病率和致死率）密切相关，因此建立严格的糖尿病患者住院管理流程对其安全十分重要，这就需要对病情管理标准进行分析研究。为了评估住院患者的救治效果，美国业界建立了健康事实（Health Facts）数据集 [86]，包含了 130 所美国医院和救治网络的糖尿病患者的数据。基于该数据集 1999 至 2008 年的 10 年间 101,721 名住院患者的数据，Mesiar 和 Sheikhi[38] 利用 CE 变量选择方法建

立预测模型，用于从其他 49 个变量预测“是否已用药”变量，取得了良好的预测效果，在仅选择使用 20 个变量的情况下就获得了 97.2% 的准确率，增进了对用药相关变量的认识，构建了合理用药评价模型。

5.12 老年医学

阿尔兹海默病（也称痴呆症）是老年人面对的主要神经退行性疾病之一，临床表现为认知能力的过度衰退等。早期筛查和诊断可以帮助痴呆症患者和家庭及早干预并管理病情发展，可以有效提高病人生活质量，降低家庭和社会成本和负担。简易精神状态量表（Mini-Mental State Examination: MMSE）是临床广泛采用的认知能力筛查工具之一。马健 [39] 通过利用 CE 分析了手指扣击运动（finger tapping）的特征和 MMSE 之间的关联强度，发现一组与 MMSE 相关联的特征，包括扣击频率（或扣击次数或扣击平均时间间隔）等。在此关联关系的基础上，他们构建了从手指扣击特征到 MMSE 的预测模型，取得了良好的预测效果。此预测模型有望用于痴呆症等疾病的认知能力筛查工作中。

跌倒是老年人面对的重大健康风险之一，需要科学管理和及早干预。跌倒预测是管理跌倒风险的重要手段之一。起立行走试验（Timed Up and Go: TUG）是一种主要的跌倒风险评估工具。马健 [40] 提出了一种结合视频分析和机器学习技术的跌倒风险预测方法。该方法首先从老年人进行 TUG 测试的视频中分析出人体 3D 姿态信息，再由一段时间的姿态信息序列计算出一组步态特征，通过利用 CE 分析步态特征和跌倒风险指数之间的关联关系，选择出一组与风险关联的步态特征（包括步幅、步态速度和步态速度的方差等），最后用此特征作为输入构建跌倒风险的预测模型。该方法在真实数据上的实验显示了良好的预测效果。此分析结果也表明了步态特征反映的行动能力与跌倒风险之间的内在联系，使得模型具有临床意义的可解释性。

在以上两个研究的基础上，马健 [41] 还利用 CE 对手指扣击运动特征数据和步态特征数据进行了联合分析，发现了某些手指运动特征与跌倒风险之间具有一定的关联性。这一发现为首次发现，揭示了衰老过程中认知能力和行动能力之间的关联，提供了科学实验证据，加深了对衰老的生理特征的认识和理解。

5.13 公共卫生学

流行病是公共卫生学的重要话题，流行病患者的及时诊断对控制流行病的传播至关重要。感染了流行病毒的病人往往伴有发热等症状，很难与正常的发热病人进行区分。目前正在流行的新型冠状病毒患者就具有这样的发热症状，基于临床数据开发能够区分病毒感染者和正常流感病人的技术成为一个紧迫的问题。然而，相关的症状有 10 几种，如何选择合适的变量集合成为研究成功的关键。Mesiar 和 Sheikhi[38] 基于 CE 变量选择方法，利用真实的临床数据，分析了新冠患者诊断相关的 19 种症状变量，发现年龄、疲劳和恶心呕吐是最重要的

诊断变量，可以使诊断达到 85% 的诊断准确率，如果将诊断变量增加到 15 个，准确率可以提高到 91.4%。

5.14 经济政策学

经济政策的评估需要定量分析，定量分析方法可以科学、客观地评估政策效果。Shan 和 Liu[42, 43] 提出了一种可以定量分析政策组合效果的决策树构建方法，CE 被用来度量非线性相关关系并构建决策树，方法的思想是利用基于 CE 定义的信息增益来构建用以区别不同政策对象群体的政策决策树，由树的叶子节点来表示不同政策组合对应的群体划分。他们将该方法应用于发展经济学领域，评估我国的减贫政策效果，研究分析了 2018 年由政府开展的贫困家庭状况普查的问卷调查数据中四川省的数据。分析发现，就业政策、新收入来源和是否有抵押贷款是影响家庭收入的主要政策因素，并揭示了这些政策组合对应的不同目标贫困群体收入结构的不同特征。该方法在无历史数据的情况下，评估验证了减贫政策的有效性，并发现了更加有效的政策组合方案。

5.15 社会学

性别不平等是社会学研究的问题之一。由性别视角，我们可以发现很多不平等现象，如两性在收入上、教育上、职业上的不平等等。分析和鉴别导致不平等现象的社会学因素是学者们关心的问题，利用定量方法分析相关社会学数据是研究的手段之一。然而各种社会因素之间的因果链条十分复杂，需要采用科学的数据分析工具加以应对。马健 [44] 提出了一种多域因果关系鉴别方法，将性别因素作为社会外在变量，将不平等问题转化为数据分析中的域迁移问题，利用基于 CE 的条件独立性测试发现社会变量之间的因果关系。他将方法应用于美国国家成人收入社会调查数据，分析了性别、教育和收入之间的因果关系链条，发现了性别导致教育不平等，进而造成收入不平等的科学证据。

5.16 政治学

政治安全事关国家安危。政治学研究关心政权领导力因素与政权危机之间的关系，并根据这些信息配置资源，开展情报收集、稳定或颠覆政权等行动。基于雪城大学莫伊尼汉全球事务研究所的国际政治领导力数据集，Card[45] 研究了 37 个领导力因素与政治安全之间的非线性关系，采用 CE（MI）作为非线性分析工具，重点关注了两个领导力变量（政权建立原因和政权结束原因）与其他因素的关系。分析结果佐证了社会学家的已有理论，分析也印证了已知的关系，发现了未知的关系和现象。

5.17 能源工程

天气是能源系统的重要影响因素，直接影响能源的生产和消费两端。特别是当可再生能源整合到能源系统中后，风速和光照等天气因素决定了风能和光伏能源的生产能力，而温度变化则会影响居民的能源消耗需求。但自然系统具有较大的随机性，给新能源系统的稳定高效运行带来了挑战。因此，新型能源网络管理系统需要建立合理的模型，以便将新能源集成到网络中。信息论为管理天气系统的随机性提供了工具。Fu 等 [46] 研究了基于信息论在集成能源系统中建立天气模型的方法。作者采用了 Copula 函数建立天气变量的联合分布模型，并采用 CE 计算的 MI 作为模型准确性的评价指标，以指导建模过程。同时，MI 还被用来衡量各种能源产出之间的关联强度。作者将得到的集成能源系统模型用于模拟中国北方某地区的能源系统运行情况，并与实际数据进行了对比。结果显示，系统模型的模拟与实际情况基本符合，说明构建的天气模型能够满足能源管理系统运行需求。

5.18 制造工程

产品质量是制造业的生命。注射成型（injection molding）是近年快速发展的工业制造技术，在航天、建筑、通讯等领域有着广泛应用。注射成型过程包括了多步复杂的物理和化学反应过程，很容易受到外部因素的影响，保证塑料产品质量的稳定性是一个难题。基于制造过程历史数据，建立产品质量预测模型是提高产品质量的手段之一。但建立模型需要首先选择有关的过程参数作为模型输入，以获得较好的预测性能。Sun 等 [47] 提出基于 CE 方法选择过程参数变量用于构建质量预测模型，并将方法应用于真实的富士康公司的注射成型生产过程数据，大幅改善了质量预测的性能。

5.19 可靠性工程

退化过程（degradation processes）在各种工程系统中普遍存在，导致系统可靠性的降低甚至失效，如金属材料的疲劳和腐蚀、半导体器件的参数漂移等。退化过程建模是评估系统和产品有效性和寿命的主要技术手段之一。由于现代系统的复杂性，影响退化过程的因素较多，因素变量本身具有非线性特征，且变量之间又相互关联，从而对退化过程建模构成了可靠性工程的一个基本难题。如果建模时忽略了因素之间的相关性，就会导致模型错误和可靠性估计误差。传统的衡量因素之间的相关性主要采用线性相关系数，难以处理复杂的相关关系。Sun 等 [48] 提出采用 copula 对过程因素之间关系建模，并用 CE 来度量退化过程因素之间的关联。他给出了一种参数化 CE 估计方法，并成功应用于微波电子组件的退化过程分析中。结果表明，该方法能够分析不同阶段的退化过程。

5.20 航空航天

航空飞行器系统日趋复杂，飞行器设计首先需要加深对其总体设计参数的认识。对各种设计参数间的耦合关系的理论分析，有助于分析设计方案可行性或优化总体设计方案。Krishnankutty 等 [49] 基于 CE 与 MI 的等价关系，提出了两种基于 Copula 的 MI 估计方法，并将方法应用于美国 22 种喷气战斗机的技术参数数据的分析，估计了飞行航程和可承受负载之间的耦合关系，验证了分析方法的有效性。

卫星是航天时代的主要航天器类型，在信息时代有着广泛的民事和军事用途。作为一种在极端环境运行的复杂系统，卫星的在轨健康状态监测十分重要。卫星遥测数据是各种传感器参数的编码，包含了卫星内部运行系统物理参数的交互关系信息。卫星的异常模式会由于这种交互而在内部传播，因此分析这种内部交互导致的故障传播链条有助于及时发现卫星异常状态，保障卫星正常运行。分析遥测参数之间的因果关系是一种解决问题的路径。Liu 等 [50] 提出直接将基于 CE 的 TE 应用于分析真实的卫星遥测数据，得到了遥测参数之间的故障传导图，结果要优于传统的 TE 方法。Zeng 等 [51] 提出了一种改进的 TE 度量，称为 NMCTE，用于分析遥测参数之间的因果关系网络，该度量利用了基于 CE 的 TE 表示和估计方法。他们又提出了基于所得因果网络的异常检测的 CN-FA-LSTM 方法。他们将 NMCTE 方法应用于真实的卫星遥测数据，得到了具有良好的可解释性的因果网络。他们又将 CN-FA-LSTM 方法在 NASA 公开的 SMAP 和 MSL 数据集上与其它 6 种方法进行了对比，验证了方法的优越性。

5.21 通信工程

通信安全是移动通讯的主要关切之一，一般通过通信层的加密技术加以解决。在资源受限的新兴网络（如 IoT、WSN 等）中，密钥分发是一个挑战。无线信道的互易性为通信双方提供了共享密钥的机制，双方可通过测量无线信道获取密钥。密钥容量概念为无线信道密钥提取提供了理论上限。然而，现实中密钥容量往往受到诸多实际物理条件（如终端移动、信道噪声等）的限制，需要对其进行定量分析。Wang 等 [52] 研究了均匀散射环境下物理因素对密钥容量的影响，将其转化为随机变量的 MI 计算问题，并基于仿真物理环境验证其理论推导的正确性，仿真实验采用了基于 CE 的 MI 估计算法估计密钥容量。仿真结果表明，理论推导得到了验证，能够指导实际应用。

5.22 测绘工程

高光谱遥感是应用广泛的前沿测绘技术，通过遥感光谱成像，能够获取不同地物的诊断性光谱信息。由于高光谱图像波段数多，数据大且存在大量冗余信息，需要利用特征提取技术对有效波段进行选择，以表征成像对象体。因此，

高光谱图像波段选择是该领域的重要问题之一，主要思想是选择一个波段子集，使得成像评价准则函数达到最大。其中，基于信息论的准则是波段选择的主要方法之一。Zeng 和 Durrani[53] 提出利用基于 CE 的 MI 选择波段的方法，并将其应用于美国印第安纳西北的 Indian Pine 处采集的真实高光谱数据，结果表明 CE 提供了一种鲁棒的 MI 波段选择方法。

5.23 金融工程

量化金融是通过对金融数据的数量关系分析指导金融决策的新兴金融学科。基于金融交易系统产生的大量金融市场交易数据，利用数学工具分析金融产品之间的数量关系，可以明晰市场规律和动态，进而管理金融资产。其中，分析市场金融变量之间的相关性是金融工程的重要问题，可以帮助交易员洞察它们之间的动态关系，进而调整投资组合和管理风险。由于金融市场变量具有非线性、非高斯性等特征，使得 MI 成为了理想的相关性度量，而 MI 估计算法则成了量化金融工具箱的重要工具之一。基于 CE 的 MI 估计算法就被量化金融算法库 `MLFinLab`[54] 实现，并得到业界广泛应用。

基于中国股票市场（沪市 A 股指数、深市 A 股指数和沪深 300 指数）真实数据，Wang[55] 研究了利用股票资产之间的相关性关系网络，优化投资组合的方法。方法采用了包括 CE 在内的线性和非线性相关性度量，基于相关性强度构建股票资产间的关系网络，进而构建投资组合。研究中估计了不同 Copula 参数函数族的 CE（MI）。

分析金融数据需要对其建模数学模型，但金融变量以及其联合分布具有非高斯性，给数据建模带来了挑战。Calsaverini 和 Vicente[56] 给出了一种巧妙的 Copula 函数模型选择方法。该方法利用 CE（MI）的边缘分布无关特性，将 Copula 鉴别问题的目标与边缘函数分开，再利用 CE 的定义，将问题转化为以 MI 为上界的模型选择问题。作者还定义了超量信息（Informaion Excess）的概念。作者将建模方法应用于 1990 至 2008 年间标普 500 指数的 150 只股票的每日对数收益率数据，利用超量信息，验证了该方法作用于 T-Copula 函数族时的有效性。

6 总结

统计独立性是统计学和机器学习领域的基础性概念，如何表示和度量统计独立性是该领域的基本问题。Copula 理论提供了统计相关性表示的理论工具，通过将随机变量的边缘函数与表示统计关联性的 Copula 函数相分离，得到了表示任何关联性的数学形式。而 CE 理论则给出了度量统计独立性的概念工具，度量了 Copula 函数表示中所有的信息量，也就是相关性的强度。CE 是一种具有诸多公理性属性的理想的统计度量工具。

本文综述了 CE 的理论和应用，介绍了 CE 基本概念定义、与 MI 等价性的定理和推论，以及 CE 的性质。介绍了 CE 的非参数估计方法。本文介绍了 CE 研究的最新进展，包括其在统计学四个基本问题（结构学习、关联发现、变量选择和时序因果发现等）上的理论应用，讨论了四个理论应用之间的关系，探讨了四个应用对应的深层次的相关性和因果性概念之间的联系，并将基于 CE 的（条件）独立性度量框架与基于核函数和距离的相关性度量框架进行了对比，指出了本理论框架在多个方面的优越性。

本文综述了 CE 在理论物理学、化学信息学、水文学、环境气象学、生态学、认知神经学、运动神经学、计算神经学、系统生物学、生物信息学、临床诊断学、老年医学、公共卫生学、经济政策学、社会学、政治学，以及能源工程、制造工程、可靠性工程、航空航天、通信工程、测绘工程和金融工程等多学科领域的实际应用。基于 CE 带来的理论和计算上的优势，在这些应用中 CE 被用来分析和度量各种类型数据中的统计关联性或因果性，通过选择变量来建立模型，以及作为评价指标评价模型，均取得了良好的应用效果。

附：软件实现

本文所述的 CE 估计算法和 TE 估计算法已在 R 和 Python 语言的算法包 `copent` 中实现 [57]，并分别在 CRAN 和 PyPI 上共享，相关源码见作者的 GitHub：`https://github.com/majianthu/`。另，CE 估计算法的 Julia 语言版本实现见 JuliaHub 上 `CopEnt` 算法包。

参考文献

[1] Karl Pearson. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of The Royal Society of London*, 60(1):489–498, 1896.

[2] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[3] Harry Joe. *Dependence modeling with copulas*. CRC press, 2014.

[4] Abe Sklar. Fonctions de repartition a n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

[5] Jian Ma and Zengqi Sun. Mutual information is copula entropy. *Tsinghua Science & Technology*, 16(1):51–54, 2011. See also arXiv preprint arXiv:0808.0845 (2008).

[6] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[7] Jian Ma and Zengqi Sun. Dependence structure estimation via copula. *arXiv preprint arXiv:0804.4451*, 2008.

[8] Jian Ma. Discovering association with copula entropy. *arXiv preprint arXiv:1907.12268*, 2019.

[9] Jian Ma. Variable selection with copula entropy. *Chinese Journal of Applied Probability and Statistics*, 37(4):405–420, 2021. See also arXiv preprint arXiv:1910.12389 (2019).

[10] Jian Ma. Estimating transfer entropy via copula entropy. *arXiv preprint arXiv:1910.04375*, 2019.

[11] Jian Ma. On thermodynamic interpretation of copula entropy. *arXiv preprint arXiv:2111.14042*, 2021.

[12] Mario Wieser, Sonali Parbhoo, Aleksander Wieczorek, and Volker Roth. Inverse learning of symmetries. In *Advances in Neural Information Processing Systems*, volume 33, pages 18004–18015, 2020.

[13] Lu Chen, Vijay P. Singh, and Shenglian Guo. Measure of correlation between river flows using the copula-entropy method. *Journal of Hydrologic Engineering*, 18(12):1591–1606, 2013.

[14] Lu Chen, Vijay P. Singh, Shenglian Guo, Jianzhong Zhou, and Lei Ye. Copula entropy coupled with artificial neural network for rainfall–runoff simulation. *Stochastic Environmental Research and Risk Assessment*, 28(7):1755–1767, 2014.

[15] Lu Chen and Vijay P. Singh. Flood forecasting and error simulation using copula entropy method. In Priyanka Sharma and Deepesh Machiwal, editors, *Advances in Streamflow Forecasting*, pages 331–368. Elsevier, 2021.

[16] 温云亮, 李艳玲, 黄春艳, and 张泽中. 基于 copula 熵理论的干旱驱动因子选择. *华北水利水电大学学报 (自然科学版)*, 40(4):51–56, 2019.

[17] 黄春艳. *黄河流域的干旱驱动及评估预测研究*. PhD thesis, 西安理工大学, 2021.

[18] Lingling Ni, Dong Wang, Jianfeng Wu, Yuankun Wang, Yuwei Tao, Jianyun Zhang, Jiufu Liu, and Fei Xie. Vine copula selection using mutual information for hydrological dependence modeling. *Environmental Research*, 186:109604, 2020.

[19] Pengcheng Xu, Dong Wang, Vijay P. Singh, Yuankun Wang, Jichun Wu, Lachun Wang, Xinqing Zou, Yuanfang Chen, Xi Chen, Jiufu Liu, Ying Zou, and Ruimin He. A two-phase copula entropy-based multiobjective optimization approach to hydrometeorological gauge network design. *Journal of Hydrology*, 555:228–241, 2017.

[20] Heshu Li, Dong Wang, Vijay P. Singh, Yuankun Wang, Jianfeng Wu, Jichun Wu, Ruimin He, Ying Zou, Jiufu Liu, and Jianyun Zhang. Developing a dual entropy-transinformation criterion for hydrometric network optimization based on information theory and copulas. *Environmental Research*, 180:108813, 2020.

[21] Heshu Li, Dong Wang, Vijay P. Singh, Yuankun Wang, Jianfeng Wu, and Jichun Wu. Developing an entropy and copula-based approach for precipitation monitoring network expansion. *Journal of Hydrology*, 598:126366, 2021.

[22] 刘磊, 高超, 王志刚, 王晓艳, 章四龙, and 陈娜. 基于非线性相关性和复杂网络的径流相似性分区. 水科学进展, 1(1):1, 2022.

[23] Victor Costa Porto, Francisco de Assis de Souza Filho, Taís Maria Nunes Carvalho, Ticiana Marinho de Carvalho Studart, and Maria Manuela Portela. A GLM copula approach for multisite annual streamflow generation. *Journal of Hydrology*, 598:126226, 2021.

[24] 黄朝君, 贾建伟, 秦赫, and 王栋. 基于 copula 熵-随机森林的中长期径流预报研究. 人民长江, 52(11):81–85, 2021.

[25] Florian H. Hodel and John R. Fieberg. Cylcop: An R package for circular-linear copulae with angular symmetry. *bioRxiv*, page 2021.07.14.452253, 2021.

[26] Robin A. A. Ince, Katarzyna Jaworska, Joachim Gross, Stefano Panzeri, Nicola J. van Rijsbergen, Guillaume A. Rousselet, and Philippe G. Schyns. The deceptively simple N170 reflects network information processing mechanisms involving visual feature coding and transfer across hemispheres. *Cerebral Cortex*, 26(11):4123–4135, 2016.

[27] Robin A.A. Ince, Bruno L. Giordano, Christoph Kayser, Guillaume A. Rousselet, Joachim Gross, and Philippe G. Schyns. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human Brain Mapping*, 38(3):1541–1573, 2017.

[28] Stephanie J. Kayser, Robin A.A. Ince, Joachim Gross, and Christoph Kayser. Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *The Journal of Neuroscience*, 35(44):14691–14701, 2015.

[29] 吴亚婷, 佘青山, 高云园, 谭同才, and 范影乐. 多尺度肌间耦合网络分析. *生物医学工程学杂志*, 38(4):742–752, 2021.

[30] Yating Wu, Qingshan She, Hongan Wang, Yuliang Ma, Mingxu Sun, and Tao Shen. R-vine copula mutual information for intermuscular coupling analysis. In *Proceedings of the 11th International Conference on Computer Engineering and Networks*, pages 526–534, 2022.

[31] David Ó' Reilly and Ioannis Delis. A network information theoretic framework to characterise muscle synergies in space and time. *Journal of Neural Engineering*, 19(1):016031, feb 2022.

[32] Johannes Leugering and Gordon Pipa. A unifying framework of synaptic and intrinsic plasticity in neural populations. *Neural Computation*, 30(4):945–986, 2018.

[33] Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables in recurrent networks. *Advances in Neural Information Processing Systems*, 2021.

[34] Agata Charzyńska and Anna Gambin. Improvement of the k-NN entropy estimator with applications in systems biology. *Entropy*, 18(1):13, 2015.

[35] Farzaneh Farhangmehr, Daniel M. Tartakovsky, Parastou Sadatmousavi, Mano R. Maurya, and Shankar Subramaniam. An information-theoretic algorithm to data-driven genetic pathway interaction network reconstruction of dynamic systems. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 214–217, 2013.

[36] Aleksander Wieczorek and Volker Roth. Causal compression. *arXiv preprint arXiv:1611.00261*, 2016.

[37] Qiang Wu and Dongxi Li. CRIA: An interactive gene selection algorithm for cancers prediction based on copy number variations. *Frontiers in Plant Science*, 13, 2022.

[38] Radko Mesiar and Ayyub Sheikhi. Nonlinear random forest classification, a copula-based approach. *Applied Sciences*, 11(15), 2021.

[39] Jian Ma. Predicting MMSE score from finger-tapping measurement. In *Proceedings of 2021 Chinese Intelligent Automation Conference*, pages 294–304, 2022. See also bioRxiv 817338 (2019).

[40] Jian Ma. Predicting TUG score from gait characteristics based on video analysis and machine learning. *bioRxiv*, page 963686, 2020.

[41] Jian Ma. Associations between finger tapping, gait and fall risk with application to fall risk assessment. *arXiv preprint arXiv:2006.16648*, 2020.

[42] Qingsong Shan and Qianning Liu. Binary trees for dependence structure. *IEEE Access*, 8:150989–150998, 2020.

[43] 罗良清, 平卫英, 单青松, and 王佳. 中国贫困治理经验总结：扶贫政策能够实现有效增收吗？. 管理世界, 38(2):70–83, 2022.

[44] Jian Ma. Causal domain adaptation with copula entropy based conditional independence test. *arXiv preprint arXiv:2202.13482*, 2022.

[45] Stuart William Card. Towards an information theoretic framework for evolutionary learning. Master's thesis, Syracuse University, 2011.

[46] Xueqian Fu, Hongbin Sun, Qinglai Guo, Zhaoguang Pan, Wen Xiong, and Li Wang. Uncertainty analysis of an integrated energy system based on information theory. *Energy*, 122(122):649–662, 2017.

[47] Yan-Ning Sun, Yu Chen, Wu-Yin Wang, Hong-Wei Xu, and Wei Qin. Modelling and prediction of injection molding process using copula entropy and multi-output SVR. In *IEEE 17th International Conference on Automation Science and Engineering*, 2021.

[48] Fuqiang Sun, Wendi Zhang, Ning Wang, and Wei Zhang. A copula entropy approach to dependence measurement for multiple degradation processes. *Entropy*, 21(8):724, 2019.

[49] Baby Alpettiyil Krishnankutty, Rajesh Ganapathy, and Paduthol Godan Sankaran. Non-parametric estimation of copula based mutual information. *Communications in Statistics - Theory and Methods*, 49(6):1513–1527, 2020.

[50] Hao Liu, Dechang Pi, Shuyuan Qiu, Xixuan Wang, and Chang Guo. Data-driven identification model for associated fault propagation path. *Measurement*, 188:110628, 2022.

[51] Zefan Zeng, Guang Jin, Chi Xu, Siya Chen, Zhelong Zeng, and Lu Zhang. Satellite telemetry data anomaly detection using causal network and feature-attention-based lstm. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022.

[52] Xu Wang, Liang Jin, Kaizhi Huang, Mingliang Li, and Yi Ming. Physical layer secret key capacity using correlated wireless channel samples. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2016.

[53] Xuexing Zeng and T S Durrani. Band selection for hyperspectral images using copulas-based mutual information. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 341–344, 2009.

[54] Hudson and Thames. Machine learning financial laboratory (MLFinLab), 2021. URL: `https://github.com/hudson-and-thames/mlfinlab`.

[55] Qiutong Wang. Social networks, asset allocation and portfolio diversification. Master's thesis, University of Waterloo, 2015.

[56] Rafael Calsaverini and Renato Vicente. An information-theoretic approach to statistical dependence: Copula information. *EPL (Europhysics Letters)*, 88(6):68003, 2009.

[57] Jian Ma. copent: Estimating copula entropy and transfer entropy in R. *arXiv preprint arXiv:2005.14025*, 2021.

[58] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):66138, 2004.

[59] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[60] Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

[61] National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey, 2013-2014.

[62] Edward I. George. The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308, 2000.

[63] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[64] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[65] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[66] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B-Methodological*, 58(1):267–288, 1996.

[67] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[68] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 67(2):301–320, 2005.

[69] Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, volume 20, pages 585–592, 2007.

[70] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 80(1):5–31, 2018.

[71] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.

[72] Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

[73] Hui Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[74] Ruth Heller, Yair Heller, Shachar Kaufman, Barak Brill, and Malka Gorfine. Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(1):978–1031, 2016.

[75] Wassily Hoeffding. A non-parametric test of independence. *Annals of Mathematical Statistics*, 19(4):546–557, 1948.

[76] Wicher Bergsma and Angelos Dassios. A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20(2):1006–1028, 2014.

[77] Wenliang Pan, Xueqin Wang, Heping Zhang, Hongtu Zhu, and Jin Zhu. Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association*, 115(529):307–317, 2020.

[78] Norbert Wiener. The theory of prediction. modern mathematics for engineers. *New York*, 165, 1956.

[79] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[80] Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

[81] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.

[82] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257, 2015.

[83] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI'11 Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.

[84] Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

[85] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022–140022, 2014.

[86] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:781670, 2014.