

---

# COPULA 縮

## 理论和应用

---

马健

2025 年 10 月 12 日



# 前言

Copula 理论是概率论领域的重要内容，经过几十年的发展已经日臻成熟。Copula 理论体系基于 Sklar 定理给出了依赖性函数表示的 Copula 函数统一形式，进而通过 Copula 函数表示来对依赖性进行推断和度量。目前，学界已经提出了很多表示不同类型依赖性的参数 Copula 函数族及其构造方法，并给出了典型的二变量依赖性度量（如 Spearman 的  $\rho$  和 Kendall 的  $\tau$  等）的 Copula 函数表示形式。基于 Copula 函数的多变量依赖性函数表示和度量一直是学界关心的问题，有待进一步发展完善。

本书内容最初起始于作者在清华大学的博士论文研究，期间作者给出了 Copula 熵的定义和估计方法，证明了其与信息论中互信息概念的等价性，同时进行了基于 Copula 理论的结构学习算法和 Copula 元分析的研究工作。十几年来，作者在博士论文的基础上不断探索，成功将 Copula 熵应用于十个统计学领域的基本问题，特别是证明了传递熵/条件独立性的 Copula 熵表示形式，形成了基于 Copula 熵的相关性和因果性相统一的理论框架，以及提出了四个基于 Copula 熵的假设检验基本问题的方法论。与此同时，学界也对 Copula 熵概念进行了理论扩展和大量的多学科实际应用。Copula 熵研究已经形成了一个理论体系，并成为了 Copula 理论领域的重要组成部分，本书的目的就是要系统地介绍 Copula 熵理论框架和最新研究进展。

Copula 熵理论给出了一个基于 Copula 函数定义的多变量独立性度量概念，具有多重理论价值，体现在

- 发展了 Copula 函数理论体系的内容，给出了理论完善的基于 Copula 函数的多变量独立性/依赖性度量；
- 通过 Copula 熵和（条件）互信息等价性的证明，在 Copula 理论/概率论和信息论之间建立了理论联系，使得数学大厦更加浑然一体；
- 构建了相关性和因果性相统一的理论框架，提出了独立性检验、条件独立性检验、多元正态性检验、双样本检验、变点检测和对称性检验等统计学核心问题的方法论工具，促进了数理统计学的基础理论和方法论体系的成熟。

本书的内容还在不断完善中，欢迎读者朋友给出不断改进的意见和建议。

马健  
北京海淀  
2025 年 5 月 31 日



# 目录

前言	i
引言	1
<b>第一章 Copula 熵</b>	<b>5</b>
1.1 背景 . . . . .	5
1.1.1 Copula 理论 . . . . .	5
1.1.2 信息论 . . . . .	6
1.2 理论 . . . . .	6
1.3 性质 . . . . .	8
1.4 估计方法 . . . . .	10
<b>第二章 理论应用</b>	<b>11</b>
2.1 结构学习 . . . . .	11
2.2 关联发现 . . . . .	12
2.3 变量选择 . . . . .	18
2.4 因果发现 . . . . .	19
2.5 系统辨识 . . . . .	24
2.6 时延估计 . . . . .	25
2.7 域自适应 . . . . .	29
2.8 正态性检验 . . . . .	32
2.9 双样本检验 . . . . .	33
2.10 变点检测 . . . . .	34
2.11 对称性检验 . . . . .	37
<b>第三章 讨论</b>	<b>39</b>
3.1 理论应用之间的联系 . . . . .	39
3.2 相关性和因果性 . . . . .	39
3.3 三种理论框架的对比 . . . . .	40

<b>第四章 仿真评测</b>	<b>43</b>
4.1 概述 . . . . .	43
4.2 独立性检验 . . . . .	43
4.3 条件独立性检验 . . . . .	47
4.4 正态性检验 . . . . .	55
4.5 双样本检验 . . . . .	57
4.6 变点检测 . . . . .	68
4.7 对称性检验 . . . . .	71
<b>第五章 数学推广</b>	<b>79</b>
5.1 Tsallis Copula 熵 . . . . .	79
5.2 生存 Copula 熵 . . . . .	80
5.3 Copula 外熵 . . . . .	81
5.4 累积 Copula 熵 . . . . .	82
5.5 累积 Copula Tsallis 熵 . . . . .	83
5.6 Copula Rényi 熵 . . . . .	84
5.7 Copula Rényi 散度和 Copula Tsallis 散度 . . . . .	86
5.8 Copula Jeffreys 散度和 Copula Hellinger 散度 . . . . .	88
5.9 Copula 分形不准确度 . . . . .	88
<b>第六章 实际应用</b>	<b>91</b>
6.1 理论物理学 . . . . .	91
6.2 天体物理学 . . . . .	91
6.3 空间科学 . . . . .	91
6.4 地质学 . . . . .	92
6.5 地球物理学 . . . . .	92
6.6 流体力学 . . . . .	93
6.7 热学 . . . . .	93
6.8 理论化学 . . . . .	94
6.9 化学信息学 . . . . .	94
6.10 材料学 . . . . .	95
6.11 水文学 . . . . .	96
6.12 气候学 . . . . .	100
6.13 气象学 . . . . .	100
6.14 环境学 . . . . .	102
6.15 生态学 . . . . .	103
6.16 动物形态学 . . . . .	105
6.17 植物学 . . . . .	105

6.18 农学 . . . . .	105
6.19 神经病学 . . . . .	106
6.20 认知神经学 . . . . .	107
6.21 运动神经学 . . . . .	109
6.22 计算神经学 . . . . .	109
6.23 心理学 . . . . .	110
6.24 系统生物学 . . . . .	110
6.25 生物信息学 . . . . .	110
6.26 临床诊断学 . . . . .	112
6.27 老年医学 . . . . .	114
6.28 精神病学 . . . . .	115
6.29 法医学 . . . . .	116
6.30 药学 . . . . .	116
6.31 公共卫生学 . . . . .	116
6.32 经济学 . . . . .	117
6.33 管理学 . . . . .	118
6.34 社会学 . . . . .	119
6.35 教育学 . . . . .	120
6.36 计算语言学 . . . . .	120
6.37 新闻传播学 . . . . .	120
6.38 法学 . . . . .	120
6.39 政治学 . . . . .	121
6.40 军事学 . . . . .	121
6.41 情报学 . . . . .	121
6.42 能源电力 . . . . .	122
6.43 纺织工程 . . . . .	126
6.44 食品工程 . . . . .	126
6.45 土木建筑 . . . . .	127
6.46 交通运输 . . . . .	128
6.47 机械制造 . . . . .	129
6.48 可靠性工程 . . . . .	131
6.49 石油工程 . . . . .	132
6.50 矿业工程 . . . . .	132
6.51 冶金工程 . . . . .	133
6.52 化学工程 . . . . .	133
6.53 医学工程 . . . . .	135
6.54 航空航天 . . . . .	135
6.55 兵器工程 . . . . .	137

6.56 车辆工程 . . . . .	137
6.57 控制工程 . . . . .	138
6.58 电子工程 . . . . .	138
6.59 通信工程 . . . . .	138
6.60 高性能计算 . . . . .	139
6.61 信息安全 . . . . .	139
6.62 测绘遥感 . . . . .	140
6.63 海洋工程 . . . . .	141
6.64 金融工程 . . . . .	141

<b>附录 A 软件实现</b>	<b>147</b>
------------------	------------

<b>参考文献</b>	<b>149</b>
-------------	------------

# 引言

统计独立性是数理统计学领域的基础性概念，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。其他学者也提出了经典的统计独立性度量，如 Spearman 的  $\rho$  [2] 和 Kendall 的  $\tau$  [3] 等。

上个世纪，在对依赖性的研究中 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 [4, 5]。根据 Sklar 定理 [6]，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。基于 Copula 函数的关联关系表示，学界给出了一些关联关系度量的 Copula 函数表示形式定义，如 Spearman 的  $\rho$  和 Kendall 的  $\tau$  等，但这些度量都是二变量的，多变量关联关系度量的 Copula 表示形式一直是学界关心的本领域基本问题，有待进一步的研究和解决。

信息论是关于信息的度量和处理的数学理论 [7]，熵和互信息是该理论的两个核心概念 [8]。其中，熵度量了随机变量的信息量，而互信息度量了两个随机变量之间的信息量。作为一个非线性相关性度量，互信息被认为包含了二变量之间相关关系的全部信息。

2008 年，马健和孙增圻定义了 Copula 熵 (Copula Entropy: CE) 的数学概念 [9, 10]。CE 概念由 Copula 密度函数定义而来，本质上是一种香农熵的形式。我们也证明了它与互信息概念是等价的。同时，我们也给出了基于秩统计量的非参数 CE 估计方法。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，那么我们认为这两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。通过 CE 的定义，我们就在 Copula 理论和信息论之间建立了一座桥梁。

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，具有很多优美的性质，包括连续性、对称性、可加性、非正性、单调变换不变性、以及在高斯变量下与相关系数等价等。

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 [11]，用来学习统计变量之间的关联关系结构。最近，我们又先后将其应用到关联发现 [12]、变量选择 [13]、因果发现 [14]、域自适应 [15]、正态性检验 [16]、时延估计 [17]、系统辨识 [18]、双样本检验 [19]、变点检测 [20] 和对称性检验 [21] 等问题上，都取得了良好的应用效果。

CE 是一个理想的统计独立性度量工具，同时它又可以用来表示和度量另一个基本的统计学概念——条件独立性 (Conditional Independence: CI)。传递熵 (Transfer Entropy: TE) [22] 是一种模型无关的因果关系度量，其本质是条件互信息——一种信息论的 CI 度量。我们证明了 TE/条件互信息可以仅由 CE 来进行表示，并在此 CE 数学表示的基础上，给出了一种基于 CE 估计方法的非参数 TE 估计方法 [14]。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。

我们也因此提出了一个基于 CE 理论的统计学方法论体系，包括独立性检验、条件独立性检验、多元正态性检验、双样本检验、变点检测和对称性检验等方法。在统计学领域已有针对这些问题的大量同类方法，而 CE 方法由于具有坚实的理论基础和非参数的估计算法，较已有方法具有理论上的优势。为了评估 CE 方法的实际性能，我们调研了以上 6 大方法的同类方法，设计了仿真评测对比实验 [23]，实验验证了 CE 方法论体系性能上的优越性。

作为一个基础性的数学概念，研究者从数学上对 CE 概念进行了理论推广，提出了 Tsallis CE [24]、生存 CE [25]、Copula 外熵 [26] 和累积 CE [27]、累积 Copula Tsallis 熵 [28]、Copula Rényi 熵 [29]、Copula Rényi 散度和 Copula Tsallis 散度 [30]、Copula Jeffreys 散度和 Copula Hellinger 散度 [31]、以及 Copula 分形不准确度 [32] 等新概念，通过对多种重要的熵概念进行 Copula 函数扩展定义得到，从而形成了以 CE 为核心的概念体系，丰富和扩展了信息论的熵概念体系 [33, 34]。

作为一种基础性的数据分析工具，CE 被提出以来，在各个不同学科领域都得到了实际应用，包括理论物理学 [35]、天体物理学 [36]、空间科学 [37]、地质学 [38]、地球物理学 [39–41]、流体力学 [42]、热学 [43–45]、理论化学 [46]、化学信息学 [47]、材料学 [48–51]、水文学 [31, 52–84]、气候学 [85, 86]、气象学 [14, 87–91]、环境学 [92–96]、生态学 [97–99]、动物形态学 [100, 101]、植物学 [102, 103]、农学 [104–106]、神经病学 [107–109]、认知神经学 [110–120]、运动神经学 [121–126]、计算神经学 [127–130]、心理学 [131]、系统生物学 [132, 133]、生物信息学 [134–144]、临床诊断学 [13, 25, 145–151]、老年医学 [152–155]、精神病学 [156–158]、法医学 [159]、药学 [160]、公共卫生学 [101, 145]、经济学 [161–167]、管理学 [168–172]、社会学 [15]、教育学 [173]、计算语言学 [174]、新闻传播学 [175]、法学 [176]、政治学 [177]、军事学 [178]、情报学 [179]，以及能源电力 [17, 180–204]、纺织工程 [205]、食品工程 [206, 207]、土木建筑 [208–213]、交通运输 [214–220]、机械制造 [221–229]、可靠性工程 [230–233]、石油工程 [234, 235]、矿业工程 [236, 237]、冶金工程 [238, 239]、化学工程 [240–247]、医学工程 [248, 249]、航空航天 [250–257]、兵器工程 [258]、车辆工程 [259, 260]、控制工程 [261]、电子工程 [262]、通信工程 [263–265]、高性能计算 [266]、信息安全 [267–270]、测绘遥感 [271–273]、海洋工程 [274] 和金融工程 [275–298] 等。在这些应用中，CE 被用来分析和度量不同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 不仅为各种应用提供了理论支撑和方法工具，同时也改进了计算的可靠性和效率。

在实际应用中，研究者同时将 CE 理论与其他理论方法相结合，提出了各种新的方法，简单列举如下。在经典信息论框架下的新方法包括：

- 基于 CE 理论的互信息估计方法，如 GCMI 方法 [111]、半参数互信息估计 [299]、非对称互信息估计 [101]、 $CE^2$  [267]、贝叶斯相关性度量 [300] 和尾部相关性度量 [289] 等；

- 基于 CE 理论的信息流分解 [120, 134];
- 基于 CE 概念对最大熵准则 [301] 进行扩展得到的最大 CE 准则 [24, 205, 302] 和最大 Tsallis CE 准则 [24] 等。

将 CE 理论与图论相结合的方法:

- 图结构之间互信息相似度估计 [100]。

将 CE 理论与信息瓶颈 (Information Bottleneck) 理论 [303] 相结合的方法:

- 信息瓶颈计算 [47]。

将 CE 理论与部分信息分解 (Partial Information Decomposition) 理论 [304] 相结合的方法包括:

- 独特信息 (Unique Information) 估计 [129]、协同 (Synergy) 估计 [130] 和  $\Omega$  信息估计 [305] 等。

Copula 理论的新方法包括:

- Copula 参数估计 [78] 和藤 Copula 结构选择 [282, 284, 285] 等。

基于 CE 理论的因果分析方法包括:

- 因果压缩 [134]、因果结构学习 [164, 206, 207]、LiNGAM-MMI [306] 和时序因果网络构建 [307, 308] 等。

基于 CE 扩展的传统机器学习方法包括:

- 聚类算法 [80, 86]、非线性主元分析 [241] 和决策树构建 [161, 162] 等。

将 CE 与神经网络相结合的方法包括:

- 图神经网络构建 [57, 151]、信息增强生成式对抗补全网络 (IEGAIN) [246] 和神经网络结构剪枝算法 [309] 等。

将 CE 与控制理论相结合的方法包括:

- 模型预测控制器设计 [261] 和网络控制系统滑模控制器设计 [198] 等。

基于 CE 的现代优化算法包括:

- 基于 CE 的进化算法 [177]、基于 CE 的多维藤分布估计算法 [310] 和基于 CE 的灰狼优化算法 [140] 等。

将 CE 与函数逼近理论相结合的方法:

- B 样条函数逼近方法 [311]。

CE 作为一种基础性的理论方法，给出了一种处理相关性和因果性的普适性基本数学概念工具，可以与其他理论和方法相结合，为更多新方法论的派生提供了可能。

本书内容安排如下：

第一章首先介绍 CE 理论的背景（Copula 理论和信息论），然后介绍 CE 理论的基本框架，给出 CE 基本概念定义、与互信息等价性的定理及其推论，以及条件互信息的 CE 表示定理，再总结 CE 的 9 条性质。最后，本章也将介绍基于经验 Copula 函数的非参数 CE 估计方法，以及由此得到的条件互信息估计方法。

第二章介绍作者在 CE 理论应用上的一系列研究，利用 CE 解决了统计学 11 个基本问题，包括变量选择和因果发现等经典问题，从而形成了系统的方法论体系。特别是，我们将给出针对 6 种基本假设检验问题（独立性检验、条件独立性检验、多元正态性检验、双样本检验、变点检测和对称性检验等）的基于 CE 理论统一的方法论体系框架。

第三章讨论 CE 理论应用之间的内在联系，探讨这些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相关的相关性度量框架进行对比，指出本理论框架在多个方面的理论优越性。

第四章通过仿真实验系统评估基于 CE 的 6 种数理统计学基本假设检验问题的方法论，调研这些方法论的同类方法的 R 语言软件实现，实验验证 CE 方法相对于各自同类方法的实际性能优越性。

第五章介绍学界对 CE 概念的数学推广，包括基于 CE 对两种广泛研究的 Rényi 熵和 Tsallis 熵的扩展，以及对信息论中散度和不准确度概念的扩展，从而形成了以 CE 为核心的熵概念体系。

第六章简要介绍 CE 在各个学科领域的实际应用。

附录A给出了 CE 方法的官方和第三方软件实现，包括 R、Python、Julia、Matlab 和 C++ 等 5 种本领域主要的编程语言，方便读者在实际中参考使用。

# 第一章 Copula 焉

## 1.1 背景

### 1.1.1 Copula 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 [4, 5]。此理论定义一类表示随机变量之间依赖关系的函数，称为 Copula 函数，定义如下：

**定义 1** (Copula 函数). 给定  $n$  维随机变量  $\mathbf{X} = (X_1, \dots, X_n)$ 。令  $\mathbf{u}$  表示  $\mathbf{X}$  的边缘分布函数  $u_i = F_i(x_i), i = 1, \dots, n$ 。则  $\mathbf{X}$  对应  $n$  维 Copula 函数  $C : I^n \rightarrow I, I = [0, 1]$  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.  $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$ .

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  $c(\mathbf{u})$ ，即  $c(\mathbf{u}) = \frac{\partial C(\mathbf{u})}{\partial u_1 \cdots \partial u_n}$ 。

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

**定理 1** (Sklar 定理). [6] 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：

$$F(\mathbf{x}) = C(F_1(x_1), \dots, F_n(x_n)). \quad (1.1)$$

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 (1.1) 两边求导，就得到相应的 Sklar 定理的概率密度函数版本：

$$p(\mathbf{x}) = c(\mathbf{u}) \prod_i p(x_i), \quad (1.2)$$

其中  $p(\cdot)$  表示联合或边缘概率密度函数。

### 1.1.2 信息论

信息论是关于信息的度量和处理的数学理论 [7]。熵 (Entropy) 是信息论的核心概念，用于衡量随机变量的信息量，其定义如下：

**定义 2** (香农熵). [8] 给定随机变量  $X \in R^n$  及其概率密度函数  $p(\mathbf{x})$ ，则香农熵的定义为

$$H(\mathbf{x}) = - \int_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}. \quad (1.3)$$

互信息 (Mutual Information: MI) 是信息论中的另一个核心概念，用于衡量两个随机变量之间的相关性，可以理解为度量了一个变量中包含另一个变量的信息量，其定义如下：

**定义 3** (互信息). [8] 给定一对随机变量  $(X, Y)$ ，及其联合概率密度函数  $p(x, y)$  和边缘密度函数  $p(x), p(y)$ ，则  $X$  和  $Y$  之间的互信息定义为

$$I(x; y) = \int_x \int_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy. \quad (1.4)$$

由 MI 的定义可以得到如下定理：

**定理 2.**  $MI$  等于边缘熵与联合熵的差，即

$$I(x; y) = H(x) + H(y) - H(x, y). \quad (1.5)$$

MI 被证明是非负的，定理如下：

**定理 3.** [8] 给定随机变量  $(X, Y)$ ，则有

$$I(x; y) \geq 0, \quad (1.6)$$

当且仅当  $X, Y$  相互独立时，等号成立。

通过将 MI 定义中的概率密度替换成条件概率密度，我们可以定义条件互信息 (Conditional Mutual Information: CMI)，用于度量给定条件下随机变量之间的互信息，定义如下：

**定义 4** (条件互信息). 给定随机变量  $(X, Y, Z)$ ，则  $(X, Y)$  在给定  $Z$  的条件下的条件互信息定义为

$$I(x; y|z) = \int_x \int_y \int_z p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} dx dy dz. \quad (1.7)$$

## 1.2 理论

通过将香农熵定义中多元随机变量的概率密度函数替换成 Copula 密度函数，我们定义如下 Copula 熵概念：

**定义 5** (Copula 熵). [9, 10] 给定多元随机变量  $\mathbf{X}$ , 及其边缘分布  $\mathbf{u}$  和 Copula 密度函数  $c(\mathbf{u})$ , 则 Copula 熵定义为

$$H_c(\mathbf{x}) = - \int_{\mathbf{u}} c(\mathbf{u}) \log c(\mathbf{u}) d\mathbf{u}. \quad (1.8)$$

CE 是一种特殊类型的香农熵, 度量了 Copula 函数中的全部信息量, 从而给出了一个多变量独立性度量的理论工具。

在信息论中, MI 和熵是两个定义不同的概念 [8]。在文献 [9] 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 4.** 多元随机变量的 MI 等价于其负的 CE, 即

$$I(\mathbf{x}) = -H_c(\mathbf{x}). \quad (1.9)$$

证明.

$$I(\mathbf{x}) = \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\prod_i p(x_i)} d\mathbf{x} \quad (1.10)$$

$$= \int_{\mathbf{x}} c(\mathbf{u}) \prod_i p(x_i) \log c(\mathbf{u}) d\mathbf{x} \quad (1.11)$$

$$= \int_{\mathbf{u}} c(\mathbf{u}) \log c(\mathbf{u}) d\mathbf{u} \quad (1.12)$$

$$= -H_c(\mathbf{x}) \quad (1.13)$$

□

由定理2和定理4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.** 多元随机变量的联合熵等于边缘熵和 CE 的和, 即

$$H(\mathbf{x}) = \sum_i H(x_i) + H_c(\mathbf{x}). \quad (1.14)$$

CMI 度量了给定条件下随机变量之间的 MI, 我们证明了其可以表示成 CE 的形式 [14], 定理如下:

**定理 5.** [14] 给定随机变量  $(X, Y, Z)$ , 其对应的 CMI 可以表示成如下 CE 的形式:

$$I(x; y|z) = H_c(x, z) + H_c(y, z) - H_c(x, y, z). \quad (1.15)$$

证明.

$$I(x; y|z) = \int_x \int_y \int_z p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} dx dy dz \quad (1.16)$$

$$= \int_x \int_y \int_z p(x, y, z) \log \frac{p(x, y, z)p(z)}{p(x, z)p(y, z)} dx dy dz \quad (1.17)$$

$$= I(x, y, z) - I(x, z) - I(y, z) \quad (1.18)$$

$$= H_c(x, z) + H_c(y, z) - H_c(x, y, z). \quad (1.19)$$

□

由 CE 的定义5得到的定理4和推论1, 以及定理5, 构成了一个完整的理论框架, 加深了我们对这些信息论基本概念及其之间关系的理解, 并在 Copula 理论和信息论之间架起了一座桥梁。

### 1.3 性质

**多变量** 香农的 MI 定义针对的是二变量情况, 但 CE 概念定义不限于二变量的情况, 适用所有多变量情况, 且多变量之间具有对称性, 扩展了 MI 的定义和适用范围。根据 Sklar 定理, 任意多元随机变量都存在对应的 Copula 函数, 因此也就存在相应的 CE。

**性质 1** (多变量). 对任意  $n$  维随机变量  $\mathbf{X} \in R^n, n > 1$ , 都存在  $H_c(\mathbf{x})$ .

**熵性** 由于 Copula 函数是定义在单位体上的概率分布函数, 且是一致连续的 [4], 因此由 Copula 函数定义的 CE 本质上是一种特殊的香农熵, 也就具有香农熵的连续性、对称性和可加性等公理性质 [312, 313]。

**性质 2** (连续性). 给定多元随机变量  $\mathbf{X}$ , 若  $\mathbf{x}_n \rightarrow \mathbf{x}$ , 则有

$$\lim_{\mathbf{x}_n \rightarrow \mathbf{x}} H_c(\mathbf{x}_n) = H_c(\mathbf{x}). \quad (1.20)$$

**性质 3** (对称性). 给定随机变量  $X$  和  $Y$ , 则有

$$H_c(x, y) = H_c(y, x). \quad (1.21)$$

**性质 4** (可加性). 给定多个统计独立的多元随机变量  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , 则有

$$H_c(\mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n H_c(\mathbf{x}_i). \quad (1.22)$$

**非正性** 由定理3得知, MI 已经被证明是非负的。因此, 根据定理4所述 MI 和 CE 的等价性, CE 就是非正的, 仅当多变量之间是相互独立时, CE 等于 0。

**性质 5 (非正性).** 给定任意多元随机变量  $\mathbf{X}$ , 则有

$$H_c(\mathbf{x}) \leq 0, \quad (1.23)$$

当且仅当  $\mathbf{X}$  相互独立时, 等号成立。

这一性质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。一般地讲, 熵度量了随机变量的不确定性, 是非负的; 而 CE 则是非正的, 因为它度量了由于变量间相关性导致减少的不确定性。

**单调变换不变性** 由于 Copula 函数具有单调变换不变性 [4], 因此基于 Copula 函数定义的 CE 天然地继承了这一不变性特性。

**性质 6 (单调变换不变性).** 给定随机变量  $(X, Y)$ , 以及分别定义其上的单调递增函数  $f$  和  $g$ , 则有

$$H_c(x, y) = H_c(f(x), g(y)). \quad (1.24)$$

**边缘函数无关** 根据 Sklar 定理, 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形式(1.14): 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。

**性质 7 (边缘函数无关).** 多元随机变量  $\mathbf{X} \in R^n$  的  $H_c(\mathbf{x})$  只由其 Copula 密度函数  $c(\mathbf{u})$  唯一决定, 与其边缘函数  $p_i(x_i), i = 1, \dots, n$  无关。

**全阶次** 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的度量。作为多元随机变量的泛函, 联合熵度量了所有阶次的随机变量关系信息, 根据其边缘熵和 CE 的分解形式(1.14)可知, 边缘熵对应于单个随机变量的各阶次信息, 而 CE 对应于所有阶次的相关关系信息。

**性质 8 (全阶次).**  $H_c(\mathbf{x})$  度量了多元随机变量所有阶次的相关关系。

**高斯分布下与相关系数等价** 相关系数是统计学传统的二阶相关性统计量。高斯分布只包含随机变量的二阶相关关系, 且这些关系由相关系数唯一决定。很容易证明, 在高斯分布的情况下, 相关系数与 CE 具有数学上的某种等价关系, 即 CE 可以由二阶相关系数矩阵来表示。此等价关系是 CE 度量全阶次相关关系的特殊情况。

**性质 9 (高斯分布下与相关系数等价).** 给定满足高斯分布的多元随机变量  $\mathbf{X} \sim N(\mu, \Sigma)$ , 则有

$$H_c(\mathbf{x}) = \frac{1}{2} \log |\Sigma|. \quad (1.25)$$

## 1.4 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<sup>1</sup> [9]。该方法仅需如下两步：

1. 估计经验 Copula 密度函数；
2. 由经验 Copula 密度函数估计 CE。

给定多元随机变量  $\mathbf{X} \in R^n$  的一组独立同分布样本  $\{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ ，可以很容易地通过秩 (rank) 统计量来完成第 1 步经验 Copula 密度函数的估计，如下

$$F_i(x_i) = \frac{1}{T} \sum_{t=1}^T \mathbf{1}(x_t^i \leq x_i), \quad (1.26)$$

其中  $i = 1, \dots, n$ ,  $\mathbf{1}(\cdot)$  表示示性函数。

在得到经验 Copula 密度函数估计后，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 [315] 提出的 k 近邻熵估计法来估计 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（秩统计量和 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秩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秩统计量的熵，内涵深刻。

根据定理5，CMI 可以表示成式(1.15)的 CE 形式。因此，我们可以很容易地得到基于 CE 估计的 CMI 估计方法。

---

<sup>1</sup>本方法已经实现为 R 和 Python 语言的 `copent` 算法包 [314]，并已分别在CRAN和PyPI上发布共享。

## 第二章 理论应用

### 2.1 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。在统计和机器学习领域，通常利用图模型（Graphical Models）[316, 317] 将这种关联结构表示成图（Graph）的形式，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。图模型是多变量概率分布的一种近似表示。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。

从数据中学习图模型结构的问题，被称为结构学习（Structure Learning），是本领域的基本问题之一 [318–320]。结构学习的算法很多，其中比较著名的有 Chow-Liu 的树结构学习方法 [321]。该方法通过最小生成树（Minimal-Spanning-Tree: MST）算法最小化树结构对应的互信息之和，来学习得到随机变量概率分布的树结构近似。给定一个多元随机变量  $\mathbf{X} = (X_1, \dots, X_m)$  对应的树结构  $T$ ，则联合概率分布函数可以表示为如下乘积形式：

$$p(\mathbf{x}|T) = \prod_{i=1}^m p(x_i|x_{\pi_i}), \quad (2.1)$$

其中  $\pi_i$  表示  $x_i$  的父节点集合。Chow-Liu 算法通过最小化如下 MI 之和来学习得到近似结构  $T$ ：

$$\min \sum_{i=1}^m I(x_i, x_{\pi_i}). \quad (2.2)$$

图结构表示与 Copula 函数表示之间具有对应关系。图结构近似表示了变量之间的依赖关系；根据 Copula 理论，Copula 函数中包含了全部的变量之间关系信息，而且这种信息是与变量无关的。式(2.1)可以表示成如下 Copula 函数形式：

$$p(\mathbf{x}|T) = \prod_{i=1}^m c_i(x_i, x_{\pi_i}) \prod_{i=1}^m p(x_i), \quad (2.3)$$

式中第一项为  $X$  的 Copula 函数的近似表示形式，给出了  $T$  结构包含的依赖关系信息，第二项边缘函数则是与  $T$  无关的。因此，我们可以利用 Copula 函数和图结构的对应关系，通过 Copula 函数来解决结构学习问题。

马健 [11] 提出了一种基于 Copula 函数的图结构学习算法框架，首先从数据中估计出经验 Copula 函数，再进行结构学习。根据式(2.3)，利用 MI 和负 CE 之间的等价性，我们可以将 Chow-Liu 算法最小化 MI 之和的优化目标公式(2.2)转化成如下最大化 CE 之和的形式：

$$\max \sum_{i=1}^m H_c(x_i, x_{\pi_i}). \quad (2.4)$$

基于此优化目标，马健给出了 Chow-Liu 算法的 CE 版本 [11]，包含两步：

1. 利用 CE 估计方法学习得到随机变量的负 CE 关联矩阵；
2. 再利用 MST 生成算法从上述矩阵得到树结构。

此算法符合 Copula 结构学习算法框架：在 CE 估计算法中，需要首先估计经验 Copula 函数，再利用经验 Copula 函数估计得到 CE 关联矩阵，进而得到 MST 结构。由于非参数 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。

藤结构（Vine Copula）是另一种基于 Copula 函数理论框架的概率分布近似图结构形式方法 [322, 323]。给定一个多元随机变量  $\mathbf{X} = (X_1, \dots, X_m)$  及其联合概率密度函数  $p(\cdot)$ ，和一个由一组树构成的藤结构  $V = (T_1, \dots, T_{m-1})$ ，其中  $T_k = (\mathbf{V}_k, \mathbf{E}_k)$ ,  $k = 1, \dots, m-1$ ，则联合概率密度可以由如下藤 Copula 结构近似表示 [324]：

$$p(\mathbf{x}|V) = \prod_{k=1}^{m-1} \prod_{e \in E_k} c_{u_e, v_e | \pi_e}(u_e, v_e | \mathbf{x}_{\pi_e}) \prod_{i=1}^m p(x_i), \quad (2.5)$$

其中  $\pi_e$  和  $u_e, v_e$  分别表示边  $e$  的条件变量集合和两个顶点变量的边缘函数。由式(2.5)可知，这种方法基于藤结构近似概率分布和 Copula 函数，结构中的每一条边对应一个二元条件 Copula 函数。基于这种近似方式的 Vine Copula 结构估计一般通过偏相关直接完成，或通过利用偏相关估计藤结构对应的最小化 MI 分布完成 [325]。但基于偏相关的藤结构估计需要满足一定的简化假设 [326]，已知满足简化假设的模型只有高斯分布和学生氏 t 分布，因此限制了这种近似方法的实际应用 [327]。而上述 Chow-Liu 树结构学习算法是基于 CE 和 MI 的等价关系，利用非参数的经验 Copula 函数估计 MI 最小化对应的近似分布函数，因此理论限制要更小得多，而树结构近似也更加通用和灵活。

马健 [11] 设计仿真实验验证了所提出算法的有效性，又将该算法应用到两个经典的 UCI 机器学习数据集 [328]：鲍鱼生长数据集和波士顿房价数据集。实验结果（见图2.1）显示，该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鲍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。<sup>1</sup>

## 2.2 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重要方法之一 [329]。

---

<sup>1</sup>实验代码：<https://github.com/majianthu/dse>

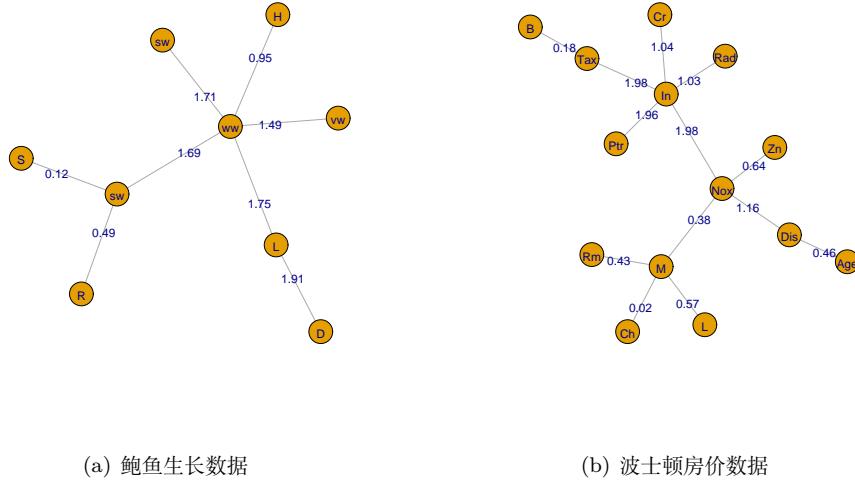


图 2.1: 结构学习算法实际数据实验分析结果.

Pearson 相关系数 [1] 是一种统计学史上重要的相关性度量概念，应用广泛。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，它只适用于线性相关关系的情况，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（对比见表 2.1）。它没有线性和高斯性的假设，且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 作为相关性度量的优点：

- 无模型假设，
- 可处理非线性关系，
- 统计独立性度量，
- 单调变换不变性，
- 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。

除了 Pearson 相关系数，还有两个常见的非参数相关系数：Spearman 的  $\rho$  [2] 和 Kendall 的  $\tau$  [3]。与 CE 通过 Copula 函数定义类似，这两个非参数相关系数可以由 Copula 函数来表示。

表 2.1: Pearson 相关系数与 CE 的对比.

性质	Pearson 相关系数	CE
相关变量个数	二变量	多变量
变量之间关系	线性	非线性
相关性阶次	二阶	全阶次
隐含假设	高斯分布	无
度量类型	相关性	独立性

**定理 6.** [4] 给定连续随机变量  $(X, Y)$  和相应的 Copula 函数  $C(u, v)$ , Spearman 的  $\rho$  的 Copula 函数表示形式为

$$\rho_{X,Y} = 12 \int_u \int_v C(u, v) dudv - 3. \quad (2.6)$$

**定理 7.** [4] 给定连续随机变量  $(X, Y)$  和相应的 Copula 函数  $C(u, v)$ , Kendall 的  $\tau$  的 Copula 函数表示形式为

$$\tau_{X,Y} = 4 \int_u \int_v C(u, v) dC(u, v) - 1. \quad (2.7)$$

与 Pearson 相关系数相同, 这两个非参数相关系数也只能衡量二变量的线性相关关系, 不同之处在于二者没有高斯分布假设。Barbe 等 [330] 和 Joe [331] 分别提出了一个 Kendall 的  $\tau$  的多变量版本, 而 Spearman 的  $\rho$  则存在两个多变量版本定义 [4, 332, 333]。给定一个多元随机变量  $\mathbf{X} = (X_1, \dots, X_n)$  和其相应的 Copula 函数  $C(\mathbf{u})$ , Joe 的多变量版本 Kendall 的  $\tau$  的定义为 [331]

$$\tau_J = \frac{1}{2^{n-1} - 1} \left\{ 2^n \int_{\mathbf{u}} C(\mathbf{u}) dC(\mathbf{u}) - 1 \right\}. \quad (2.8)$$

两个 Spearman 的  $\rho$  的多变量版本是基于 Copula 函数的表示形式, 给定一个多元随机变量  $\mathbf{X} = (X_1, \dots, X_n)$  和其相应的 Copula 函数  $C(\mathbf{u})$ , Wolff 的定义为 [332]

$$\rho_W = \frac{n+1}{2^n - (n+1)} \left\{ 2^n \int_{\mathbf{u}} C(\mathbf{u}) d\mathbf{u} - 1 \right\}, \quad (2.9)$$

而 Joe [333] 和 Nelson [4] 的定义为

$$\rho_{JN} = \frac{n+1}{2^n - (n+1)} \left\{ 2^n \int_{\mathbf{u}} u_1 \cdots u_n dC(\mathbf{u}) - 1 \right\}. \quad (2.10)$$

另一个同类的相关系数是 Gini 的  $\gamma$  [334], 其定义也是基于秩统计量, 也可以表示成 Copula 函数形式 [4]。

**定理 8.** [4] 给定连续随机变量  $(X, Y)$  和相应的 Copula 函数  $C(u, v)$ , Gini 的  $\gamma$  的 Copula 函数表示形式为

$$\gamma_{X,Y} = 2 \int_u \int_v (|u + v - 1| - |u - v|) dC(u, v). \quad (2.11)$$

Behboodian 等 [335] 给出了一个多变量版本的 Gini 的  $\gamma$ , 给定一个多元随机变量  $\mathbf{X} = (X_1, \dots, X_n)$  和其相应的 Copula 函数  $C(\mathbf{u})$ , 其定义为

$$\gamma_B = \frac{1}{b(n) - a(n)} \left\{ \int_{\mathbf{u}} (A(\mathbf{u}) + \bar{A}(\mathbf{u})) dC(\mathbf{u}) - a(n) \right\}, \quad (2.12)$$

其中  $A(\mathbf{u}) = \{M(\mathbf{u}) + W(\mathbf{u})\}/2$ ,  $\bar{A}$  为  $A$  相应的生存函数,  $M, W$  为 Fréchet 上、下界函数,  $b(n), a(n)$  为归一化因子。

Schweizer 和 Wolff [336] 定义了一种基于 Copula 函数的  $L_1$  范数的新度量  $\sigma$  (见定义6)。

**定义 6.** 给定连续随机变量  $(X, Y)$  和相应的 Copula 函数  $C(u, v)$ , Schweizer 和 Wolff 的  $\sigma$  定义为

$$\sigma_{X,Y} = 12 \int_u \int_v |C(u, v) - uv| du dv. \quad (2.13)$$

此度量可以被扩展到多变量的情况, 给定一个多元随机变量  $\mathbf{X} = (X_1, \dots, X_n)$  和其相应的 Copula 函数  $C(\mathbf{u})$ , Schweizer 和 Wolff 的  $\sigma$  多变量版本为 [332]

$$\sigma_W = \frac{2^n(n+1)}{2^n - (n+1)} \int_{\mathbf{u}} |C(\mathbf{u}) - \prod_{i=1}^n u_i| d\mathbf{u}. \quad (2.14)$$

Rényi [337] 曾经提出过著名的相关性度量的公理系统, 给出了一个理想的相关性度量  $\delta$  应该具备的公理性质, 包括如下七条:

1.  $\delta$  定义在任何成对的连续随机变量  $(X, Y)$  上;
2.  $\delta_{X,Y} = \delta_{Y,X}$ ;
3.  $0 \leq \delta_{X,Y} \leq 1$ ;
4. 若  $X, Y$  相互独立, 则  $\delta_{X,Y} = 0$ ;
5. 若  $X, Y$  之间存在严格的单调函数关系, 则  $\delta_{X,Y} = 1$ ;
6. 若  $f, g$  分别是  $X, Y$  上的严格单调函数, 则  $\delta_{X,Y} = \delta_{f(X),g(Y)}$ ;
7. 若随机变量  $(X, Y)$  符合正态分布, 则  $\delta_{X,Y} = |\rho_{X,Y}|$ ,  $\rho_{X,Y}$  为相关系数。

Schweizer 和 Wolff [336] 对 Rényi 公理进行了修正, 并证明了他们提出的度量  $\sigma$  满足修正后的公理性质。Schmid 等 [338] 总结讨论了以上几种关联度量满足的 Rényi 公理性质的情况。

通过与第1.3节所述的 CE 性质对比可以看出, CE 仍然满足 Rényi 的这七条性质中的第 2、4 和 6 条, 不同之处在于 1) 突破了第 1 条二变量的限制, 定义在任意多变量上; 2) 不再满足第 3 和 5 条, 而是满足式(1.23), 当且仅当随机变量独立时等于 0; 3) 不再满足第 7 条, 而是变成当随机变量为高斯分布时, 满足式(1.25)的等价关系。同时, CE 还具有可加性等其他度量不具备的性质。因此, CE 是一个理论完善的关联关系度量。由上述可见, 已有关联度量的多变量扩

展定义多样，不是自然且唯一的。与之相比，CE 是一个基于 Copula 函数严格定义的多变量关联关系度量，自然且唯一，可以衡量任意多变量的非线性的相关关系。

多元随机变量之间的关联度量是比多变量关联度量更宽泛、更一般的问题。关于这方面的研究已经有很多，如传统相关系数扩展、距离相关法或 Kernel 函数法等，读者可以参考综述论文 [339] 里的介绍。而基于 CE 概念，多元随机变量之间关联度量可很方便地通过如下定理得到：

**定理 9.** 给定一组多元随机变量  $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ ，其中  $\mathbf{X}_1 \in R^{d_1}, \dots, \mathbf{X}_n \in R^{d_n}, d_1, \dots, d_n \geq 1$ ，则基于 CE 的多元随机变量之间关联度量为

$$H_c(\mathbf{x}_1; \dots; \mathbf{x}_n) = H_c(\mathbf{x}) - \sum_{i=1}^n H_c(\mathbf{x}_i). \quad (2.15)$$

证明.

$$H_c(\mathbf{x}_1; \dots; \mathbf{x}_n) = -I(\mathbf{x}_1; \dots; \mathbf{x}_n) \quad (2.16)$$

$$= - \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{p(\mathbf{x}_1) \cdots p(\mathbf{x}_n)} d\mathbf{x} \quad (2.17)$$

$$= -I(\mathbf{x}) + \sum_{i=1}^n I(\mathbf{x}_i) \quad (2.18)$$

$$= H_c(\mathbf{x}) - \sum_{i=1}^n H_c(\mathbf{x}_i). \quad (2.19)$$

□

当  $d_i = 1$  时，则  $H_c(\mathbf{x}_i) = 0$ . 这一定理可以直观地理解为多向量关联等于从所有变量关联的信息量中减去每个随机向量内部的关联信息量，也就得到了随机变量之间的关联信息量，是一种自然且易于理解的多向量关联度量。与同类度量相比，基于 CE 的多向量关联度量既简洁又自然，且很容易通过非参数 CE 估计方法来计算。

综上所述，我们就基于 CE 概念给出了适用于任意情况的关联度量及其估计方法，包括多变量关联度量和多向量关联度量。这两种度量自然且唯一，具有完美关联度量的公理性质，且易于计算，与同类度量方法相比具有显著理论优势和实际优越性。

鉴于 CE 具有完美的理论性质，马健 [12] 提出利用其来发现关联关系，以充分发挥 CE 的理论优势和非线性关联发现上的优异性能。我们将上述六种关联度量应用于著名的 NHANES 医学体检数据 [340]，以计算数据中各个实验室体检化验结果变量之间的内在关联关系<sup>2</sup>，实验结果（见图2.2）表明，CE 得到关联关系矩阵能够更清晰地反映体检化验变量之间的非线性关系，能够进行生物医学意义上的解释 [12]，说明了 CE 相对于上述五种典型关联度量的显著优越性。更多关于独立性度量方法的仿真对比评测，请见第4.2节。

---

<sup>2</sup>实验代码：<https://github.com/majianthu/nhanes>



图 2.2: 基于 NHANES 数据的 6 种关联度量对比实验结果.

## 2.3 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 [341,342]。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），同时降低模型的复杂度。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或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 [343] 和 BIC [344] 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 [345]、岭回归 (Ridge Regression) [346] 和弹性网络 (Elastic Net) [347] 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，通常是模型无关的。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。其它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) [348,349] 和距离相关 (Distance Correlation: DC) [350,351] 等。

马健 [13] 提出了一种基于 CE 的变量选择方法，以取代 LASSO 等传统方法。该方法利用 CE 度量自变量和目标变量之间的关联强度，根据 CE 绝对值从大到小依次选择变量。在变量选择问题上，CE 被真实数据实验证明优于以下主流变量选择方法 [13]：

- LASSO / Ridge Regression / Elastic Net [345–347],
- AIC / BIC [343,344],
- Adaptive LASSO [352],
- Hilbert-Schmidt Independence Criterion (HSIC) [348,349],
- Distance Correlation [350,351],
- Heller-Heller-Gorfine Tests of Independence [353],
- Hoeffding's D test [354],
- Bergsma-Dassios T\* sign covariance [355],
- Ball correlation [356].

实验<sup>3</sup>采用了著名的 UCI 心脏病数据集 [328]，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，用来研究如何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

---

<sup>3</sup>实验代码：<https://github.com/majianthu/aps2020>

一个参照标准。实验结果表明，与其他方法相比，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2.3。

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：

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

该方法做变量选择是模型无关的，这是与基于似然函数的方法相比，方法无需考虑模型及其复杂度等因素，具有明显的普适性优势。作为一种关联度量工具，CE 与其他度量工具相比定义更坚实，具有很多理想的独立性度量公理属性，因此也就具有了明显的理论优势。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 的估计方法基于秩统计量，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。

生存分析 (Survival Analysis) [357] 是一类特殊的回归问题，其预测目标是事件发生时间 (time-to-event)，也即是未来某一事件发生所需要的时间。这类问题的特殊性还在于一种删失 (Censoring) 机制，用于当某一事件在观察期未发生时的处理。生存分析在医学、可靠性和社会科学等领域具有广泛的应用。建立生存分析模型也需要进行变量选择，用于筛选与事件发生时间相关的变量 [358,359]。马健 [146] 提出将基于 CE 的变量选择方法应用于此类问题，通过计算变量与事件发生时间之间的 CE 来选择变量。他将方法应用于两个公开的肺癌数据，与常用的随机生存森林 (Random Survival Forest) 和 Lasso-Cox 两种典型的生存分析变量选择方法进行了对比，发现该方法能够在保证模型可解释性的同时获得更好的预测性能，验证了方法的优越性<sup>4</sup>。

## 2.4 因果发现

因果关系普遍存在于自然界当中，发现因果关系是科学哲学的核心命题 [360–362]，也是各门学科的主要课题之一 [363]。从一组随机变量的观测数据中发现变量之间的因果关系，被称为因果发现 (Causal Discovery)，是统计学的经典问题 [364–367]。因果发现方法一般由以因果关系判据为基础的因果关系结构搜索算法构成，如典型的 PC 算法 [368]。因果发现方法在不同学科领域都有重要的应用价值。

---

<sup>4</sup>实验代码：<https://github.com/majianthu/survival>



(a) CE



(b) dCor



(c) dHSIC

图 2.3: 三种统计独立性度量选择的变量.

如何判别因果关系是因果发现问题解决的基础性问题，因果关系度量则是因果发现算法的核心组成部分。控制论学者维纳提出了一种因果关系判定的哲学准则，表述为因变量必须能够改善对果变量预测的能力 [369]。在此维纳准则的基础上，格兰杰提出了著名的格兰杰因果关系 (Granger Causality: GC) 检验 [370,371]。GC 的定义非常简单：

**定义 7** (Granger Causality). 给定时序随机变量  $X_t, Y_t$ , 当利用预测函数  $f_p$  预测  $Y_{t+1}$  时, 如果将  $X_t$  加入自变量集合, 预测误差的方差函数  $var(\cdot)$  变小, 即:

$$var[y_{t+1} - f_p(y_{t+1}|y_t)] > var[y_{t+1} - f_p(y_{t+1}|y_t, x_t)], \quad (2.20)$$

则认为  $X$  和  $Y$  之间存在格兰杰意义上的因果关系,  $X$  是  $Y$  的因变量。

GC 检验是经典的因果关系判别工具, 但不难从定义7中得知, 它只适用于线性和高斯假设的情况 [372], 因而限制了其广泛应用。

Schreiber [22] 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, 如下

**定义 8** (Transfer Entropy). 给定时序变量  $X_t, Y_t, t = 1, \dots, T$ , 由  $X$  到  $Y$  的 TE 定义为

$$TE_{X \rightarrow Y} = \sum_t p(y_{t+1}, y_t, x_t) \log \frac{p(y_{t+1}|y_t, x_t)}{p(y_{t+1}|y_t)}. \quad (2.21)$$

TE 是一种因果关系度量, 本质上是信息论的条件互信息 (Conditional Mutual Information: CMI), 也是在检验和度量条件独立性 (Conditional Independence) 关系。定义式(2.21)可写成如下 CMI 的形式:

$$TE_{X \rightarrow Y} = I(y_{t+1}; x_t|y_t). \quad (2.22)$$

TE 可以认为是 GC 的非线性推广, 在高斯变量条件下与 GC 等价 [373]。TE 度量了因变量向果变量传递的有助于消除预测不确定性的新信息量, 是一个模型无关的信息论度量, 适用于任何情况的因果关系检验, 较之 GC 等带有高斯模型假设的传统因果关系推断方法更科学合理, 因而具有普适性。

CE 是统计独立性度量, 而 TE 是条件独立性度量。通过并不复杂的数学变换, 马健 [14] 证明了 TE 可以表示为只包含 CE 的数学形式。这一数学表示形式为从 CE 估计 TE 提供了理论基础。

**定理 10.** 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:

$$TE_{X \rightarrow Y} = H_c(y_{t+1}, y_t) + H_c(y_t, x_t) - H_c(y_{t+1}, y_t, x_t). \quad (2.23)$$

证明.

$$TE_{X \rightarrow Y} = \sum_t p(y_{t+1}, y_t, x_t) \log \frac{p(y_{t+1}|y_t, x_t)}{p(y_{t+1}|y_t)} \quad (2.24)$$

$$= \sum_t p(y_{t+1}, y_t, x_t) \log \frac{p(y_{t+1}, y_t, x_t)p(y_t)}{p(y_{t+1}, y_t)p(y_t, x_t)} \quad (2.25)$$

$$= I(y_{t+1}, y_t, x_t) - I(y_{t+1}, y_t) - I(y_t, x_t) \quad (2.26)$$

$$= -H_c(y_{t+1}, y_t, x_t) + H_c(y_{t+1}, y_t) + H_c(y_t, x_t). \quad (2.27)$$

□

因为 TE 本质上是条件独立性关系  $Y_{t+1} \perp\!\!\!\perp X_t|Y_t$  的度量，因此定理10的证明本质上与定理5完全相同，是给出了一种时序条件独立性度量的 CE 表示。

在过去的研中，因果关系的估计往往是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。马健 [14] 基于以上 TE 的 CE 表示形式，利用非参数的 CE 估计算法，提出了一种简单优雅、易于理解和实现的非参数 TE 估计方法。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<sup>5</sup>:

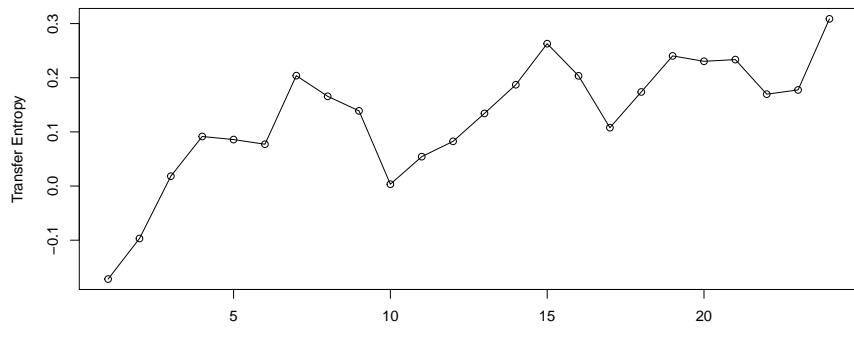
1. 利用非参数 CE 估计方法，估计式(2.23)中的 3 个 CE 子项；
2. 由 3 个 CE 估计值计算得到 TE。

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<sup>6</sup>。实验采用了 UCI 机器学习数据集仓库中的北京 PM2.5 数据 [374]，包含了北京地区 2010-2014 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 1-24 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。

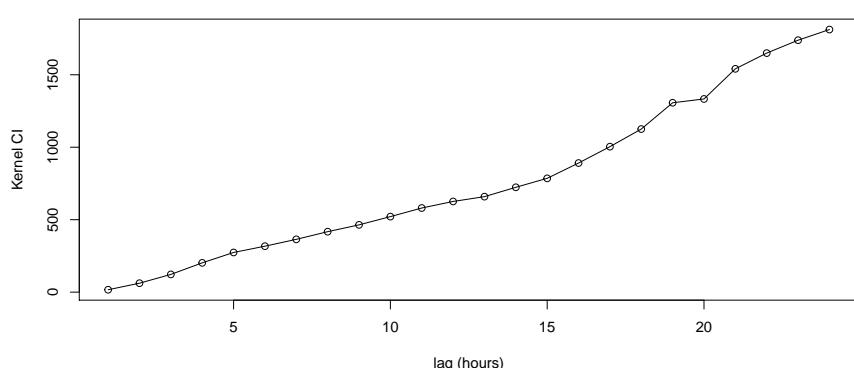
同样在上述实验数据的基础上，我们将提出的 TE 估计方法与另外两种条件独立性度量进行了对比实验，估计从气象因素到 PM2.5 的因果关系 24 小时走势。这两种度量分别是基于核函数的条件独立性度量 (Kernel-based Conditional Independence: KCI) [375] 和条件距离相关 (Conditional Distance Correlation: CDC) [376]。论文通过将用 CE 估计 TE 与其它两种方法进行了对比，结果（见图2.4）显示 TE 的估计效果更好。更多关于条件独立性度量方法的仿真对比评测，请见第4.3节。

<sup>5</sup>此方法已在 R 和 Python 语言的 copent 包 [314] 中实现。

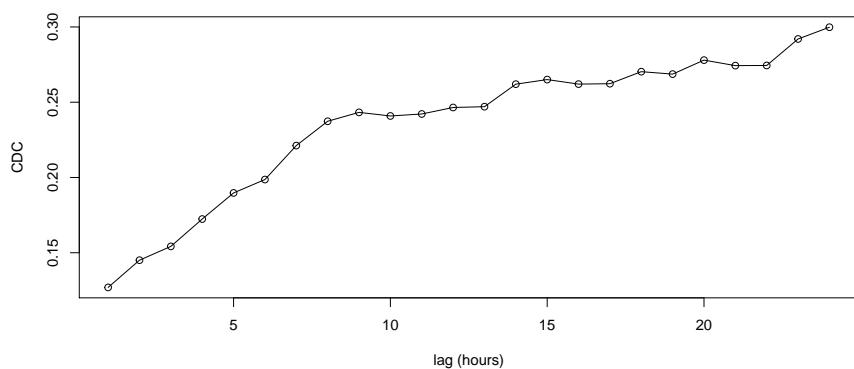
<sup>6</sup>实验代码: <https://github.com/majianthu/transferentropy>



(a) TE



(b) KCI



(c) CDC

图 2.4: 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.

## 2.5 系统辨识

微分方程是描述动态系统的主要数学工具，在不同学科具有广泛的应用。从数据中学习微分方程是动态系统领域的一个重要问题，也称系统辨识或方程发现 [377–379]，近年来得到了大量的研究 [380, 381]。

方程发现问题通常可以被当作一个回归问题来对待，即从数据学习一组从系统状态到状态微分的回归方程。给定一个一般的动态系统微分方程形式，如下：

$$\frac{dx_i}{dt} = f_i(\mathbf{x}, t), \quad (2.28)$$

其中  $x_i, i = 1, \dots, n$  表示系统状态变量，则方程发现问题就是从数据辨识  $f_i$ 。从数据辨识  $f_i$  需要确定该方程包含的未知自变量，一旦自变量确定则方程的对应关系就知道了，这是典型的变量选择问题。很多经典回归模型方法被应用到此问题，如基于稀疏性的方法（如 SINDy [382]）和核函数方法 [383] 等。

熵概念在微分方程领域的已有研究涉及热力学系统的平衡态和可逆性等问题 [384]。而在动态系统领域，熵被用来表述确定性混沌系统表现出的“随机性”，其中最著名的就是在系统状态空间上定义的 Kolmogorov-Sinai 熵 [385]。Nardone 和 Sonnino [386] 提出了一种描述时间序列复杂度的差分熵（Entropy of Difference）。目前也有一些学者将 MI 方法应用到系统辨识领域，如 Chernyshov 和 Jharko [387] 提出了一种利用 Tsallis MI 来度量动态系统输入和输出之间的相关性的系统辨识方法，又比如 Stoer Vogel 和 van Schuppen [388] 提出通过最小化系统估计误差和白噪声之间的 MI 率来辨识线性动态系统。

马健 [18] 提出了一种基于 CE 的微分方程发现方法，将问题理解为变量选择问题，基于系统状态变量和状态差分变量之间关系的分析，利用第2.3节所述的基于 CE 的变量选择方法解决了此方程发现问题。该方法利用了动态系统中的混沌特性，利用熵概念度量状态变量及其差分之间表现出来的随机性关系，解决系统辨识问题。该方法包含了两个主要步骤：

1. 利用差分算子近似计算状态变量的微分；
2. 计算状态微分和状态变量之间的 CE，根据 CE 来选择方程的变量。

该方法中的差分算子可以由以下非参数方式计算得到：

$$\left. \frac{dx}{dt} \right|_{t=t_0} \approx \frac{x_{t_1} - x_{t_0}}{t_1 - t_0}. \quad (2.29)$$

而 CE 可以由非参数估计方法得到。因此，所提出的方法是非参数的，不做任何假设，适用于任何动态系统的辨识。

作者将方法应用于两个经典的 3 维动态系统：Lorenz 系统 [389] 和 Rössler 系统 [390, 391]，系统中都包含了由一阶和二阶的状态变量组成的 3 个方程，分别如下式所示：

**Lorenz 系统**

$$\begin{aligned}\frac{dx}{dt} &= \sigma(y - x), \\ \frac{dy}{dt} &= \rho x - y - xz, \\ \frac{dz}{dt} &= -\beta z + xy,\end{aligned}\tag{2.30}$$

其中  $\sigma, \rho, \beta$  分别为系统的普朗特数、瑞利数和几何因子参数，在仿真实验中分别设为 10, 28, 8/3。

**Rössler 系统**

$$\begin{aligned}\frac{dx}{dt} &= -(y + z), \\ \frac{dy}{dt} &= x + ay, \\ \frac{dz}{dt} &= b + z(x - c),\end{aligned}\tag{2.31}$$

其中  $a, b, c$  为系统参数，在仿真实验中分别设为 0.38, 0.2, 5.7。

我们生成了这两个系统的仿真数据，应用该方法从仿真数据中辨识出的状态变量和状态微分变量之间的关系由估计出的 CE 值所反映，数值大说明存在函数关系，从实验结果（见图2.5和图2.6）可以看出，辨识结果基本与系统方程相符合，证明了该方法的有效性<sup>7</sup>。

## 2.6 时延估计

时延 (Time Lag) 是一种动态系统中普遍存在的特性参数，指一个变量作用于另一个变量需要的时间。由于物质、能量或信息的传输时间，时延存在于所有物理、社会和生物系统中的因果效应发生的时间先后关系上。因此，时延参数估计是时间序列分析领域重要的理论问题 [392, 393]，在科学和工程等诸多领域具有广泛的应用价值，比如可以用来分析交通系统中的拥堵传播、太阳活动对地球系统的影响、政策效应分析等问题。

时延估计 (Time Lag Estimation) 是指从一对系统测量信号中，估计出源信号到目标信号的作用发生时间。给定源信号  $x_t$  和目标信号  $y_t$ ,  $t = 1, \dots, T$ ，假设它们之间存在的系统函数关系为  $f(\cdot)$ ，包含了  $x$  到  $y$  的时延  $l$ ，则时延估计的一般模型为

$$y_t = f_L(x_{t-l}, \mathbf{z}_t) + \epsilon, \tag{2.32}$$

其中  $\mathbf{z}_t$  为系统中其他变量， $\epsilon$  为系统噪声。时延估计问题的目标就是要从观测数据  $x_t, y_t$  估计时延参数  $l$ 。这往往需要首先对  $f_L$  作出一定的假设，通常假设系统为线性，但实际系统则通常为更加复杂的非线性系统。

传统的时延估计的主要方法是基于时序变量的自相关系数 [394]，但其具有线性假设，因此作用范围十分有限。另一种主要方法是时延互信息 (Time-delayed MI) [395]，可以去除线性假设的限制，适用于具有非线性特性的时序变量。但这两种方法本质上都是对称关系的度量，而系统时延由于是因果关系的属性，因此是非对称的关系。

<sup>7</sup> 实验代码：<https://github.com/majianthu/sysid>



(a) 仿真数据



(b) 实验结果

图 2.5: 基于 Lorenz 系统的系统辨识仿真实验结果.



(a) 仿真数据



(b) 实验结果

图 2.6: 基于 Rössler 系统的系统辨识仿真实验结果.

时延参数是从源信号到目标信号的因果过程的作用时间，因此可以利用因果关系度量工具来进行估计。TE 作为一种非对称的因果关系度量，量化了从因变量到果变量作用关系的信息量，因而更适用于估计因果时延特性。给定上述时延模型公式(2.32)，可以首先估计一定时间窗口内不同时延取值  $l = 1, \dots, L$  下  $X_{t-l}$  到  $Y_t$  的 TE 值  $TE_{X \rightarrow Y}(l)$ ，再取最大 TE 值对应的时延作为时延参数的估计值  $\hat{l}$ ，表示如下：

$$\hat{l} = \arg \max_l TE_{X \rightarrow Y}(l). \quad (2.33)$$

动态系统的时延参数估计可以通过上述 TE 方法加以解决，但传统的 TE 估计问题被认为十分困难，往往需要对时延系统函数  $f_L$  作出一定假设才能完成 TE 估计，因而难以普遍应用于所有系统情况。马健 [17] 提出利用第2.4节所述的基于 CE 的非参数 TE 估计方法来解决时延估计问题，方法包括两个步骤：

1. 利用基于 CE 的 TE 估计方法，估计时延窗口内的因变量到果变量的一组 TE 值；
2. 再将 TE 的最大值对应的时延作为时延参数的值。

由于基于 CE 的 TE 估计器是非参数的，使得该方法不对动态系统做任何假设，不仅适用于线性系统，也适用于非线性系统，具有普适性。

作者仿真了具有不同动态特性的线性和非线性时延动态系统以验证方法的有效性，仿真的五个系统的状态方程分别如下：

### 随机输入系统

$$\begin{aligned} x_i &= \xi_1, \\ y_{i+l} &= x_i + \xi_2; \end{aligned} \quad (2.34)$$

### 非线性输入系统

$$\begin{aligned} x_i &= \sin(2\pi i/m) + \xi_1, \\ y_{i+l} &= x_i + \xi_2; \end{aligned} \quad (2.35)$$

### 一阶维纳过程系统

$$\begin{aligned} x_i &= x_{i-1} + \xi_1, \\ y_{i+l} &= x_i + \xi_2; \end{aligned} \quad (2.36)$$

### 二阶维纳过程系统

$$\begin{aligned} x_i &= \alpha x_{i-1} + \beta x_{i-l} + \xi_1, \\ y_i &= x_i + \xi_2; \end{aligned} \quad (2.37)$$

### 二阶非线性维纳过程系统

$$\begin{aligned} x_i &= \alpha x_{i-1} + \beta x_{i-l} + \xi_1, \\ y_i &= x_i^2 + \sin(x_i) + \xi_2. \end{aligned} \quad (2.38)$$

其中,  $x_i$  表示系统状态变量,  $y_i$  表示系统输出变量,  $\xi_1 \sim N(0, 0.1), \xi_2 \sim N(0, 0.001)$  表示高斯噪声, 系统变量系数  $\alpha = 0.2, \beta = 0.8, l = 1, 2, 3, 4$  表示仿真实验中的时延参数。仿真实验的仿真轨迹数据如图2.7所示。将该时延估计方法应用于仿真数据, 以估计系统状态  $X$  到系统输出  $Y$  的时延参数, 估计结果如图2.8所示, 可以发现该方法准确地从每种系统的仿真数据中估计出了四种时延情况的时延参数值。无论是线性系统和非线性系统, 该方法都给出了准确的估计结果, 特别是在两个二阶维纳过程系统中, 该方法不仅通过 TE 最大值估计出了时延参数, 而且 TE 估计值还反映了时延因果效应以时延参数为周期的周期性特征和随时延增加而减弱的衰减性特征, 证明了基于非参数 TE 估计的方法对各种系统的适用性和合理性。

作者又将方法应用于摩洛哥缔头万 (Tétouan) 城的电力负荷数据, 分析五种天气因素对该城三个区域电力负荷影响的时延特征, 发现了不同天气因素对负荷产生影响的时延长度, 以及影响的每日变化特征<sup>8</sup>。

## 2.7 域自适应

域自适应 (Domain Adaptation: DA) 是一类常见的机器学习问题范式, 是指由于外部因素的变化, 导致训练模型的数据分布与应用部署模型时的数据分布产生了不同, 需要让训练的模型适应分布偏移的情况。常见的 DA 问题解决思路就是将源问题域上学习的知识迁移到目标问题域上, 根据方法的不同, 可以分为示例迁移、特征迁移、参数迁移和关系知识迁移等几类 [396]。

DA 问题具有重要的现实意义和价值。比如, 将在一个医院采集的数据上训练好的模型应用到其他医院时, 可能由于数据采集设备的不同导致采集的数据发生分布偏移, 从而导致模型性能下降。同样的情况也会发生在其他领域 (如社会学) 的问题中, 比如由于人群的社会属性的不同, 由一个人群研究得到的模型结论在另一个人群上就会发生模型偏差。

基于因果关系视角来解决 DA 问题是迁移学习领域一个重要的研究方向, 它将源域和目标域的数据分布迁移视为外部因素导致的结果, 通过学习外部因素到问题变量的因果关系来构建模型。基于 CE 的条件独立性度量作为一种基本的因果关系发现工具, 可以用于解决 DA 问题。基于此, 马健 [15] 提出了一种从因果角度解决 DA 问题的方法。他假设自变量  $X$  到预测变量  $Y$  在不同域  $D_i$  上的关系是不变的, 将数据分布迁移视为一个由外在条件变量  $Z$  在  $D_i$  上作用不同导致的结果, 这样 DA 问题就转化为学习自变量  $X$ 、预测变量  $Y$  和外在变量  $Z$  之间统计关系的问题, 需要发现  $X, Y$  之间不变的依赖关系, 二者的依赖关系以外在变量为条件, 即判断是否如下条件独立性关系是否成立:

$$X \not\perp\!\!\!\perp Y | Z. \quad (2.39)$$

这时, 利用第2.4节所述的基于 CE 的条件独立性测试就能发现域迁移条件  $Z$  背后的  $X$  和  $Y$  之间不变的因果关系, 从而很好地解决了 DA 问题。

---

<sup>8</sup>实验代码: <https://github.com/majianthu/timelag>



图 2.7: 时延估计仿真的仿真变量轨迹 ( $l = 4$ ).



图 2.8: 时延估计仿真实验的算法估计结果.

作者设计了仿真实验验证了方法的有效性，并将方法成功应用于社会学的男女收入不平等的社会原因分析问题<sup>9</sup>。

## 2.8 正态性检验

正态分布（亦称高斯分布）是一类非常重要的概率分布函数，由于中心极限定理，其在概率论中居于基础性中心地位，且在自然和社会现象中普遍存在 [397,398]。正态性是很多统计模型和方法的假设条件，因此在应用中检验样本分布的正态性假设十分必要。正态性检验（Normality Test）是一类检验样本分布正态性的基本假设检验方法 [399,400]，分为单变量和多变量两类，这里我们关注多元正态性检验问题。

给定一个多元随机变量  $\mathbf{X}$ ，多元正态性检验的零假设为

$$H_0 : \mathbf{X} \sim N(\mu, \Sigma); \quad (2.40)$$

对立假设为

$$H_1 : \mathbf{X} \not\sim N(\mu, \Sigma), \quad (2.41)$$

其中  $N(\mu, \Sigma)$  表示正态分布函数。传统的正态性检验方法很多，比如基于矩、特征函数、熵或最优传输等概念工具的方法等 [401–405]。

CE 作为衡量变量间全阶次相关关系的度量工具，也可以用于检验二阶相关特性的正态性检验问题。我们都知道，根据最大熵原理 [301]，在一、二阶统计量相同的情况下，在所有分布中正态分布的熵最大。由于正态分布完全由其一阶和二阶统计量决定，因此正态分布的 CE 中包含的是二阶相关性对应的信息。可以很容易推导，在正态分布的条件下，CE 与二阶统计量具有某种等价关系，如式(1.25)所示。而在非正态分布中，相关关系不仅是二阶的，也有高阶的，因而 CE 包含的信息除了二阶相关关系对应的信息外，也有高阶相关关系对应的信息，且非高斯性越强，CE 中高阶相关的信息越多。同时，由于熵是全阶次的信息量，因此 CE 度量了包括二阶相关信息在内的全部阶次相关性信息。

马健 [16] 利用高斯分布的 CE 与二阶统计量之间等价关系提出了一种多变量正态性检验 (Multivariate Normality Test: MVNT) 的方法，通过计算分布与同方差的高斯分布在 CE 统计量的差值来衡量联合分布的正态性。给定多元随机变量的样本  $\mathbf{X}_T$ ，则由此得到的正态性检验的统计量定义为

$$T_{mvn}(\mathbf{X}_T) = H_c(\mathbf{x}) - H_c(\mathbf{x}_G), \quad (2.42)$$

其中， $\mathbf{X}_G$  是与  $\mathbf{X}$  具有相同方差的高斯分布多元随机变量。易知，当多元随机变量  $\mathbf{X}$  的分布为高斯分布时  $T_{mvn} = 0$ ；且多元分布的非高斯性越强， $T_{mvn}$  的绝对值越大。

他同时给出了此统计量的估计方法<sup>10</sup>，包括了十分简单的两部分：式(2.42)的第一项可以由 CE 的非参数估计方法得到；第二项即是式(1.25)的值，可先估计协方差矩阵  $\Sigma$ ，再解析计算得到。

---

<sup>9</sup>实验代码：<https://github.com/majianthu/cda>

<sup>10</sup>此方法已在 R 和 Python 语言的 `copent` 包 [314] 中实现。

作者设计了两组仿真实验，仿真了两类非高斯性的情况，并将此检验方法与 5 种经典的同类方法进行了对比，证明了此检验方法的有效性和对传统 5 种经典方法的优越性<sup>11</sup>。更多关于多元正态性检验方法的仿真对比评测，请见第4.4节。

## 2.9 双样本检验

双样本检验（Two-Sample Test）是统计学中另一类重要的假设检验方法，用于测试两组样本是否来自同一个分布函数 [406–409]。很多统计学的理论方法可以转化成双样本检验问题，如对称性测试就可以转化成检验对称变换的样本是否同分布的问题，又如变点检测（Change Point Detection）其实就是寻找一组双样本检测中样本间差异最大的点。同时，双样本检验又具有广泛的应用价值，比如可以检测临床治疗、政策实施等人为干预前后目标变量是否发生了变化等。

传统的双样本检验方法很多，如双样本 T 检验 [410]，Kolmogorov–Smirnov 检验 [411] 和基于核函数的检验 [412] 等。但这些方法都有各自的不足之处，比如 T 检验需要正态分布假设，Kolmogorov–Smirnov 检验只能作用于单变量情况，而核函数方法需要超参数的调节等。

马健 [19] 提出了一种基于 CE 的双样本检验方法<sup>12</sup>，思想是基于样本与检验标注之间的相关性程度来定义检验统计量。给定两组样本  $\mathbf{X}_0 = \{X_{01}, \dots, X_{0m}\} \sim P_0$  和  $\mathbf{X}_1 = \{X_{11}, \dots, X_{1n}\} \sim P_1$ ，该检验的零假设为

$$H_0 : P_0 = P_1, \quad (2.43)$$

对立假设为

$$H_1 : P_0 \neq P_1. \quad (2.44)$$

定义联合样本集  $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$ ，以及零假设和对立假设分别对应的标注变量  $Y_0 = (1_1, \dots, 1_{m+n})$  和  $Y_1 = (1_1, \dots, 1_m, 2_1, \dots, 2_n)$ ，则零假设和对立假设对应的样本与标注变量之间的 CE 分别为

$$H_0 : H_c(\mathbf{x}; y_0) = H_c(\mathbf{x}, y_0) - H_c(\mathbf{x}), \quad (2.45)$$

$$H_1 : H_c(\mathbf{x}; y_1) = H_c(\mathbf{x}, y_1) - H_c(\mathbf{x}), \quad (2.46)$$

用于度量两样本分别与  $H_0$  和  $H_1$  的符合程度。则该双样本检验方法的检验统计量可通过零假设和对立假设相应的 CE 的差来定义：

$$T_{tst}(\mathbf{X}_0, \mathbf{X}_1) = H_c(\mathbf{x}, y_0) - H_c(\mathbf{x}, y_1). \quad (2.47)$$

易知，当  $H_0$  为真时，则  $T_{tst}$  较小；而当  $H_1$  为真时，则  $T_{tst}$  较大。

作者给出了基于 CE 非参数估计的统计量  $T_{tst}$  估计方法。因此，该检验方法是多变量非参数检验，且估计算法无需调参。由于  $T_{tst}$  是基于 CE 的信息论度量，因此是问题无关的。

作者在 3 组由正态分布和正态 Copula 仿真的数据上验证该方法的有效性，并将方法与基于 MI、核函数和 dCor 的三种多变量非参数检验方法进行了对比，发现该方法有效检验了仿真实验

<sup>11</sup>实验代码：<https://github.com/majianthu/mvnt>

<sup>12</sup>此方法已在 R 和 Python 语言的 copent 包 [314] 中实现。

中的双样本假设，与同类方法相比具有同等或更好的检验性能<sup>13</sup>。更多关于双样本检验方法的仿真对比评测，请见第4.5节。

## 2.10 变点检测

变点检测（Change Point Detection）[413, 414] 是统计学中一个典型的时间序列分析任务，是指在一个时间序列中检测发生的系统状态突变。自此问题在上个世纪 50 年代被提出以来 [415, 416]，学界对其开展了长期的研究，目前已经给出了大量的检测算法 [417, 418]。根据不同角度区分，变点检测算法可以是离线检测或在线检测，单点检测或多点检测，检测对象可以是单变量数据或多变量数据。变点检测的应用领域十分广泛，可以用于检测自然系统、生命系统、社会系统、或工业系统中发生的各种类型突变。

变点检测问题可以转化为双样本检验问题加以解决，即在时间序列的每个时间点上，对该点前后的数据做双样本检验测试，由此得到的检验统计量最大值对应的时间点即可认为是发生了状态改变的变点。给定一个时间序列  $\mathbf{X}_t, t = 1, \dots, T$ ，判断是否存在这样一个变点  $t_c$ ，其前后时间序列  $\mathbf{X}_b, \mathbf{X}_a$  的特性发生了变化，基于双样本检验的变点检测就是要验证如下假设是否成立，若如下零假设成立：

$$H_0 : P_b(\mathbf{x}_b) = P_a(\mathbf{x}_a), \quad (2.48)$$

则不存在变点；若如下对立假设成立：

$$H_1 : P_b(\mathbf{x}_b) \neq P_a(\mathbf{x}_a), \quad (2.49)$$

则变点存在，其中  $P_b, P_a$  分别为变点前后时间序列变量的概率分布函数。

马健 [20] 根据这一原理，利用第2.9节提到的基于 CE 的双样本检验，提出了一种非参数多变量的单变点检测方法。该方法通过在  $X_t$  上的每个时间点上进行前述基于 CE 的双样本检验，得到一组检验统计量  $T_{tst}(t), t = 1, \dots, T$ ，设  $t_c$  为统计量最大值的所在位置，即

$$t_c = \arg \max_t T_{tst}(\mathbf{X}_b, \mathbf{X}_a; t). \quad (2.50)$$

若  $T_{tst}(t_c)$  大于某一设定的阀值，则认为  $H_1$  成立， $t_c$  为变点；否则认为  $H_0$  成立，不存在变点。

他又进一步结合上述单变点检验方法和二分割策略提出了一种多变点检测方法<sup>14</sup>，该方法包括以下步骤：

1. 将整个时间序列加入待检测时间序列队列；
2. 对下一个待检测时间序列进行单变点检测，若检测的检验统计量大于预设阀值，则认为存在一变点；否则认为不存在变点；
3. 若变点存在，则将该检测序列中变点前后的两个子序列分别加入待检测时间序列队列；

---

<sup>13</sup>实验代码：<https://github.com/majianthu/tst>

<sup>14</sup>此单变点和多变点检测方法均已在 R 和 Python 语言的 `copent` 包 [314] 中实现。



图 2.9: 在尼罗河年径流变化数据上的单变点检测实验结果。

4. 继续对待检测序列进行以上 2、3 步检测，直至队列中所有待检测序列检测完毕。

该方法采用预设阀值来判断某一段序列上是否存在变点，从而能够自动估计变点的个数。由于统计量  $T_{tst}$  是基于 CE 定义的信息论度量，与具体问题无关，从而使得该方法的统计量阀值可以做到统一设定、普遍适用，无需针对具体问题的阀值参数调优环节。

作者在一组仿真数据上验证了该方法，并与传统经典方法进行了对比，证明了该方法的有效性和优越性。更多关于变点检测方法的仿真对比评测，请见第4.6节。

作者又在两个典型的变点检测实际测试数据（尼罗河年径流数据和英国煤矿矿难数据）上验证了该方法<sup>15</sup>。尼罗河数据记录了 1871-1970 年尼罗河在阿斯旺的年径流量数据，由于 1898 年阿斯旺水坝的修建，导致了年径流量随后明显变小。英国煤矿矿难数据则记录了英国 1851-1962 年每年发生的严重煤矿矿难数，一般认为由于 1887 年煤矿监管立法，年矿难数随着法律的实施而明显减少。实验结果（见图2.9和图2.10）显示，该方法分别成功检测到了尼罗河数据中年径流变化中的突变点（1898 年）和英国年矿难数变化中的突变点（1892 年），验证了该方法的有效性。

<sup>15</sup> 实验代码：<https://github.com/majianthu/cpd>



图 2.10: 在英国矿难数据上的单变点检测实验结果.

## 2.11 对称性检验

对称性是自然科学的根本准则之一，在理论物理中发挥着重要作用 [419, 420]，而数学的对称性则定义为群变换下的不变性 [421]。对称性也是概率统计学中分布函数的基本属性假设之一 [422, 423]，体现为概率分布函数在对称变换下的不变性。根据变换的不同，分布对称性有四种类型：平稳性、可收缩性、可交换性和可旋转性 [424]。很多统计学模型和方法都基于概率分布函数的某种对称性假设 [425]，如 T 检验 [426] 和 Wilcoxon 符号秩检验 [427, 428]。因此，检验分布的对称性成为了数理统计学假设检验领域的基本问题之一 [429]，目前已经有大量的对称性检验方法被提出 [430–441]。正态性检验就是检验分布的高斯性——一种具有某种旋转变换不变性的分布对称性；双样本检验可用于检验样本变换对应的分布对称性；而变点检测问题本质上就是一种对称性检验——平稳性检验，即平移变换下的分布对称性。

概率密度函数的对称性表现为某种变换不变性。给定来自概率密度函数  $p(x) \in R$  的样本  $X = \{X_i, i = 1, \dots, N\}$ ， $u$  为  $p(x)$  的均值（或中值），则旋转对称性检验就是要从  $X$  判断  $p$  是否关于  $u$  对称，也即是  $p$  关于  $u$  的变换不变性。因此，此对称性检验问题的零假设为

$$H_0 : p(u - x) = p(u + x); \quad (2.51)$$

相应的备择假设为

$$H_1 : p(u - x) \neq p(u + x). \quad (2.52)$$

基于此问题定义，对称性检验就转化为双样本检验问题，通过对  $u$  的对称性变换得到的成对样本进行双样本检验来检验  $p$  对称性的正确与否。若  $u - x$  与  $u + x$  来自同一分布，即  $H_0$  成立，则  $p$  是对称的，否则  $H_1$  成立，则  $p$  就是非对称的。基于此思想，马健 [21] 提出了一个利用基于 CE 的双样本检验进行对称性检验的方法。假设  $\tilde{X}$  是由  $X$  减去其均值估计  $\tilde{u}$  而得到，则问题就变为检验  $\tilde{X}$  和  $-\tilde{X}$  是否来自同一分布。因此，我们定义检验统计量为

$$T_{sym}(X) = T_{tst}(\tilde{X}, -\tilde{X}). \quad (2.53)$$

若  $p$  对称，则  $T_{sym} = 0$ ；否则， $T_{sym} > 0$ ，且  $p$  越不对称， $T_{sym}$  越大。

由此得到的检验方法分为两步：

1. 均值化  $X$  得到  $\tilde{X} = X - \tilde{u}$ ；
2. 根据式(2.53)计算统计量  $T_{sym}$ 。

由于基于 CE 的双样本检验的统计量  $T_{tst}$  可以进行非参数估计，因此  $T_{sym}$  也可以由非参数估计方法得到。

我们设计了仿真实验<sup>16</sup>验证本检验方法的有效性，并将其与同类方法进行了对比，具体请见第4.7节。

本方法是基于 CE 双样本检验测试样本集对称变换不变性的思想，可以很容易地进行扩展。通过将  $p$  扩展为多变量分布，可以得到多变量概率密度函数的对称性检验 [442]；通过将本方法

---

<sup>16</sup>实验代码：<https://github.com/majianthu/symmetry>

中的最基本的镜像对称替换为更复杂的对称变换 [443]，可以得到面向各种分布对称性的对称性检验，如基于置换变换的可交换性检验等。

# 第三章 讨论

## 3.1 理论应用之间的联系

前一章介绍的 CE 的前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。区别在于这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。

基于 CE 概念的独立性检验和条件独立性检验是其他理论应用的基础，通过问题之间内在的上下游关系构成了一个方法论体系，为各种问题的求解提供了一个功能丰富的工具箱。独立性检验可以用来解决建立模型过程中的变量选择问题，这样的模型可以是特定类型的函数，如生存分析函数，也可以是动态过程模型，如我们在系统辨识应用中估计的微分方程。条件独立性检验可以直接用来估计 TE，进而进行时序变量之间的因果分析；利用 TE 方法又可以估计动态系统中的时延参数。基于 CE 概念可以用来解决假设检验问题，包括面向多元正态性的单样本检验和普适的双样本检验，而得到的双样本检验又可以解决时序分析中的变点检测问题以及对称性检验问题。

## 3.2 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 [14] 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 [14]。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3.1），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。

### 3.3 三种理论框架的对比

我们提出了一个基于 CE 概念，能够将独立性和条件独立性两个基本概念相统一的理论框架。与此类似，核函数的方法 [349,375] 和距离相关的方法 [351,376] 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3.1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 DC 也具有类似的等价关系 [350]，HSIC 则未知。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。

三种度量框架都发展出了一套系统的方法论体系 [444,445]，包含了独立性检验、条件独立性检验、正态性检验、双样本检验和变点检测等方法。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独立性检验和条件独立性检验方法。实验结果表明了 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。三个理论框架都包含有正态性检验和双样本检验等假设检验方法，但基于 CE 的方法理论更严格，也因此在仿真数据对比实验上表现出了更优越的检验能力。CE 框架和核函数框架都基于各自的双样本检验方法发展出了多变量非参数变点检测方法，仿真实验表明，前者具有更优越的检测性能。CE 和距离相关框架还包含对称性检验，而核函数框架尚不包含此方法论。

需要指出的是，核函数方法和距离相关方法在独立性检验 [446] 和条件独立性检验 [447] 问题上具有某种等价性。同时，也有学者提出二者的 Copula 版定义 [448–450]。



图 3.1: 对温度到 PM2.5 的 TE 变化的分解.

表 3.1: 三种统计独立性度量框架的对比.

框架	CE	DC	HSIC
定义	基于 Copula 函数	相关性的非线性扩展	基于核函数
多变量	是	[451]	[349]
不变性	单调变换不变	线性变换不变	无
Pearson 相关	高斯性假设下等价	高斯性假设下等价	未知
计算复杂度	$O(n^2)$	$O(n^4)$	$O(n^4)$
独立性检验	[9]	[350, 451]	[348, 349]
条件独立性检验	[14]	[376]	[375]
正态性检验	[16]	[452]	[453]
双样本检验	[19]	[454]	[412]
变点检测	[20]	—	[455]
对称性检验	[21]	[456]	—

# 第四章 仿真评测

## 4.1 概述

我们提出了 CE 理论，利用其解决了统计学领域的几个基本问题，包括独立性/条件独立性检验、多元正态性检验、双样本检验、变点检测和对称性检验等，进而形成了一个系统的方法论体系。同时，本领域内已经存在大量的针对这些问题的同类方法。为了将基于 CE 的方法与这些同类方法进行对比，我们进行了针对不同问题同类方法的评估实验，实验以仿真为主。我们在 R 语言 `copent` 算法包 [314] 中实现了基于 CE 的 6 个方法，并调研整理了 R 语言和 Python 语言实现的每个问题的 CE 同类方法，利用这些方法实现完成了评估对比实验。本部分将给出这些评估实验的设计和结果，以便读者了解基于 CE 的方法相对于其各自同类方法的性能优势。

从理论的角度来看，CE 具有着坚实的数学基础和性能良好的非参数估计方法，从而使其方法论体系体现出了科学性和普适性。在所有仿真对比实验中，基于 CE 的方法都展现出了所有方法中最好的评估结果。基于这些仿真实验结果，作者认为 CE 理论给出了解决这些统计学基本问题最为科学且有效的方法论体系。

## 4.2 独立性检验

独立性是概率统计领域的基本性概念，具有基础性的重要地位。从统计学初期的皮尔逊相关系数开始，如何度量这种统计学的概念就一直是本学科关注的核心问题之一，有大量的度量方法根据不同的思想或原则被提出来 [457]。这其中，就包括前述的 CE 等三种理论框架的方法。那么哪一种方法是最理想的度量呢？理论上，为了回答此问题，Rényi [337] 曾经提出了著名的独立性度量的公理系统，包括了 7 条公理。Schweizer 和 Wolff [336] 在提出他们基于 Copula 的度量时，对 Rényi 的公理系统又做了修正。

如何从实验的角度评估对比这些度量方法是一个重要的问题。马健 [23] 设计了一组仿真实验<sup>1</sup>，对现有的 16 种独立性度量进行了对比（度量方法及实现见表4.1），仿真实验考虑了变量为线性/非线性、高斯性/非高斯性的情况，也考虑了度量方法为二元/多元、单变量相关/多变量相关等情况，对多种不同角度进行了组合，其中二元独立性度量 16 种，多元独立性度量 6 种。实验中对比的独立性度量如下：

---

<sup>1</sup>实验代码：<https://github.com/majianthu/eval>

**Kendall's  $\tau$**  Kendall 的  $\tau$  [3] 是一种广泛采用的二元非参数独立性度量，基于秩统计量定义。给定符号函数  $s$ ，则 Kendall 的  $\tau$  定义为样本对组合的一致性和非一致性的差值，如下：

$$\tau = \frac{1}{n^2} \sum_{i,j=1}^n s(x_i - x_j)s(y_i - y_j), \quad (4.1)$$

其中  $n$  是样本数。

**Hoeffding's D** Hoeffding 的 D [354] 也是一种二元非参数独立性度量。独立性假设检验测试如下等式是否成立：

$$F_{XY} = F_X F_Y, \quad (4.2)$$

其中  $F_{XY}$  和  $F_X, F_Y$  分别是随机变量  $X, Y$  的联合分布和边缘分布。Hoeffding 定义了如下泛函  $D$  作为检验统计量：

$$D(x, y) = \int \{F_{XY}(x, y) - F_X(x)F_Y(y)\}^2 dF_{XY}(x, y). \quad (4.3)$$

**Bergsma-Dassios's  $\tau^*$**  受 Kendall 的  $\tau$  启发，Bergsma 和 Dassios 定义了一个新的二元独立性检验统计量  $\tau^*$  [355]。给定随机变量  $X, Y$ ，该统计量定义在两组样本对上，如下

$$\tau^*(X, Y) = \frac{1}{n^4} \sum_{i,j,k,l=1}^n a(x_i, x_j, x_k, x_l)a(y_i, y_j, y_k, y_l), \quad (4.4)$$

其中符号函数  $a$  定义如下

$$a(z_1, z_2, z_3, z_4) = s(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|). \quad (4.5)$$

**HHG** Heller 等 [353] 提出了一种基于划分的二元独立性检验统计量，定义基于 Hoeffding 的  $D$  的统计量公式(4.3)。他们给出了 Pearson 指数 (HHG.chisq) 和似然比指数 (HHG.lr) 两种统计量定义。HHG 统计量基于秩统计量，因而是分布无关的。

**Ball** Pan 等 [458] 提出了一种独立性度量，称为球方差 (Ball Covariance)。球方差定义为 Borel 概率测度和边缘 Borel 概率测度的平均距离，是 Banach 空间中 Hoeffding 的  $D$  的对应。球方差统计量表示为

$$BCov(x_1, \dots, x_K) = \frac{1}{N^2} \sum_{i,j=1}^N (P_{ij}^X - \prod_{k=1}^K P_{ij}^{x_k})^2, \quad (4.6)$$

其中， $N$  为样本数， $P_{ij}^X, P_{ij}^{x_k}$  分别为联合经验 Borel 概率测度和经验边缘 Borel 概率测度。球方差在变量独立的情况下等于 0。

**BET** Zhang [459] 提出了一种独立性检验的框架，称为 BET (Binary Expansion Testing)，其通过多尺度二次展开来近似二元 Copula 函数。他定义了二次展开上的交叉 OR 值  $\lambda$  来度量二元独立性，类似于列联表的 OR 值。

**QAD** Trutschnig [460] 提出了一种非对称依赖性的度量 (QAD)  $\zeta$ , 定义为 Copula 函数与独立 Copula  $\Pi$  之间的距离

$$\zeta(C, \Pi) = D_1(C, \Pi), \quad (4.7)$$

其中  $D_1$  是通过马尔科夫核函数定义的 Copula 函数之间的距离度量。

**CODEC** 若随机变量  $X, Y$  相互独立, 则有

$$E(Y|X) = E(Y). \quad (4.8)$$

Chatterjee [461] 基于式(4.8)定义了一个简单的相关系数 (CODEC)。给定随机变量  $(X, Y)$ , 当  $X_1 \leq \dots \leq X_n$ ,  $r_i$  为相应的  $Y_i$  的秩, 则 CODEC 统计量定义为

$$\xi_n(X, Y) = 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}. \quad (4.9)$$

**Mixed** Genest 等 [462] 提出了一个基于经验 Copula 的独立性度量, 定义如下

$$G_c(\hat{C}, \Pi) = \int_{I^d} n(\hat{C} - \Pi)^2 d\mathbf{u}, \quad (4.10)$$

其中  $\hat{C}$  为经验 Copula 函数,  $\Pi$  为独立 Copula 函数。

**Subcopula** Erdely [463] 提出了一种基于 subcopula 的独立性度量, 定义为

$$d(S) = \sup\{S - \Pi\} - \sup\{\Pi - S\}, \quad (4.11)$$

其中  $S$  为二元 subcopula 函数。

**dCor** 距离相关 (Distance Correlation: dCor) 是 Székely 等 [350, 351] 提出的一种独立性度量, 是传统线性相关系数的非线性扩展。给定随机变量  $(X, Y)$ , dCor 定义为

$$dCor(X, Y) = \frac{\nu^2(X, Y)}{\sqrt{\nu^2(X)\nu^2(Y)}}, \quad (4.12)$$

其中  $\nu^2(X, Y)$  是距离协方差 (Distance Covariance), 定义如下

$$\nu^2(X, Y; w) = \|f_{X,Y}(t, s) - f_X(t)f_Y(s)\|_w^2, \quad (4.13)$$

$\|\cdot\|_w$  是基于加权 2 范数定义的度量。

**MDC** Shao 和 Zhang [464] 提出了一种 dCor 的变体, 称为 MDC (Martingale Difference Correlation), 其本质是检验式(4.8)是否成立。

**HSIC** 希尔伯特-施密特独立性准则 (HSIC) [348] 是一种广泛使用二元独立性度量, 基于核函数定义在再生核希尔伯特空间 (RKHS) 上。HSIC 也有多个变量定义版本, 称为 dHSIC [349]。

**NNS** Voile 和 Nawrocki [465] 提出了一种基于偏矩 (Partial Moments) 的非线性相关系数, 称为 NNS (Nonlinear Nonparametric Statistic)。

我们设计了 6 组仿真实验生成实验数据, 其中前三组为二元独立性实验, 第四、五组为三元独立性实验, 第六组为基于四元正态分布仿真实验, 仿真两个二元随机变量之间的独立性, 具体实验设置如下:

1. 随机变量  $\mathbf{X} = (X_1, X_2)$  满足二元正态分布  $\mathbf{X} \sim N(\mathbf{u}, \rho)$ , 其中协方差  $\rho$  从 0 以 0.1 步长增加到 0.9;
2. 随机变量  $\mathbf{X} = (X_1, X_2)$  满足二元正态 copula 函数  $C_N(u, v; \rho)$ , 两个固定的边缘函数分别为正态分布  $u \sim N(0, 1)$  和指数分布  $v \sim E(\lambda = 2)$ , copula 函数参数  $\rho$  从 0 以 0.1 步长增加到 0.9;
3. 随机变量  $\mathbf{X} = (X_1, X_2)$  满足二元阿基米德 copula 函数, 包括 Clayton, Frank 和 Gumbel 三种, 定义如下:

$$C_{\alpha}^{Clayton}(u, v) = \max \left( [u^{\alpha} + v^{\alpha} - 1]^{-\frac{1}{\alpha}}, 0 \right), \quad (4.14)$$

$$C_{\alpha}^{Frank}(u, v) = -\frac{1}{\alpha} \ln \left( 1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{e^{-\alpha} - 1} \right), \quad (4.15)$$

$$C_{\alpha}^{Gumbel}(u, v) = \exp \left\{ -[(-\ln u)^{\alpha} + (-\ln v)^{\alpha}]^{\frac{1}{\alpha}} \right\}, \quad (4.16)$$

参数  $\alpha$  从 1 增加到 10, 边缘函数与上同;

4. 随机变量  $\mathbf{X} = (X_1, X_2, X_3)$  满足三元正态分布  $\mathbf{X} \sim N(\mathbf{u}, \Sigma)$ , 其中协方差矩阵  $\Sigma$  为

$$\begin{vmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{vmatrix}, \quad (4.17)$$

$\rho$  从 0 以 0.1 步长增加到 0.9;

5. 随机变量  $\mathbf{X} = (X_1, X_2, X_3)$  满足三元 Gumbel copula 函数, 定义如下:

$$C_{\alpha}^{Gumbel}(u, v, w) = \exp \left\{ -[(-\ln u)^{\alpha} + (-\ln v)^{\alpha} + (-\ln w)^{\alpha}]^{\frac{1}{\alpha}} \right\}, \quad (4.18)$$

参数  $\alpha$  从 1 增加到 10, 一个边缘函数为正态分布  $u \sim N(0, 2)$ , 另两个边缘函数为指数分布  $v \sim E(\lambda = 2), w \sim E(\lambda = 0.5)$ , 边缘函数参数固定不变;

6. 随机变量  $\mathbf{X} = (X_1, X_2, X_3, X_4)$  满足四元正态分布  $\mathbf{X} \sim N(\mathbf{u}, \Sigma)$ , 用于仿真两组变量间强度变化的独立性关系, 4 个变量分为两组, 协方差矩阵  $\Sigma$  为

$$\begin{vmatrix} 1 & \rho_{12} & \rho & \rho \\ \rho_{12} & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho_{34} \\ \rho & \rho & \rho_{34} & 1 \end{vmatrix}, \quad (4.19)$$

表 4.1: 评估的独立性度量方法及其软件实现.

算法包	独立性度量	语言
copent	CE [9]	R
stats	Ktau [3]	R
energy	dCor [350]	R
dHSIC	dHSIC [349]	R
HHG	HHG.chisq, HHG.lr [353]	R
independence	Hoeff [354], BDtau [355]	R
Ball	Ball [356]	R
qad	QAD [460]	R
BET	BET [459]	R
MixedIndTests	Mixed [462]	R
subcopem2D	subcopula [463]	R
EDMeasure	MDM [464]	R
FOCI	CODEC [461]	R
NNS	NNS [465]	R

组内变量间协方差固定不变  $\rho_{12} = 0.8, \rho_{34} = 0.75$ , 两组变量间的 4 个协方差  $\rho$  从 0 以 0.1 步长增加到 0.8。

六组仿真实验的结果分别见图4.1、图4.2、图4.3、图4.4和图4.5。同时，作者也在两组实际数据（心脏病数据和葡萄酒数据）上对上述度量的性能进行了对比，结果见 [23]。以上实验结果表明，基于 CE 的独立性度量在所有情况中都表现了最好的性能，给出了最合理的独立性度量估计值。

### 4.3 条件独立性检验

条件独立性是另一个统计学的基础性概念，与很多其他理论问题密切相连 [466]。我们证明了 TE 可以由 CE 来表示的结论，并给出了相应的估计算法 [14]。因为 TE 本质上是条件互信息，因此我们也同时给出了一个基于信息论的条件独立性度量方法。本领域内也存在一些基于其他理论的同类度量方法 [467]（见表4.2），如前述的基于距离相关的方法 [376]、基于核函数的方法 [375, 468]、基于 copula 的方法 [469] 等等。

为了对比这些方法，我们设计了两组仿真实验并采用了一组实际数据来评估表4.2中包括 CE 在内的 17 种条件独立性方法的效果<sup>2</sup>。实验中对比的条件独立性检验方法如下：

**CDC** Wang 等 [376] 提出了一种条件独立性检验方法，称为 CDC (Conditional Distance Correlation)，将 dCor 独立性检验扩展到条件独立性的情况。CDC 的定义主要是将式(4.13)扩展到

<sup>2</sup>实验代码：<https://github.com/majianthu/eval>



图 4.1: 二元独立性检验对比实验 (二元正态分布) .



图 4.2: 二元独立性检验对比实验 (二元正态 Copula 函数) .



图 4.3: 二元独立性检验对比实验 (二元阿基米德 Copula 函数) .

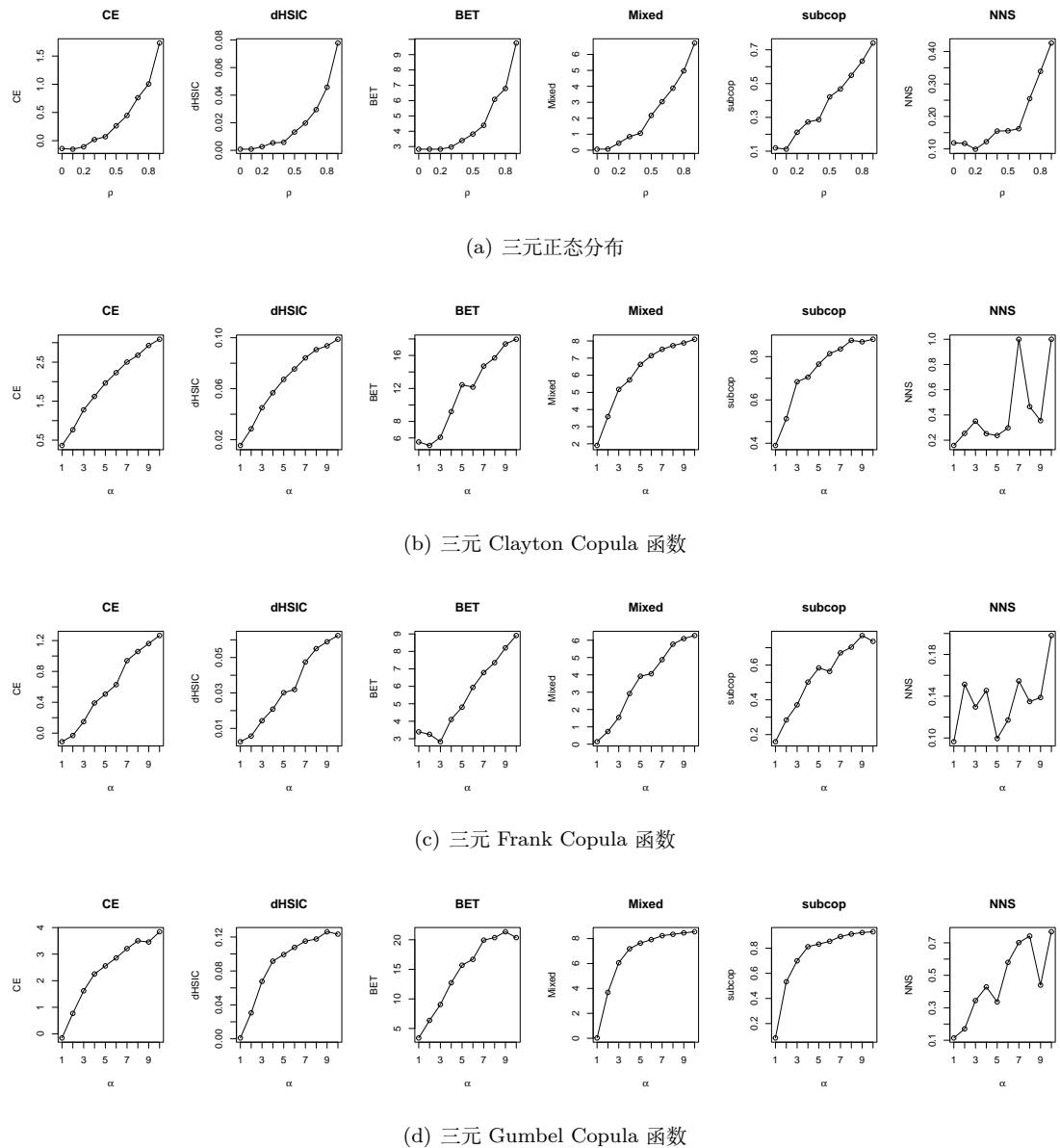


图 4.4: 多元独立性检验对比实验.



图 4.5: 多元随机变量独立性检验对比实验 (四元正态分布) .

条件独立性的情况, 如下

$$\nu^2(X, Y | Z; w) = \|f_{X, Y | Z}(t, s) - f_{X | Z}(t)f_{Y | Z}(s)\|_w^2. \quad (4.20)$$

**CMD** Part 等 [470] 通过扩展 MDC 方法提出了一种基于 dCor 的条件独立性检验方法。

**CODEC** 式(4.8)可以扩展到条件独立性检验的问题。给定随机变量  $(X, Y, Z)$ ,  $X, Y$  相对于  $Z$  的条件独立性可以通过下式进行判断:

$$E(Y | X, Z) = E(Y, Z). \quad (4.21)$$

基于此, Azadkia 和 Chatterjee [471] 提出了一种条件独立性度量, 扩展了 CODEC 方法。

**KPC** Huang 等 [472] 提出了一种基于核函数的条件独立性检验方法, 称为 KPC (Kernel Partial Correlation), 用于检验式(4.21)是否成立。KPC 方法可以认为是 CODEC 方法在 RKHS 的对应。

**FCIT** Chalupka 等 [473] 提出了一种基于预测方法检验式(4.21)的条件独立性检验方法, 称为 FCIT (Fast Conditional Independence Test)。

**PCIT** Burkart 和 Király [474] 提出了另一种基于预测检验式(4.21)的条件独立性检验方法, 称为 PCIT (Predictive Conditional Independence Test)。

**CCIT** Sen 等 [475] 提出了一种基于分类器的条件独立性检验方法，称为 CCIT (Classifier Conditional Independence Test)。给定随机变量  $(X, Y, Z)$ ，该方法检验如下命题：

$$p_{X,Y,Z}(x, y, z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z), \quad (4.22)$$

其中  $p$  为概率密度函数。

**CMI1** CMI 是一种基于信息论的条件独立性度量。Runge [476] 提出了一种基于近邻法的 CMI 估计方法，通过估计下式得到 CMI 的估计值：

$$I_{X,Y|Z} = H_{XZ} + H_{YZ} - H_Z - H_{XYZ}. \quad (4.23)$$

**CMI2** Mesner 和 Shalizi [477] 提出了另一种估计式(4.23)的 CMI 估计方法，用于离散和连续变量混合的情况。

**PCor** 偏相关 (Partial Correlation) 是一种普遍采用的简便的条件独立性度量 [478,479]。偏相关可以通过回归模型的残差之间的相关性估计得到。

**GCM** Shah 和 Peters [480] 提出了一种条件独立性度量方法，称为 GCM (Generalised Covariance Measure)，其由回归函数残差之间的正则协方差得到。

**wGCM** Scheidegger 等 [481] 提出了一种基于加权函数的 GCM 的扩展方法，称为 wGCM。

**KCIT** Zhang 等 [375] 提出了一种基于核函数的条件独立性方法，称为 KCIT。KCIT 可以理解为在核函数映射的 RKHS 上的偏相关测试。

**RCoT** Strobl 等 [468] 提出了两种 KCIT 的近似估计方法，称为 RCIT 和 RCoT，用于解决 KCIT 计算量大的问题。

**PCop** 给定随机变量  $(X, Y, Z)$ ，偏 Copula (Partial Copula) 是条件边缘函数  $U_X, U_Y$  的联合函数  $(U_X, U_Y)$ ，其中  $U_X, U_Y$  定义如下

$$U_X = F_{X|Z}(X|Z), U_Y = F_{Y|Z}(Y|Z). \quad (4.24)$$

Petersen 和 Hansen [469] 提出了一种通过测试  $U_X, U_Y$  之间独立性来进行条件独立性测试的方法。

仿真实验利用如下概率分布生成仿真数据，包括

表 4.2: 评估的条件独立性度量方法及其软件实现.

算法包	条件独立性度量	语言
<code>copent</code>	CE [14]	R
<code>EDMeasure</code>	CMDM [470]	R
<code>FOCI</code>	CODEC [471]	R
<code>RCIT</code>	RCoT [468]	R
<code>cdcsis</code>	CDC [376]	R
<code>GeneralisedCovarianceMeasure</code>	GCM [480]	R
<code>weightedGCM</code>	wGCM [481]	R
<code>comets</code>	PCM [482]	R
<code>KPC</code>	KPC [472]	R
<code>ppcor</code>	PCor [478]	R
<code>parCopCITest</code>	PCop [469]	R
<code>causallearn</code>	KCI [375]	Python
<code>pycit</code>	CMI1 [476]	Python
<code>knnncmi</code>	CMI2 [477]	Python
<code>fcit</code>	FCIT [473]	Python
<code>CCIT</code>	CCIT [475]	Python
<code>pcit</code>	PCIT [474]	Python

1. 随机变量  $(X, Y, Z)$  满足三元正态分布  $N(\mathbf{u}, \Sigma)$ , 其中协方差矩阵  $\Sigma$  为

$$\begin{vmatrix} 1 & \rho_{xy} & \rho_{xz} \\ \rho_{xy} & 1 & \rho_{yz} \\ \rho_{xz} & \rho_{yz} & 1 \end{vmatrix}, \quad (4.25)$$

两个协方差固定不变  $\rho_{xy} = 0.7, \rho_{yz} = 0.6$ , 条件变量对应的第三个协方差  $\rho_{xz}$  从 0 以 0.1 步长增加到 0.9, 以模拟条件独立性强度变化;

2. 随机变量  $(X, Y, Z)$  满足三元正态 copula 函数  $C_N(u_x, u_y, u_z; \Sigma)$ , 其中协方差矩阵  $\Sigma$  与上同, copula 函数中条件变量对应的第三个协方差  $\rho_{xz}$  从 0 以 0.1 步长增加到 0.9, 以模拟条件独立性强度变化。

这样, 我们就基于三元正态分布和三元正态 copula 函数得到的仿真数据生成强度逐渐变化的一组条件独立性关系。我们将这 17 种条件独立性度量方法用于这两组仿真实验数据估计条件独立性强度。结果表明, CE 方法在仿真数据上能够评估出渐次变化的条件独立性强度 (结果见图4.6和图4.7), 给出较同类方法同样或更优的估计结果。



图 4.6: 在三元正态分布上仿真条件独立性关系的评估实验结果。



图 4.7: 在三元正态 copula 函数上仿真条件独立性关系的评估实验结果。

## 4.4 正态性检验

正态性假设是统计学分析中最常见的理论假设之一，其假设检验方法一直以来都是本领域的重要研究课题，大量正态性检验方法被提出来，如 BHEP 方法、基于偏度和散度的方法、基于距离相关的方法等等 [401]。我们基于 CE 概念提出了一个多元正态性检验的方法，并给出了统计量的估计算法 [16]。为了与同类算法进行对比，我们调研得到了基于 R 语言实现的 29 种重要的多元正态性检验方法（见表4.3），并与基于 CE 的方法进行仿真实验对比。仿真实验中对比的方法包括：

**Mardia** Mardia 提出了一种基于偏度（Skewness）和峰度（Kurtosis）多变量扩展的多元正态性检验（MVNT）方法 [483, 484]。其多元偏度和峰度定义为

$$S = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n m_{ij}^3, \quad (4.26)$$

$$K = \frac{1}{n} \sum_{i=1}^n m_{ii}^2, \quad (4.27)$$

其中  $m_{ij}$  为马氏距离的平方。

**Henze-Zirkler** Henze 和 Zirkler [485] 提出了一种基于两个概率分布函数之间的非负泛函距离的 MVNT 方法，该方法具有仿射变换不变性。

**Royston** Royston [486–488] 提出了一种基于 Shapiro-Wilk 统计量的 MVNT 方法。

**Doornik-Hansen** Doornik 和 Hansen [489] 提出了一种 MVNT 方法，先利用变换将数据变为相互独立的变量，再利用偏度和峰度统计量进行检验。

**Energy** Székely 和 Rizzo [452] 提出了一种基于能量统计量的 MVNT 方法，该方法具有仿射变换不变性。

**Anderson-Darling** Anderson-Darling 法 [490] 是一种通用的拟合优度检验方法，可以用于 MVNT 问题 [491]。

**Cramér-von Mises** Koziol [492] 提出了一种基于 Cramér-von Mises 统计量的 MVNT 方法。

**Nikulin-Rao-Robson** Vionov 等 [493] 提出了一种基于 Nikulin-Rao-Robson 统计量的  $\chi^2$  式 MVNT 法。

**McCulloch** McCulloch [494] 提出了一种  $\chi^2$  拟合优度统计量，并应用于 MVNT 问题。

**Dzhaparidze-Nikulin** Dzhaparidze 和 Nikulin [495] 提出了一种  $\chi^2$  拟合优度统计量，可用于 MVNT 问题。

**BHEP** Henze 和 Wagner [496] 提出了一种基于两个函数之间距离的 MVNT 统计量，称为 BHEP 测试。

**Cox-Small** Cox 和 Small [497] 提出了一种基于回归的 MVNT 方法。

**DEHT 和 DEHU** Dörr 等 [498, 499] 基于多元正态分布是一类偏微分方程初始值问题的唯一解的特性，提出了两种 MVNT 问题的统计量（DEHT 和 DEHU）。

**EHS** Ebner 等 [500] 提出了一种基于偏微分方程初始值求解问题的 MVNT 统计量。

**HJG** Henze 和 Jiménez-Gamero [501] 提出了一种基于加权  $L^2$  距离泛函的统计量。

**HV** Henze 和 Visagie [502] 提出了一种基于偏微分方程求解原理的 MVNT 统计量。

**HZ** 同 Henze-Zirkler。

**KKurt** Kozoil [503] 提出了一种多变量峰度统计量。

**MAKurt 和 MASkew** Malkovich 和 Afifi [504] 提出了一种多变量偏度和峰度统计量。

**MKurt 和 MSkew** 同 Mardia。

**MQ1 和 MQ2** Manzotti 和 Quiroz [505] 提出了两种基于球谐函数（Spherical Harmonics）二次型的 MVNT 统计量。

**MRSSkew** Móri 等 [506] 提出了一种多变量的偏度和峰度。

**PU** Pudelko [507] 提出了一种基于经验特征函数的 MVNT 统计量。

**SR** 同 Energy。

**Shapiro-Wilk** 同 Royston。

我们设计了三组仿真实验<sup>3</sup>来模拟不同的非正态性变化情况：

---

<sup>3</sup>实验代码：<https://github.com/majianthu/mvnt>

表 4.3: 评估的多元正态性检验方法及其 R 语言软件实现.

算法包	检验方法
<code>copent</code>	CE [16]
<code>MVN</code> [508]	Mardia, Royston, Henze-Zirkler Dornik-Haansen, Energy
<code>mvnTest</code> [509]	Anderson-Darling, Cramér-von Mises Nikulin-Rao-Robson, McCulloch Dzhaparidze-Nikulin
<code>mnt</code> [401]	BHEP, Cox-Small, DEHT, DEHU, EHS, HJG HV, HZ, KKurt, MAKurt, MASkew, MKurt MQ1, MQ2, MRSSkew, MSkew, PU, SR
<code>mvnormtest</code>	Shapiro-Wilk [486]

1. 第一组实验仿真二元随机变量  $\mathbf{X} = (X_1, X_2)$  满足二元正态 copula 函数  $C_N(u, v)$ , 对应两个边缘函数为正态分布  $u \sim N(0, 1)$  和指数分布  $v \sim E(\lambda)$  生成, 其中由指数边缘分布的参数  $\lambda = 1, \dots, 10$ , 当  $\lambda = 2$  时最接近正态分布, 而在 2 以后随着其值的增加, 分布的非正态性逐渐变大;
2. 第二组实验仿真二元随机变量  $\mathbf{X} = (X_1, X_2)$  满足二元  $t$  分布函数  $\mathbf{X} \sim t_\nu(\mathbf{0}, \Sigma)$ , 矩阵  $\Sigma$  中的非对角元素  $\rho = 0.5$ , 而通过自由度参数  $\nu = 1, \dots, 10$  逐渐增加来仿真非正态性的逐渐减小。
3. 第三组实验首先仿真生成两个不同的二维正态分布  $X_1 \sim N_1(\mu_1, \rho_1)$  和  $X_2 \sim N_2(\mu_2, \rho_2)$  的同样数量的样本, 其中  $\mu_1 = \mathbf{0}, \mu_2 = \mathbf{3}, \rho_1 = 0.5, \rho_2 = 0.8$ , 再将两个样本集的对应样本进行加权  $\{(\beta - 1)X_1 + (10 - \beta)X_2\}/9$ , 从而得到一个混合模型样本集, 其中  $\beta = 1, \dots, 10$ , 使得混合样本分布的非正态性随着  $\beta$  的变化经历由小变大再变小的过程。

我们将 CE 和其他 29 种检验方法应用于这三组仿真数据得到检验统计量。实验结果见图4.8、图4.9和图4.10, 可以看出基于 CE 的方法的统计量能够刻画出数据分布非正态性的三种预设变化, 比所有对比方法的检验结果更准确合理。

## 4.5 双样本检验

双样本检验是统计学中基础性的假设检验问题之一, 很多其他的理论问题, 如对称性检验、单样本检验和变点检测等问题都可以转化成此类问题。同时, 对比两组分布的差异也是实际应用中经常面对的问题之一。双样本检验可分为单变量问题和多变量问题两种情况, 相应地都有很多经典的方法, 单变量的有 Wilcoxon 方法、Kruskal-Wallis 方法和 Kolmogorov-Smirnov 方法等 [510], 双样本检验有基于统计量的方法 (如距离相关 [454]、核函数 [412]、HHG 统计量 [353]、

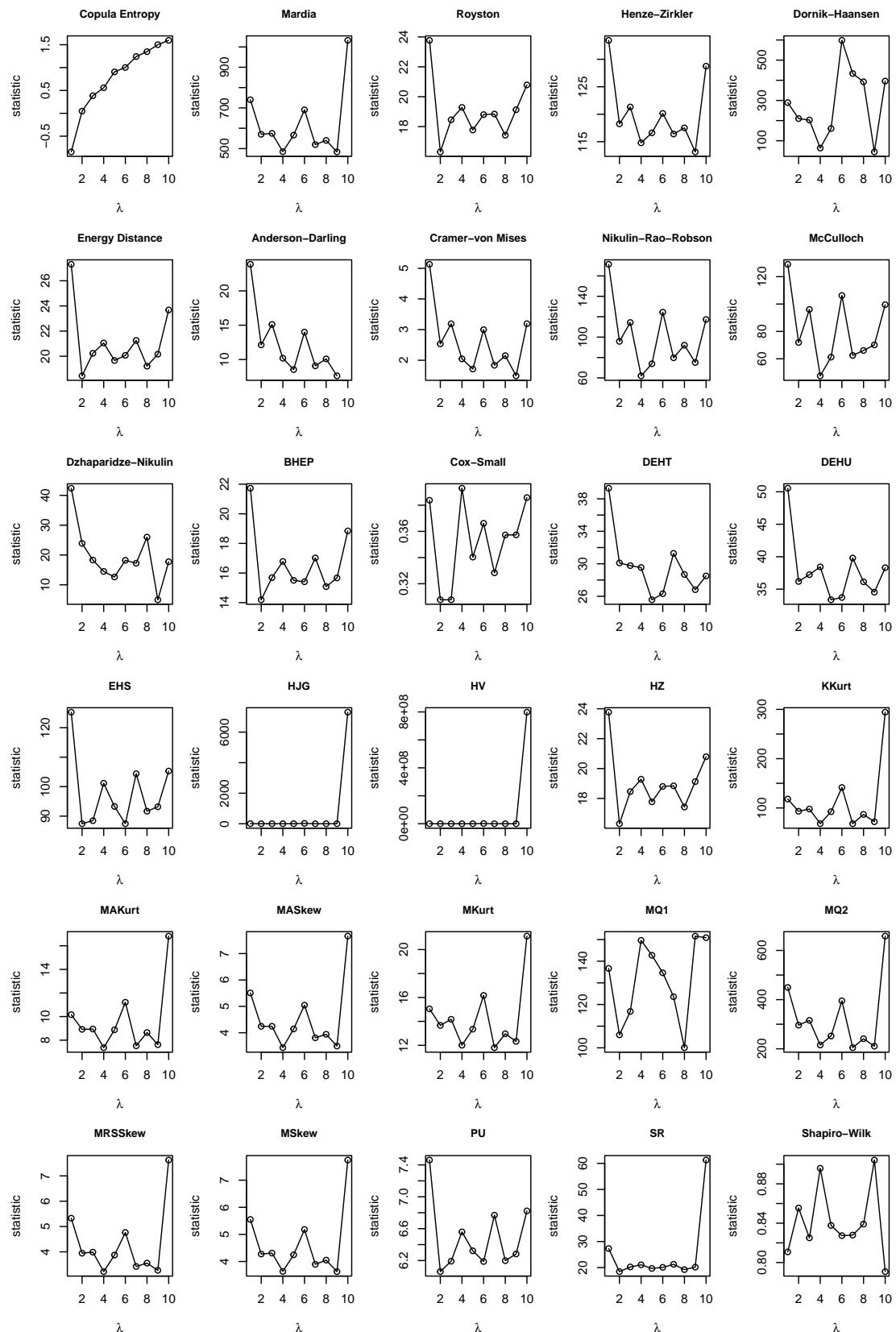


图 4.8: 基于边缘分布变化的多元正态性检验评估实验结果。



图 4.9: 基于 copula 函数变化的多元正态性检验评估实验结果。



图 4.10: 基于混合模型分布变化的多元正态性检验评估实验结果。

Ball 统计量 [458] 等) 和基于统计学习的方法 (如随机森林 [511]) 等。我们提出了一种基于 CE 的双样本检验方法，并给出了统计量的估计算法 [19]。为了与同类方法进行对比，我们调查整理了 R 语言环境下实现的单变量和多变量同类方法 (见表4.4和表4.5)。

仿真实验中对比的单变量双样本检验方法包括：

**Wilcoxon1** Wilcoxon 秩和检验 [427] (又称 Mann-Whitney 检验 [512]) 是一种经典的基于秩统计量的单变量非参数双样本检验方法，用于检验两组样本是否来自同一分布。

**Kruskal-Wallis** Kruskal-Wallis 法由 Kruskal 和 Wallis 于 1952 年提出 [513]，是一种非参数双样本检验方法，用于检验多组样本的均值是否存在统计意义上的差异。

**Kolmogorov-Smirnov** Kolmogorov-Smirnov 法 [514,515] 是一种单变量非参数的假设检验方法，用于检验两个样本是否来自同一分布。该方法的统计量是双样本经验分布函数差值绝对值的最大值。

**Cramér-von Mises** Cramér-von Mises 检验 [516,517] 是一种单变量非参数的假设检验方法，由 Cramér 和 von Mises 分别于 1927 和 1931 年提出。该方法的统计量是双样本经验分布函数差值绝对值的和。

**Kuiper** Kuiper 法 [518] 是一种单变量非参数的双样本检验方法，由 Kuiper 于 1960 年提出。

**WASS** Wasserstein 法是一种单变量非参数的双样本检验方法，其统计量定义为双样本经验分布函数的 Wasserstein 距离 [519]。

**DTS** Dowd [520] 提出了一种单变量非参数的双样本检验方法，其统计量是双样本经验分布函数的加权和。

**AD** Pettitt [521] 提出了一种基于 Anderson-Darling 秩统计量的双样本检验方法，其统计量是双样本经验分布函数差值的加权和。

**Wilcoxon2** 同 Wilcoxon1。

**Vartest** Ammous 等 [522] 提出了一种基于 James-Welch ANOVA 的双样本检验方法，检验双样本是否具有相同的方差。

**LR** LR 法是一种由 Lehmann 和 Rosenblatt 分别在 1951 和 1952 年提出的基于混合样本秩的双样本检验方法 [523,524]。

**ZA,ZK 和 ZC** Zhang [525] 提出了三种基于似然比的双样本检验方法。

**TNL** Aliev 等 [526] 提出了一种基于秩统计量的非参数双样本检验方法。

仿真实验中对比的多变量双样本检验方法包括：

**MI** 基于 MI 的双样本检验 [527,528] 是一类非参数检验方法，利用样本和标注之间的互信息作为检验统计量。

**Kernel** Gretton 等 [412] 提出一种基于核函数的双样本检验，其统计量定义为样本分布在 RKHS 的距离，称为最大均值差异（Maximum Mean Discrepancy）。

**Energy** Székely 和 Rizzo [454] 提出了一种基于能量统计量的双样本检验方法。

**Ball** Pan 等 [458] 提出了一种基于球散度（Ball Divergence）的非参数双样本检验方法。

**Random Forest** Hediger 等 [511] 提出了一种基于随机森林分类器的双样本检验方法。

**HHG** Heller 等 [353] 提出了一种基于 HHG 独立性检验的 k 样本检验方法，其统计量是样本和类别标注之间的独立性度量值。

**Cramer** Baringhaus 和 Franz [529] 提出了一种多变量双样本检验方法，其统计量是双样本随机变量欧式空间距离和的差值。

**TST.HD** Cousido-Rocha 等 [530] 提出了一种多变量双样本检验方法，其统计量是双样本经验特征函数之间的距离。

**F-F** Fasano 和 Franceschini [531] 提出了一种 Kolmogorov-Smirnov 双样本检验的多变量扩展版本。

**Peacock** Peacock [532] 提出了一种二变量情况的 Kolmogorov-Smirnov 双样本检验方法。

**RPT** Lopes 等 [533] 提出了一种多变量双样本检验方法，其首先利用随机映射将高维样本变换为低维样本，再利用 Hotelling  $T^2$  检验得到统计量。

**Depth** Liu 和 Singh [534] 提出了一种基于深度（Depth）度量的双样本检验方法。

**AD** 同前。

**QN** Kruskal [535] 提出了一种基于秩统计量的非参数 k 样本检验方法。

**BM** Neubert 和 Brunner [536] 提出了一种基于学生氏化秩统计量的双样本检验方法。

**WASS** 同前。

**SWD** Wasserstein 距离 [537] 可以被用于多变量双样本检验问题。

**Graph** Bai 和 Chu [538] 提出了一种基于图 (Graph) 方法的多变量双样本检验方法。

**KMD** Huang 和 Sen [539] 提出了一种非参数 k 样本检验方法，其统计量为基于核函数的 k 样本相似度度量。

我们设计了单变量和多变量两组对比仿真实验<sup>4</sup>:

**单变量实验** 单变量情况的对比仿真实验设计为检验两个不同的单变量正态分布，有两组：

1. 第一组仿真首先生成满足正态分布的随机变量  $x_0 \sim N(\mu_0, \delta_0)$ ，其中均值  $\mu_0 = 0$  和方差  $\delta_0 = 1$ ，再生成满足正态分布的随机变量  $x_2 \sim N(\mu_1, \delta_1)$ ，其均值  $\mu_1$  的值由 0 增加到 9， $\delta_1 = 1$ ；
2. 第二组仿真首先生成同样的随机变量  $x_0$ ，第二个随机变量  $x_1 \sim N(\mu_1, \delta_1)$  的方差  $\delta_1$  的值则由 1 增加到 10， $\mu_1 = 0$ 。

我们利用所有双样本检验方法从仿真数据估计检验统计量。实验结果见图4.11。结果表明，CE 方法的统计量正确地反映了实验设定情况，取得了与对比方法相同或更好的结果。

**多变量实验** 多变量情况的对比仿真实验也是设计为检验对比两个二元分布，有三组：

1. 第一组仿真均值差变化的两个满足正态分布的二元随机变量  $\mathbf{x}_0 \sim N(u_0, \rho_0)$  和  $\mathbf{x}_1 \sim N(u_1, \rho_1)$ ，第一个均值  $u_0 = \mathbf{0}$ ，第二个均值为  $u_1 = (i, i)$ ， $i$  从 0 以 1 为步长增加到 9，两个分布的协方差相同  $\rho_0, \rho_1 = 0.5$ ；
2. 第二组仿真首先仿真一个满足二元正态分布的二元随机变量  $\mathbf{x}_0 \sim N(u_0, \rho_0)$ ，其中  $u_0 = \mathbf{0}, \rho_0 = 0$ ，第二个二元随机变量  $\mathbf{x}_1$  满足的正态分布  $N(u_1, \rho_1)$  的协方差  $\rho_1$  则由 0 以 0.1 为步长增加到 0.9， $u_1 = \mathbf{0}$ ；
3. 第三组仿真首先仿真一个满足二元正态分布的随机变量  $\mathbf{x}_0 \sim N(u_0, \rho_0)$ ，协方差  $\rho_0 = 0$ ，第二个二元随机变量  $\mathbf{x}_1$  分布则满足二元正态 copula 函数  $c(u, v)$ ，其两个边缘函数为正态函数  $u \sim N(0, 2)$  和指数函数  $v \sim E(\lambda)$ ，正态 copula 函数和边缘正态函数参数不变，指数函数的参数  $\lambda$  从 1 增加到 10，以仿真不断变大的二元非正态性。

我们利用这些检验方法从仿真数据估计检验统计量。三组实验结果分别见图4.12、图4.13和图4.14，从中可见基于 CE 的方法在多变量的情况下仍然具有良好的检验性能，很好地反映了不同情况下分布差异的变化，具有与对比方法相同或更优的检验能力。

---

<sup>4</sup>实验代码：<https://github.com/majianthu/tst>



(a) 均值变化实验



(b) 方差变化实验

图 4.11: 单变量双样本检验评估实验结果。



图 4.12: 多变量双样本检验评估实验（均值变化）结果。



图 4.13: 多变量双样本检验评估实验 (协方差变化) 结果。



图 4.14: 多变量双样本检验评估实验 (正态 Copula 函数参数变化) 结果。

表 4.4: 评估的单变量双样本检验方法及其 R 语言软件实现.

算法包	方法
<code>copent</code>	CE [19]
<code>stat</code>	Wilcoxon1 [510] Kruskal-Wallis [510]
<code>twosamples</code> [520]	CVM, KS, Kuiper WASS, DTS, AD
<code>robusTest</code> [522]	Wilcoxon2, Vartest
<code>R2sample</code>	LR [523,524], ZA,ZK,ZC [525]
<code>tnl.Test</code>	TNL [526]

## 4.6 变点检测

变点检测是时序数据分析领域重要的问题之一，可分为在线和离线、单变量和多变量、参数和非参数等不同类型。它在很多不同领域（如工业过程、神经数据分析、计量经济学等）都有重要应用。此问题很早就被提出，研究者提出了很多解决方法，相关综述见 [417, 418, 544]。我们基于 CE 提出了一种非参数多变点检测方法 [20]，可以在无分布假设的情况下应用到任何场合。为了与同类方法进行对比，我们调研整理了基于 R 语言实现的变点检测方法（见表4.6），并设计了对比仿真实验<sup>5</sup>。

仿真实验对比的 R 算法包中的变点检测算法包括：

**changepoint** Scott 和 Knott [545] 提出的基于二分割（Binary Segmentation）的单变量变点检测算法被用于检测 3 种单变量变点的情况。

**ecp** James 和 Matteson [546] 提出的基于核函数的多变量变点检测算法被用于 3 种多变量变点的情况。

**rid** Fan 和 Wu [547] 提出的 random interval distillation (RID) 算法被用于所有 6 种变点的情况。

**CptNonPar** McGonigle 和 Cho [548] 提出的非参数多变点检测算法被用于所有 6 种变点的情况。

**npwbs** Ross [549] 提出的基于 Wild Binary Segmentation (WBS) 的变点检测算法被用于单变量变点的 3 种情况。

<sup>5</sup>实验代码：<https://github.com/majianthu/cpd>

表 4.5: 评估的多变量双样本检验方法及其 R 语言软件实现.

算法包	方法
<code>copent</code>	CE [19] MI [527,528]
<code>kernlab</code> [540]	Kernel [412]
<code>energy</code>	Energy [454]
<code>Ball</code>	Ball divergence [458]
<code>hypoRF</code>	Random Forest [511]
<code>HHG</code> [353]	HHG sum.chisq HHG sum.lr HHG max.chisq HHG max.lr
<code>cramer</code>	Cramer [529]
<code>TwoSampleTest.HD</code> [541]	TST.HD
<code>fasano.franceschini.test</code> [542]	F-F
<code>Peacock.test</code>	Peacock [532]
<code>RandomProjectionTest</code>	RPT [533]
<code>DepthProc</code>	Depth [534]
<code>kSamples</code>	AD [521] QN [535]
<code>lawstat</code>	BM [536]
<code>T4transport</code>	WASS [543] SWD [537]
<code>rgTest</code>	Graph [538]
<code>KMD</code>	KMD [539]

**MFT** Messer 等 [550] 提出的 multiple filter test (MFT) 算法被用于单变量均值变点的情况。

**jcp** Messer [551] 提出的单变量变点检测算法被用于 3 种单变量变点的情况。

**InspectChangepoint** Wang 和 Samworth [552] 提出的基于稀疏映射的变点检测算法被用于所有 6 种变点的情况。

**hdbinseg** Cho 和 Fryzlewicz [553] 提出的基于稀疏化二分割的变点检测算法被用于 3 种多变量变点的情况。

**changepoint.np** Killick 等 [554] 提出的低计算量的最优变点检测算法被用于 3 种单变量变点的情况。

**changepoint.geo** Grundy 等 [555] 提出的基于几何降维映射的多变量变点检测算法被用于 3 种多变量变点的情况。

**mosum** Eichinger 和 Kirch [556] 提出的基于滑动和统计量的变点检测算法被用于 3 种单变量变点的情况。

**SNSeg** Zhao 等 [557] 提出的基于自归一方法的变点检测算法被用于所有 6 种变点的情况。

**offlineChange** Ding 等 [558] 提出的多变量变点检测算法被用于 3 种多变量变点的情况。

**IDetect** Anastasiou 和 Fryzlewicz [559] 提出的基于变点隔离的算法被用于单变量均值变点情况。

**wbs** Fryzlewicz [560] 提出的基于 WBS 的变点检测算法被用于单变量均值变点情况。

**breakfast** Fryzlewicz [561] 提出的基于 WBS2 的变点检测算法被用于单变量均值变点情况。

**mscp** Levajkovic 和 Messer [562] 提出的基于自适应滑动和过程的变点检测算法被用于单变量均值变点情况。

**L2hdchange** Li 等 [563] 提出的基于双向滑动和的变点检测算法被用于多变量变点的 3 种情况。

**gfpop** Hocking 等 [564] 提出的基于动态规划求解的最优变点检测算法被用于 3 种单变量变点的情况。

**HDCD** Moen 等 [565] 提出的稀疏自适应变点检测算法被用于所有 6 种变点的情况。

**HDDchangepoint** Drikvandi 和 Modarres [566] 提出的非参数变点检测算法被用于多变量变点的 3 种情况。

**HDcpDetect** Li 等 [567] 提出的多变量变点检测算法被用于多变量变点的 3 种情况。

**decp** Ryan 和 Killick [568] 提出的基于随机矩阵理论的协方差变点检测算法被用于多变量变点的 3 种情况。

**cpss** Zou 等 [569] 提出的基于样本分割策略估计变点数的算法被用于单变量变点的 3 种情况。

仿真实验的设计如下：首先生成 4 组均值-（协）方差作出相应变化的单/双变量正态分布的随机样本，再将仿真数据顺次连接以模拟分布变化具有 3 个变点的序列数据，最后我们将表4.6 中的 26 个算法包中适用于仿真变点情况的方法进行检测实验。

对比仿真实验模拟了 6 种不同类型的变点情况，分别是

1. 单变量下均值变点、均值-方差变点和方差变点；
2. 双变量下均值变点、均值-方差变点和方差变点。

六种变点情况的仿真实验中四个正态分布的均值和（协）方差的设定值见表4.7。我们通过多次重复以上仿真实验，计算在每种变点情况下各个算法每次检验到的变点个数的平均值，以作为算法的检测性能指标。结果（见图4.15）表明，基于 CE 的方法能够检测出所有 6 种情况下的变点位置，获得了与对比方法同等或更好的检测性能结果。特别是在检测难度大的单/双变量方差变点的情况下，基于 CE 的方法性能明显优于对比方法。值得一提的是，基于 CE 的方法是唯一适用于所有 6 种不同情况且在所有情况下均性能良好的算法，同时其参数调节量是最小的，体现了良好的普适性和实用性。而一些对比方法则只适用于某种特定情况（单变量或多变量、均值或方差）的变点，需要在算法中预先假定变点类型，且需要对算法的相关参数进行调优。

## 4.7 对称性检验

我们设计了三组仿真实验以验证基于 CE 的对称性检验的有效性，并将其与同类方法进行对比 [21]。我们选择了 R 包 `symmetry` 中实现的 22 种对称性检验方法进行对比，包括

**MI** Mira 检验统计量 [436]；

**CM** Cabilio–Masaro 检验统计量 [437]；

**MGG** Miao-Gel-Gastwirth 检验统计量 [433]；

**B1**  $\sqrt{b_1}$  检验统计量 [435]；

表 4.6: 多变点检测方法评估实验中的 R 语言算法软件实现.

算法包	单变量			多变量		
	均值	均值方差	方差	均值	均值方差	方差
copent [20]	✓	✓	✓	✓	✓	✓
changepoint [570]	✓	✓	✓			
ecp [546]				✓	✓	✓
rid [547]	✓	✓	✓	✓	✓	✓
CptNonPar [548]	✓	✓	✓	✓	✓	✓
npwbs [549]	✓	✓	✓			
MFT [550]	✓					
jcp [551]	✓	✓	✓			
InspectChangepoint [552]	✓	✓	✓	✓	✓	✓
hdbinseg [553]				✓	✓	✓
changepoint.np [554]	✓	✓	✓			
changepoint.geo [555]				✓	✓	✓
mosum [571]	✓	✓	✓			
SNSeg [572]	✓	✓	✓	✓	✓	✓
offlineChange [558]				✓	✓	✓
IDetect [559]	✓					
wbs [560]	✓					
breakfast [573]	✓					
mscp [562]	✓					
L2hdchange [563]				✓	✓	✓
gfpop [574]	✓	✓	✓			
HDCD [565,575]	✓	✓	✓	✓	✓	✓
HDDchangepoint [566]				✓	✓	✓
HDcpDetect [567]				✓	✓	✓
decp [568]				✓	✓	✓
cpss [569]	✓	✓	✓			



图 4.15: 变点检测算法评估仿真实验中各个算法检测到的平均变点数.

表 4.7: 多变点检测方法评估实验中不同变点情况数据仿真中正态分布参数的设定值.

均值, (协) 方差	单变量			双变量		
	均值	均值方差	方差	均值	均值方差	方差
$(\mu_1, \rho_1)$	(0,1)	(0,1)	(0,1)	(0,0.2)	(0,0.2)	(0,0.2)
$(\mu_2, \rho_2)$	(5,1)	(5,3)	(0,10)	(10,0.2)	(10,0.8)	(0,0.8)
$(\mu_3, \rho_3)$	(10,1)	(10,1)	(0,5)	(5,0.2)	(5,0.1)	(0,0.1)
$(\mu_4, \rho_4)$	(3,1)	(3,10)	(0,1)	(1,0.2)	(1,0.9)	(0,0.9)

**KS** Kolmogorov-Smirnov 检验统计量 [435];

**SGN** 符号检验统计量 [435];

**WCX** Wilcoxon 检验统计量 [435];

**FM** 基于特征函数的检验统计量 [438];

**RW** Rothman-Woodrooffe 检验统计量 [439];

**BHI** Litvinova 检验统计量 [440];

**BHK** Baringhaus-Henze 上确界式检验统计量 [431];

**BH2** Baringhaus-Henze 检验统计量 [431];

**MOI 和 MOK** Milošević-Obradović 检验统计量 [434];

**NAI 和 NAK** Nikitin-Ahsanullah 检验统计量 [441];

**K2 和 K2U** 基于 V 和 U 统计量的 Božin-Milošević-Nikitin-Obradović Kolmogorov 式统计量 [432];

**NAC1, NAC2, BHC1 和 BHC2** Allison-Pretorius 检验统计量 [430]。

仿真实验<sup>6</sup>采用三个非对称分布函数族来生成对称性连续变化的样本，包括

<sup>6</sup>实验代码: <https://github.com/majianthu/symmetry>

**Beta 分布** Beta 分布是定义在  $[0, 1]$  或  $(0, 1)$  区间的连续分布函数，具有两个参数  $a, b > 0$  来控制函数的对称性，函数定义如下：

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, \quad (4.28)$$

其中  $\Gamma$  是伽玛 (Gamma) 函数。

**非对称拉普拉斯分布** 非对称拉普拉斯分布是扩展了拉普拉斯分布的连续分布函数族，其定义如下：

$$f(x; \mu, \delta, k) = \frac{1}{\delta(k+k^{-1})} e^{-(x-\mu)k^s s/\delta}, \quad (4.29)$$

其中， $s = \text{sgn}(x - \mu)$ ， $\mu, \delta$  分别是位置和尺度参数，参数  $k$  控制分布的对称性。

**双模态正态分布** 双模态正态分布由两个正态分布混合而成，两个分布  $X_1 \sim N(\mu_1, \delta_1)$  和  $X_2 \sim N(\mu_2, \delta_2)$  的混合由一个混合比例参数  $p$  控制，从而由  $p$  来控制混合分布的对称性。

在 Beta 分布的仿真实验中，生成 9 组样本，样本生成条件为  $a+b=10$  且  $b=1, \dots, 9$ 。在非对称拉普拉斯分布的仿真实验中，生成 9 组样本，样本生成条件为  $\mu=0, \delta=1$  且  $k=0.1, \dots, 0.9$ 。在双模态正态分布的仿真实验中，生成 9 组样本，样本生成条件为  $\mu_1=0, \mu_2=5, \delta_1=\delta_2=1$  且  $p=0.1, \dots, 0.9$ 。每组样本的大小均为 300。通过这样的仿真，我们模拟了这三种分布函数的对称性由弱变强再变弱的过程。

三组仿真实验结果分别见图4.16、图4.17和图4.18。由实验结果可以看出，基于 CE 的对称性检验方法成功地检测到了三组实验仿真的分布对称性变化，而其他 22 种对比方法中只有少数方法对部分情况进行了成功检测。特别是在基于双模态正态分布的实验中，我们的方法表明对称性在两端和中间的样本中变强，与实验设计完全一致，而其他对比方法大多未得出此结果。



图 4.16: 基于 Beta 分布的对称性检验仿真实验结果。



图 4.17: 基于非对称拉普拉斯分布的对称性检验仿真实验结果。



图 4.18: 基于双模态正态分布的对称性检验仿真实验结果。

# 第五章 数学推广

## 5.1 Tsallis Copula 熵

香农熵是满足连续性、对称性和可加性等公理性质的信息量度量 [312,313]，CE 作为一类特殊类型的香农熵也满足这些公理属性。香农熵有很多数学扩展定义，其中一个具有代表性的是通过扩展可加性得到的 Tsallis 熵 (Tsallis Entropy) [576–578]。

给定随机变量  $(X, Y) \sim P(x, y)$ ，其香农熵定义为

$$H(x, y) = - \int_x \int_y p(x, y) \log p(x, y) dx dy. \quad (5.1)$$

相应地，Tsallis 熵的数学定义如下

**定义 9** (Tsallis Entropy). 给定随机变量  $(X, Y) \sim P(x, y)$ ，则 Tsallis 熵的定义为

$$H_T(x, y) = \frac{1}{q-1} \left( 1 - \int_x \int_y p^q(x, y) dx dy \right), \quad (5.2)$$

其中  $q \in R$ 。

当  $q \rightarrow 1$  时，Tsallis 熵对应于香农熵，即

$$\lim_{q \rightarrow 1} H_T(x, y) = H(x, y). \quad (5.3)$$

Mortezaejad 等 [24] 通过将式(5.2)中的概率密度函数  $p(x, y)$  换成 copula 密度函数  $c(u, v)$ ，得到了 CE 的扩展定义 Tsallis CE，如下

**定义 10** (Tsallis Copula Entropy). 给定随机变量  $(X, Y)$  及其 copula 密度函数  $c(u, v)$ ，则 Tsallis Copula 熵定义为

$$H_{TC}(x, y) = \frac{1}{q-1} \left( 1 - \int_u \int_v c^q(u, v) du dv \right), \quad (5.4)$$

其中， $q \in R$ 。

当  $q \rightarrow 1$  时，Tsallis CE 对应于 CE，即

$$\lim_{q \rightarrow 1} H_{TC}(x, y) = H_C(x, y). \quad (5.5)$$

Mortezaejad 等在 Tsallis CE 定义的基础上，给出了最大 Tsallis CE 准则的 Copula 函数求解方法 [24]。给定各阶矩约束条件

$$\int_u \int_v c(u, v) dudv = 1, \quad (5.6)$$

$$\int_u \int_v u^r c(u, v) dudv = \frac{1}{r+1}, \quad (5.7)$$

$$\int_u \int_v v^r c(u, v) dudv = \frac{1}{r+1}, \quad (5.8)$$

$r = 1, \dots, n$ , 和 Spearman 的  $\rho$  约束条件

$$\int_u \int_v uv c(u, v) dudv = \frac{\rho+3}{12}, \quad (5.9)$$

最大 Tsallis CE 准则的拉格朗日方程为

$$L(c, \Lambda) = \frac{1}{q-1} \left( 1 - \int_u \int_v c^q(u, v) dudv \right) \quad (5.10)$$

$$- \lambda_0 \left( \int_u \int_v c(u, v) dudv - 1 \right) \quad (5.11)$$

$$- \sum_{r=1}^n \lambda_r^u \left( \int_u \int_v u^r c(u, v) dudv - \frac{1}{r+1} \right) \quad (5.12)$$

$$- \sum_{r=1}^n \lambda_r^v \left( \int_u \int_v v^r c(u, v) dudv - \frac{1}{r+1} \right) \quad (5.13)$$

$$- \lambda_\rho \left( \int_u \int_v uv c(u, v) dudv - \frac{\rho+3}{12} \right), \quad (5.14)$$

其中  $\Lambda = (\lambda_0, \lambda_1^u, \dots, \lambda_n^u, \lambda_1^v, \dots, \lambda_n^v, \lambda_\rho)$ 。将此拉格朗日方程对  $c$  求导，令导数为 0 可得：

$$\frac{\partial L(c, \Lambda)}{\partial c} = -\frac{q}{q-1} c^{q-1} - \lambda_0 - \sum_{r=1}^n \lambda_r^u u^r - \sum_{r=1}^n \lambda_r^v v^r - \lambda_\rho uv = 0. \quad (5.15)$$

则由最大 Tsallis CE 准则得到的 copula 密度函数为：

$$c(u, v) = \sqrt[q-1]{-\frac{q-1}{q}(\lambda_0 + \sum_{r=1}^n (\lambda_r^u u^r + \lambda_r^v v^r) + \lambda_\rho uv)}, \quad (5.16)$$

式中  $\Lambda$  各元素可以由将式(5.16)代回式(5.6)、式(5.7)、式(5.8)和式(5.9)求得。

## 5.2 生存 Copula 熵

生存函数是一类特殊的概率分布函数，定义为

$$\bar{F}(x, y) = P(X > x, Y > y), \quad (5.17)$$

其中  $X, Y$  为随机变量。相应地，可以容易定义二元生存 Copula 函数

$$\bar{C}(u, v) = C(U > u, V > v). \quad (5.18)$$

其中  $u, v$  为随机变量的边缘分布函数。Nair 和 Sunoj [25] 通过将 CE 定义中的 Copula 函数替换为生存 Copula 函数，定义了生存 CE (Survival CE: SCE) 的概念，如下

**定义 11** (Survival Copula Entropy). 给定随机变量  $(X, Y)$ ，及其生存 Copula 函数  $\bar{C}(u, v)$ ，则其生存 Copula 熵定义为

$$H_{\bar{C}}(x, y) = - \int_0^1 \int_0^1 \bar{C}(u, v) \log \bar{C}(u, v) dudv. \quad (5.19)$$

Nair 和 Sunoj 还讨论了 SCE 的上下界、有序性、与依赖性概念 PQD 的关系和单调变换不变性等性质。

### 5.3 Copula 外熵

外熵 (Extropy) 是香农熵的对偶性概念 [579]。给定随机变量  $(X, Y) \sim P(x, y)$ ，其香农熵定义为

$$H(x, y) = - \int_x \int_y p(x, y) \log p(x, y) dx dy, \quad (5.20)$$

则其相应的外熵定义如下 [580]

**定义 12** (Extropy). 给定随机变量  $(X, Y) \sim P(x, y)$ ，则外熵的定义为

$$J(x, y) = \frac{1}{4} \int_0^\infty \int_0^\infty p^2(x, y) dx dy. \quad (5.21)$$

Saha 和 Kayal [26] 利用此外熵定义扩展了 CE 概念，定义了 Copula 外熵 (Copula Extropy: CEx) 的概念，如下

**定义 13** (Copula Extropy). 给定随机变量  $(X, Y)$  及其 Copula 函数  $C(u, v)$ ，则其 Copula 外熵定义为

$$J_c(x, y) = \frac{1}{4} \int_0^1 \int_0^1 c^2(u, v) du dv, \quad (5.22)$$

其中  $c(u, v)$  为二元 copula 密度函数。

Saha 和 Kayal 还定义了累积 Copula 外熵 (Cumulative Copula Extropy: CCEx) 的概念，如下

**定义 14** (Cumulative Copula Extropy). 给定随机变量  $(X, Y)$  及其 Copula 函数  $C(u, v)$ ，则其累积 Copula 外熵定义为

$$J_C(x, y) = \frac{1}{4} \int_0^1 \int_0^1 C^2(u, v) du dv. \quad (5.23)$$

类比 SCE，他们定义了生存 Copula 外熵 (Survival Copula Extropy: SCEx) 的概念，如下

**定义 15** (Survival Copula Extropy). 给定随机变量  $(X, Y)$  及其生存 Copula 函数  $\bar{C}(u, v)$ ，则其生存 Copula 外熵定义为

$$J_{\bar{C}}(x, y) = \frac{1}{4} \int_0^1 \int_0^1 \bar{C}^2(u, v) dudv. \quad (5.24)$$

他们还讨论了 CEx 等概念的有界性、有序性、与独立性度量之间的关系等性质，以及 CEx 与 CE 的关系问题。

## 5.4 累积 Copula 熵

Arshad 等 [27] 通过累积 Copula 函数定义了累积 CE (Cumulative Copula Entropy: CCE) 的概念，如下

**定义 16** (Cumulative Copula Entropy). 给定多元随机变量  $\mathbf{X}$  及其累积 Copula 函数  $C(\mathbf{u})$ ，则其累积 Copula 熵定义为

$$H_{CCE}(C) = - \int_{\mathbf{u}} C(\mathbf{u}) \log C(\mathbf{u}) d\mathbf{u}. \quad (5.25)$$

他们讨论了 CCE 与多变量相关系数之间的关系、算术加权平均 Copula 函数的 CCE 的性质、有序性；定义了 CCE 的生成函数并讨论其性质，以及几何加权平均 Copula 函数的 CCE 生成函数性质。

他们还定义了分形累积 CE (Fractional Cumulative Copula Entropy: FCCE)，如下

**定义 17** (Fractional Cumulative Copula Entropy). 给定多元随机变量  $\mathbf{X}$  及其累积 Copula 函数  $C(\mathbf{u})$ ，则其分形累积 Copula 熵定义为

$$H_{FCCE}(C) = \int_{\mathbf{u}} C(\mathbf{u})(-\log C(\mathbf{u}))^r d\mathbf{u}, \quad (5.26)$$

其中  $0 < r < 1$ 。

他们讨论了 CCE 和 FCCE 之间的关系、算术加权平均 Copula 函数的 FCCE 的性质和 FCCE 的有序性等问题。

他们给出了经验 Copula 的定义，进而定义了经验 Beta CCE (Empirical Beta CCE) 的概念，如下

**定义 18** (Empirical Beta Cumulative Copula Entropy). 给定多元随机变量  $\mathbf{X}$  及其经验 Copula 函数  $\hat{C}(\mathbf{u})$ ，则其经验 Beta 累积 Copula 熵定义为

$$H_{EBCC E}(\hat{C}) = - \int_{\mathbf{u}} \hat{C}(\mathbf{u}) \log \hat{C}(\mathbf{u}) d\mathbf{u}. \quad (5.27)$$

进而可以定义分形经验 Beta CCE 的概念，如下

**定义 19** (Fractional Empirical Beta Cumulative Copula Entropy). 给定多元随机变量  $\mathbf{X}$  及其经验 Copula 函数  $\hat{C}(\mathbf{u})$ , 则其分形经验 Beta 累积 Copula 熵定义为

$$H_{FEBCE}(\hat{C}) = \int_{\mathbf{u}} \hat{C}(\mathbf{u})(-\log \hat{C}(\mathbf{u}))^r d\mathbf{u}, \quad (5.28)$$

其中  $0 < r < 1$ 。

他们还定义了累积 Copula KL 散度 (Cumulative Copula Kullback-Leibler Divergence: CCKL) 的概念, 并将 CCKL 应用到 Copula 函数的拟合优度检验问题上。

**定义 20** (Cumulative Copula Kullback-Leiber Divergence). 给定两个 Copula 函数  $C_1(\mathbf{u}), C_2(\mathbf{u})$ , 则累积 Copula KL 散度定义为

$$D_{CCKL}(C_1, C_2) = \int_{\mathbf{u}} C_1(\mathbf{u}) \log \frac{C_1(\mathbf{u})}{C_2(\mathbf{u})} d\mathbf{u} - \frac{\rho_k(C_1) - \rho_k(C_2)}{2^k n(k)}, \quad (5.29)$$

其中  $\rho_k(C) = n(k)(2^k \int_{\mathbf{u}} C(\mathbf{u}) d\mathbf{u} - 1)$  为  $k$  维 Spearman 相关系数,  $n(k) = \frac{k+1}{2^k - k - 1}$ 。

## 5.5 累积 Copula Tsallis 熵

给定多元随机变量  $\mathbf{X}$ , Tsallis 熵的定义式(5.2)也可以写成如下形式 [576]:

$$H_T(\mathbf{x}) = \frac{1}{q-1} \int_{\mathbf{x}} (p(\mathbf{x}) - p(\mathbf{x})^q) d\mathbf{x}, \quad (5.30)$$

其中  $q \in R$ 。定义式(5.30)进而可以写成类似香农熵的形式:

$$H_T(\mathbf{x}) = - \int_{\mathbf{x}} p(\mathbf{x}) \log_q(p(\mathbf{x})) d\mathbf{x}, \quad (5.31)$$

其中  $\log_q(r) = \frac{r^{q-1}-1}{q-1}$ ,  $q > 0, q \neq 1$ .

作为 CE 的 Tsallis 熵扩展, Zachariah 等 [28] 利用累积 Copula 函数提出了累积 Copula Tsallis 熵 (Cumulative Copula Tsallis Entropy: CCTE) 的概念, 定义如下

**定义 21** (Cumulative Copula Tsallis Entropy). 给定多元随机变量  $\mathbf{X}$  及其 Copula 函数  $C(\mathbf{u})$ , 则其累积 Copula Tsallis 熵的定义为

$$H_{CCTE}(\mathbf{x}) = - \int_{\mathbf{u}} C(\mathbf{u}) \log_q(C(\mathbf{u})) d\mathbf{u}, \quad (5.32)$$

其中  $q > 0, q \neq 1$ 。

他们又定义了经验累积 Copula Tsallis 熵 (Empirical Cumulative Copula Tsallis Entropy: ECCTE), 如下

**定义 22** (Empirical Cumulative Copula Tsallis Entropy). 给定多元随机变量  $\mathbf{X}$  及其 Copula 函数  $C(\mathbf{u})$ ,  $\hat{C}(\mathbf{u})$  为相应的经验累积 Copula 函数, 则其经验累积 Copula Tsallis 熵的定义为

$$H_{ECCTE}(\mathbf{x}) = - \int_{\mathbf{u}} \hat{C}(\mathbf{u}) \log_q(\hat{C}(\mathbf{u})) d\mathbf{u}, \quad (5.33)$$

其中  $q > 0, q \neq 1$ 。

他们证明了 ECCTE 向 CCTE 的收敛性，从而给出了 CCTE 的非参数估计方法。

他们定义了累积 Copula Tsallis 不准确度的概念，如下

**定义 23** (Cumulative Copula Tsallis Inaccuracy). 给定两个连续多元随机变量  $\mathbf{X}_1, \mathbf{X}_2 \in R^n$  及其 Copula 函数  $C_1(\mathbf{u})$  和  $C_2(\mathbf{u})$ ，则累积 Copula Tsallis 不准确度定义为

$$D_{CCTI}(C_1, C_2) = - \int_{\mathbf{u}} C_1(\mathbf{u}) \log_q C_2(\mathbf{u}) d\mathbf{u}. \quad (5.34)$$

受前述 CCKL [27] 启发，他们还定义了累积 Copula Tsallis 散度，如下

**定义 24** (Cumulative Copula Tsallis Divergence). 给定两个连续多元随机变量  $\mathbf{X}_1, \mathbf{X}_2 \in R^n$  及其 Copula 函数  $C_1(\mathbf{u})$  和  $C_2(\mathbf{u})$ ，则累积 Copula Tsallis 散度定义为

$$D_{CCTD}(C_1, C_2) = \int_{\mathbf{u}} C_1(\mathbf{u}) \log_q \frac{C_1(\mathbf{u})}{C_2(\mathbf{u})} d\mathbf{u} - \frac{\rho_k(C_1) - \rho_k(C_2)}{2^k n(k)}, \quad (5.35)$$

其中  $\rho_k(C) = n(k)(2^k \int_{\mathbf{u}} C(\mathbf{u}) d\mathbf{u} - 1)$  为  $k$  维 Spearman 相关系数， $n(k) = \frac{k+1}{2^k - k - 1}$ 。

很容易得知

$$\lim_{q \rightarrow 1} D_{CCTD}(C_1, C_2) = D_{CCKL}(C_1, C_2). \quad (5.36)$$

基于 KL 散度和 MI 的关系，他们在 CCTD 的基础上进一步定义了累积 Copula Tsallis MI 的概念，如下

**定义 25** (Cumulative Copula Tsallis Mutual Information). 给定多元随机变量  $\mathbf{X}$  及其 Copula 函数  $C(\mathbf{u})$  及其相应的独立 Copula 函数  $\Pi(\mathbf{u}) = \prod_{i=1}^k u_i$ ，则累积 Copula Tsallis MI 定义为

$$I_{CCTMI}(C) = D_{CCTD}(C, \Pi) = \int_{\mathbf{u}} C(\mathbf{u}) \log_q \frac{C(\mathbf{u})}{\Pi(\mathbf{u})} d\mathbf{u} - \frac{\rho_k(C)}{2^k n(k)}, \quad (5.37)$$

其中  $\rho_k(C) = n(k)(2^k \int_{\mathbf{u}} C(\mathbf{u}) d\mathbf{u} - 1)$  为  $k$  维 Spearman 相关系数， $n(k) = \frac{k+1}{2^k - k - 1}$ 。

## 5.6 Copula Rényi 熵

Rényi 熵 (Rényi Entropy) 是一类重要的熵定义，由 Alfréd Rényi 在 1961 年提出 [581]。Rényi 熵给出了满足可加性的信息度量的最一般形式，Hartley 熵、香农熵、collision 熵和 min-entropy 等都是其特例 [582, 583]。

**定义 26** (Rényi Entropy). 给定多元随机变量  $\mathbf{X}$ ，则 Rényi 熵定义为

$$H_R(\mathbf{x}) = \frac{1}{1-\alpha} \log \int_{\mathbf{x}} p^\alpha(\mathbf{x}) d\mathbf{x}, \quad (5.38)$$

其中  $\alpha > 0, \alpha \neq 1$ 。

当  $\alpha \rightarrow 1$  时, Rényi 熵退化为香农熵, 即

$$\lim_{\alpha \rightarrow 1} H_R(\mathbf{x}) = H(\mathbf{x}). \quad (5.39)$$

类似地, 可以定义 Copula Rényi 熵如以下形式:

**定义 27** (Copula Rényi Entropy). 给定多元随机变量  $\mathbf{X}$  及其 copula 密度函数  $c(\mathbf{u})$ , 则 Copula Rényi 熵定义为

$$H_{CR}(\mathbf{x}) = \frac{1}{1-\alpha} \log \int_{\mathbf{u}} c^\alpha(\mathbf{u}) d\mathbf{u}, \quad (5.40)$$

其中  $\alpha > 0, \alpha \neq 1$ .

当  $\alpha \rightarrow 1$  时, Copula Rényi 熵退化为 CE, 即

$$\lim_{\alpha \rightarrow 1} H_{CR}(\mathbf{x}) = H_c(\mathbf{x}). \quad (5.41)$$

Saha 和 Kayal [29] 基于 Copula 函数扩展了 Rényi 熵, 定义了累积 Copula Rényi 熵 (Cumulative Copula Rényi Entropy: CCRE) 和生存 Copula Rényi 熵 (Survival Copula Rényi Entropy: SCRE), 如下

**定义 28** (Cumulative Copula Rényi Entropy). 给定多元随机变量  $\mathbf{X}$  及其累积 Copula 函数  $C(\mathbf{u})$ , 则多变量累积 Copula Rényi 熵的定义为

$$H_{CCRE}(\mathbf{x}) = \frac{1}{1-\alpha} \log \int_{\mathbf{u}} C^\alpha(\mathbf{u}) d\mathbf{u}, \quad (5.42)$$

其中  $\alpha > 0, \alpha \neq 1$ .

**定义 29** (Survival Copula Rényi Entropy). 给定多元随机变量  $\mathbf{X}$  及其生存 Copula 函数  $\bar{C}(\mathbf{u})$ , 则多变量生存 Copula Rényi 熵的定义为

$$H_{SCRE}(\mathbf{x}) = \frac{1}{1-\alpha} \log \int_{\mathbf{u}} \bar{C}^\alpha(\mathbf{u}) d\mathbf{u}, \quad (5.43)$$

其中  $\alpha > 0, \alpha \neq 1$ .

当  $\alpha > 1$  时,  $H_{CCRE}(\mathbf{x}) > 0$ ; 当  $0 < \alpha < 1$  时,  $H_{CCRE}(\mathbf{x}) < 0$ .

当  $\alpha > 1$  时,  $H_{SCRE}(\mathbf{x}) \geq 0$ .

当  $\alpha \rightarrow 1$  时, CCRE 和 SCRE 分别退化为 CCE 和 SCE, 即

$$\lim_{\alpha \rightarrow 1} H_{CCRE}(\mathbf{x}) = H_{CCE}(\mathbf{x}) \quad (5.44)$$

和

$$\lim_{\alpha \rightarrow 1} H_{SCRE}(\mathbf{x}) = H_{SCE}(\mathbf{x}). \quad (5.45)$$

他们还定义了累积 Copula Rényi 不准确度的概念, 如下

**定义 30** (Cumulative Copula Rényi Inaccuracy). 给定多元随机变量  $\mathbf{X}, \mathbf{Y} \in R^n$  以及相应的 Copula 函数  $C_{\mathbf{X}}(\mathbf{u}), C_{\mathbf{Y}}(\mathbf{u})$  和边缘函数  $F_i(x_i), G_i(y_i), i = 1, \dots, n$ , 则累积 Copula Rényi 不准确度定义为

$$D_{CCRI}(C_{\mathbf{X}}, C_{\mathbf{Y}}) = \frac{1}{1-\alpha} \log \int_{\mathbf{u}} C_{\mathbf{X}}(\mathbf{u}) \{C_{\mathbf{Y}}(G_i(F_i^-(u_i)))\}^{\alpha-1} d\mathbf{u}, \quad (5.46)$$

其中  $\alpha > 0, \alpha \neq 1$ .

当  $C_{\mathbf{X}} = C_{\mathbf{Y}}$  时,  $D_{CCRI} = H_{CCRE}$ .

类似地, 他们还定义了生存 Copula Rényi 不准确度的概念, 如下

**定义 31** (Survival Copula Rényi Inaccuracy). 给定多元随机变量  $\mathbf{X}, \mathbf{Y} \in R^n$  以及相应的生存 Copula 函数  $\bar{C}_{\mathbf{X}}(\mathbf{u}), \bar{C}_{\mathbf{Y}}(\mathbf{u})$  和边缘函数  $\bar{F}_i(x_i), \bar{G}_i(y_i), i = 1, \dots, n$ , 则生存 Copula Rényi 不准确度定义为

$$D_{SCRI}(\bar{C}_{\mathbf{X}}, \bar{C}_{\mathbf{Y}}) = \frac{1}{1-\alpha} \log \int_{\mathbf{u}} \bar{C}_{\mathbf{X}}(\mathbf{u}) \{\bar{C}_{\mathbf{Y}}(\bar{G}_i(\bar{F}_i^-(u_i)))\}^{\alpha-1} d\mathbf{u}, \quad (5.47)$$

其中  $\alpha > 0, \alpha \neq 1$ .

当  $\bar{C}_{\mathbf{X}} = \bar{C}_{\mathbf{Y}}$  时,  $D_{SCRI} = H_{SCRE}$ .

## 5.7 Copula Rényi 散度和 Copula Tsallis 散度

散度 (Divergence) 是概率论和信息论等领域的重要概念, 其中最著名的是 Kullback-Leibler (KL) 散度 [584], 也称相对熵 (Relative Entropy)。KL 散度度量两个概率密度之间的“距离”, 其数学定义如下

**定义 32** (Kullback-Leibler Divergence). 给定两个概率密度函数  $p_X, p_Y$ , 则 KL 散度定义为

$$D_{KL}(p_X, p_Y) = \int_x p_X(x) \log \frac{p_X(x)}{p_Y(x)} dx. \quad (5.48)$$

对于任意满足条件的  $p_X, p_Y$ ,  $D_{KL}(p_X, p_Y) \geq 0$ ; 当  $p_X = p_Y$  时,  $D_{KL}(p_X, p_Y) = 0$ 。 $D_{KL}$  不具有对称性, 即  $D_{KL}(p_X, p_Y) \neq D_{KL}(p_Y, p_X)$ , 因此其不是真正意义的距离。

受 KL 散度启发, 学者又定义了 Rényi 散度 (Rényi Divergence) [581] 和 Tsallis 散度 (Tsallis Divergence) [585, 586], 分别如下:

**定义 33** (Rényi Divergence). 给定两个概率密度函数  $p_X, p_Y$ , 则 Rényi 散度的定义为

$$D_R(p_X, p_Y) = \frac{1}{\alpha-1} \log \int_x p_X^\alpha(x) p_Y^{1-\alpha}(x) dx, \quad (5.49)$$

其中  $\alpha > 0, \alpha \neq 1$ .

**定义 34** (Tsallis Divergence). 给定两个概率密度函数  $p_X, p_Y$ , 则 Tsallis 散度的定义为

$$D_T(p_X, p_Y) = \frac{1}{q-1} \left( \int_x p_X^q(x) p_Y^{1-q}(x) dx - 1 \right), \quad (5.50)$$

其中  $q > 0, q \neq 1$ .

KL 散度是 Rényi 散度和 Tsallis 散度的特殊情况 [581, 585]，即

$$\lim_{\alpha \rightarrow 1} D_R(p_X, p_Y) = \lim_{q \rightarrow 1} D_T(p_X, p_Y) = D_{KL}(p_X, p_Y). \quad (5.51)$$

给定随机变量  $(X, Y)$ ，则 KL 散度与 MI 和 CE 有如下关系：

$$D_{KL}(p_{XY}, p_X p_Y) = I(x, y) = -H_c(x, y). \quad (5.52)$$

在此 KL 散度与 CE 的理论联系的基础上，Mohammadi 和 Emadi [30] 基于 Copula 函数对 Rényi 散度和 Tsallis 散度概念进行了扩展，得到  $D_R(p_{XY}, p_X p_Y)$  和  $D_T(p_{XY}, p_X p_Y)$  分别对应的 Copula Rényi 散度和 Copula Tsallis 散度概念，其定义分别如下

**定义 35** (Copula Rényi Divergence). 给定随机变量  $(X, Y)$  以及其 copula 密度函数  $c(u, v)$ ，则 Copula Rényi 散度定义为

$$D_{CR}(c_{XY}) = \frac{1}{\alpha - 1} \log \int_u \int_v c^\alpha(u, v) dudv, \quad (5.53)$$

其中  $\alpha > 0, \alpha \neq 1$ 。

**定义 36** (Copula Tsallis Divergence). 给定随机变量  $(X, Y)$  以及其 copula 密度函数  $c(u, v)$ ，则 Copula Tsallis 散度定义为

$$D_{CT}(c_{XY}) = \frac{1}{q - 1} \left( \int_u \int_v c^q(u, v) dudv - 1 \right), \quad (5.54)$$

其中  $q > 0, q \neq 1$ 。

他们给出了这两种散度之间的关系，如下

$$D_{CT}(c_{XY}) = \frac{1}{\alpha - 1} (e^{(\alpha - 1)D_{CR}(c_{XY})} - 1). \quad (5.55)$$

他们证明了以下结论：

1.  $D_{CR}(c_{XY}) \geq 0, D_{CT}(c_{XY}) \geq 0$ ;
2. 当  $X, Y$  相互独立时， $D_{CR}(c_{XY}) = D_{CT}(c_{XY}) = 0$ ；以及
3.  $D_{CR}(c_{XY})$  和  $D_{CT}(c_{XY})$  具有单调变换不变性。

他们给出了 Copula Rényi 散度和 Copula Tsallis 散度的非参数估计算法，并证明了估计算法的渐进一致性。

这两种 Copula 散度分别与 Copula Rényi 熵和 Tsallis Copula 熵具有等价关系。结合式(5.40)和式(5.53)，可知

$$D_{CR}(c_{XY}) = -H_{CR}(x, y). \quad (5.56)$$

同理，结合式(5.4)和式(5.54)，可知

$$D_{CT}(c_{XY}) = -H_{TC}(x, y). \quad (5.57)$$

## 5.8 Copula Jeffreys 散度和 Copula Hellinger 散度

KL 散度的定义不具有对称性，而 Jeffreys 散度 [587] 和 Hellinger 散度 [588] 则是信息论中两种满足对称性的散度定义，它们同属于  $f$  散度族 [589]。

**定义 37** (Jeffreys Divergence). 给定两个概率密度函数  $p_X, p_Y$ ，则 Jeffreys 散度定义为

$$D_J(p_X, p_Y) = \int_x (p_X(x) - p_Y(x))(\log p_X(x) - \log p_Y(x))dx. \quad (5.58)$$

**定义 38** (Hellinger Divergence). 给定两个概率密度函数  $p_X, p_Y$ ，则 Hellinger 散度定义为

$$D_H(p_X, p_Y) = \int_x (\sqrt{p_X(x)} - \sqrt{p_Y(x)})^2 dx. \quad (5.59)$$

我们知道，CE 与 KL 散度有如下对应关系：

$$D_{KL}(p_{XY}, p_X p_Y) = -H_c(x, y). \quad (5.60)$$

受此启发，Mohammadi 等 [31] 给出了  $D_J(p_{XY}, p_X p_Y)$  和  $D_H(p_{XY}, p_X p_Y)$  分别对应的 Copula Jeffreys 散度和 Copula Hellinger 散度概念，其定义分别如下

**定义 39** (Copula Jeffreys Divergence). 给定随机变量  $(X, Y)$ ，以及其 copula 密度函数  $c(u, v)$ ，则 Copula Jeffreys 散度定义为

$$D_{CJ}(c_{XY}) = \int_u \int_v (c(u, v) - 1) \log c(u, v) dudv. \quad (5.61)$$

**定义 40** (Copula Hellinger Divergence). 给定随机变量  $(X, Y)$ ，以及其 copula 密度函数  $c(u, v)$ ，则 Copula Hellinger 散度定义为

$$D_{CH}(c_{XY}) = \int_u \int_v (\sqrt{c(u, v)} - 1)^2 dudv. \quad (5.62)$$

作者证明了这两种散度的单调变换不变性，给出了它们高斯分布下的情况，并给出了基于 Copula 密度估计的估计方法。

## 5.9 Copula 分形不准确度

不准确度 (Inaccuracy) [590] 是一种度量一个分布和另一个分布之间散度或偏差的数学概念，其定义为

**定义 41** (Inaccuracy). 给定两个连续多元随机变量  $\mathbf{X}, \mathbf{Y} \in R^n$  及其密度函数  $p_X, p_Y$ ，则不准确度定义为

$$D_I(p_X, p_Y) = - \int_{\mathbf{x}} p_X(\mathbf{x}) \log p_Y(\mathbf{x}) d\mathbf{x}. \quad (5.63)$$

不准确度是香农熵的一种泛化，且香农熵为其最小值，即  $D_I(p_{\mathbf{X}}, p_{\mathbf{Y}}) \geq H(\mathbf{x})$ 。显然，当  $p_{\mathbf{X}} = p_{\mathbf{Y}}$  时， $D_I(p_{\mathbf{X}}, p_{\mathbf{Y}}) = H(\mathbf{x})$ 。

不准确度与 KL 散度具有如下关系：

$$D_{KL}(p_{\mathbf{X}}, p_{\mathbf{Y}}) = D_I(p_{\mathbf{X}}, p_{\mathbf{Y}}) - H(\mathbf{x}). \quad (5.64)$$

在此关系的基础上，不准确度被用作一些优化问题的目标函数，此时其又被称为交叉熵（Cross Entropy）[591]。

通过将不准确度定义式(5.63)中的密度函数  $p_{\mathbf{X}}, p_{\mathbf{Y}}$  换为 Copula 密度函数  $c_{\mathbf{X}}, c_{\mathbf{Y}}$ ，我们可以定义 Copula 不准确度的概念，如下

**定义 42** (Copula Inaccuracy). 给定两个连续多元随机变量  $\mathbf{X}, \mathbf{Y} \in R^n$  及其 Copula 密度函数  $c_{\mathbf{X}}(\mathbf{u}), c_{\mathbf{Y}}(\mathbf{u})$ ，则 Copula 不准确度定义为

$$D_{CI}(c_{\mathbf{X}}, c_{\mathbf{Y}}) = \int_{\mathbf{u}} c_{\mathbf{X}}(\mathbf{u}) \log c_{\mathbf{Y}}(\mathbf{u}) d\mathbf{u}. \quad (5.65)$$

根据此定义，当  $c_{\mathbf{X}} = c_{\mathbf{Y}}$  时，Copula 不准确度就变为 CE，即  $D_{CI} = H_c$ 。

通过将不准确度定义式(5.63)中的密度函数  $p_{\mathbf{X}}, p_{\mathbf{Y}}$  换为累积 Copula 函数  $C_{\mathbf{X}}, C_{\mathbf{Y}}$ ，我们可以定义累积 Copula 不准确度的概念，如下

**定义 43** (Cumulative Copula Inaccuracy). 给定两个连续多元随机变量  $\mathbf{X}, \mathbf{Y} \in R^n$  及其累积 Copula 函数  $C_{\mathbf{X}}(\mathbf{u}), C_{\mathbf{Y}}(\mathbf{u})$ ，则累积 Copula 不准确度定义为

$$D_{CCI}(C_{\mathbf{X}}, C_{\mathbf{Y}}) = \int_{\mathbf{u}} C_{\mathbf{X}}(\mathbf{u}) \log C_{\mathbf{Y}}(\mathbf{u}) d\mathbf{u}. \quad (5.66)$$

根据此定义，当  $C_{\mathbf{X}} = C_{\mathbf{Y}}$  时，累积 Copula 不准确度就变为累积 Copula 熵，即  $D_{CCI} = H_{CCCE}$ 。

前面已经介绍了 Zachariah 等 [28] 提出的 CCTI 概念，以及 Saha 和 Kayal [29] 提出的 CCRI 和 SCRI 的概念。很容易证明，

$$\lim_{q \rightarrow 1} D_{CCTI}(C_{\mathbf{X}}, C_{\mathbf{Y}}) = \lim_{\alpha \rightarrow 1} D_{CCRI}(C_{\mathbf{X}}, C_{\mathbf{Y}}) = D_{CCI}(C_{\mathbf{X}}, C_{\mathbf{Y}}). \quad (5.67)$$

受这些工作启发，Pandey 和 Kundu [32] 提出了累积 Copula 分形不准确度 (Cumulative Copula Fractional Inaccuracy: CCFI) 的概念，定义如下

**定义 44** (Cumulative Copula Fractional Inaccuracy). 给定两个连续多元随机变量  $\mathbf{X}, \mathbf{Y} \in R^n$  及其累积 Copula 函数  $C_{\mathbf{X}}(\mathbf{u}), C_{\mathbf{Y}}(\mathbf{u})$  和边缘累积函数  $F_i(x_i), G_i(y_i), i = 1, \dots, n$ ，则多变量累积 Copula 分形不准确度定义为

$$D_{CCFI}(C_{\mathbf{X}}, C_{\mathbf{Y}}) = \int_{\mathbf{u}} C_{\mathbf{X}}(\mathbf{u}) [ -\log C_{\mathbf{Y}}(G_i(F_i^-(u_i))) ]^r d\mathbf{u}, \quad (5.68)$$

其中  $0 < r < 1$ 。

易知, 当  $C_X = C_Y$  时,  $D_{CCFI} = H_{FCCE}$ 。

通过将累积 Copula 函数替换成生存 Copula 函数, 他们又由 CCFI 定义了生存 Copula 分形不准确度 (Survival Copula Fractional Inaccuracy: SCFI) 的概念, 如下

**定义 45** (Survival Copula Fractional Inaccuracy). 给定两个连续多元随机变量  $\mathbf{X}, \mathbf{Y} \in R^n$  及其生存 Copula 函数  $\bar{C}_X(\mathbf{u}), \bar{C}_Y(\mathbf{u})$  和边缘生存函数  $\bar{F}_i(x_i), \bar{G}_i(y_i), i = 1, \dots, n$ , 则多变量生存 Copula 分形不准确度定义为

$$D_{SCFI}(\bar{C}_X, \bar{C}_Y) = \int_{\mathbf{u}} \bar{C}_X(\mathbf{u}) [ -\log \bar{C}_Y(\bar{G}_i(\bar{F}_i^-(u_i))) ]^r d\mathbf{u}, \quad (5.69)$$

其中  $0 < r < 1$ 。

他们还证明, 如果  $\mathbf{X}, \mathbf{Y}$  有共同的支撑且径向对称, 则有

$$D_{CCFI}(C_X, C_Y) = D_{SCFI}(\bar{C}_X, \bar{C}_Y). \quad (5.70)$$

# 第六章 实际应用

## 6.1 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 [35] 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。

## 6.2 天体物理学

红移是宇宙星体的最重要信息之一，因其反映了星体距离地球的宇宙距离，可以用于研究星系演化和宇宙学。测光红移是一种从宇宙星体光学观测估计其红移的方法。由于光学观测相比于光谱观测更易于施行，因此测光红移是现代天文学巡天观测的主要手段之一，一般在获得测光红移信息后再对感兴趣的星体进行光谱观测。机器学习方法已经成为构建测光红移预测模型的主要方法之一，但其预测准确度仍有待提高。马健 [36] 提出利用基于 CE 的变量选择方法构建此类估计模型，以提高预测模型的准确度。该方法首先估计光学观测和红移之间的 CE 作为观测变量的重要性度量，再将重要的观测变量作为模型的输入来预测红移。他将方法应用于斯隆巡天类星体观测数据，结果表明，利用 CE 选择后得到的模型的准确度要高于未经过选择的模型，特别是在高红移 ( $z > 4$ ) 的星体上，预测准确度得到了明显提升<sup>1</sup>。同时，方法也选择出了具有可解释性的一组光学观测变量，包括光度星等、紫外波段亮度和标准差、和其他四个波段亮度等，为进一步天体物理研究和光学观测仪器设计提供了科学证据。

## 6.3 空间科学

电离层是指大气层从 50 公里到 1000 公里高度由大量离子和自由电子构成的区域。它的物理特性受太阳辐射活动的影响，同时也影响着通过其的微波传播。电子总量（Total Electron

---

<sup>1</sup>实验代码：<https://github.com/majianthu/quasar>

Content: TEC) 是电离层的主要物理参数, 对空间运行的卫星通讯系统和全球定位系统具有重要影响。多普勒频移 (Doppler Frequency Shift: DFS) 是检测电离层的技术之一, 可以从通过传输信号的能量谱估计得到, 它被研究证明可以反映电离层的短期变化。研究 TEC 和 DFS 之间的复杂关系, 可以加深对电离层特性变化了解, 对相关空间系统运行意义重大。但这 e 种关系具有非线性复杂性, 给研究构成了挑战。TE 作为一种研究非线性因果关系的数学工具, 是研究这种关系的理想方法。Akerele 等 [37] 利用一组分析工具研究了赤道地区高频微波信号传输中 TEC 和 DFS 之间的非线性关系, 其中利用了基于 CE 的 TE 作为因果关系分析工具。他们利用 2020-2021 年尼日利亚 Abuja 和 Lagos 两地的高频多普勒系统测量数据和 NASA 的 TEC 数据, 分析了上述关系。结果发现, TE 分析结果显示 TEC 的变化会导致 DFS 的显著变化, 这一因果关系的强度会随着电离层季节变化而变化, 而相关性分析并没有揭示这一因果关系。这一结论对高频微波通讯系统的管理具有重要参考价值。

## 6.4 地质学

岩相是地质学的基本概念, 指在特定沉积环境中形成的具有相同特性的岩石类型。岩相分类是地质调查和勘探中的重要问题, 对于岩相形成评价和储量识别等任务具有重要意义。随着地质数据的增加, 利用机器学习方法对岩相进行分类成为了重要课题, 得到了越来越多的研究。但已有的研究大多直接利用机器学习方法构建分类模型, 未对模型进行选择, 导致模型有待改进; 同时, 由于采用的分类器基本都是黑箱模型, 缺乏可解释性, 使得得到的模型难以为地质学家所理解。Ma [38] 提出利用 CE 构建岩相分类模型, 首先利用 CE 来计算地质变量与岩相之间的相关性, 再基于相关性强度对变量进行选择来得到最终的分类模型。他将该方法应用于被广泛研究的美国堪萨斯地质调查得到的测井岩相分类数据, 对地质变量进行选择后构建岩相分类模型。实验结果表明, 该方法可以选择更少的地质变量作为分类模型输入, 同时不降低分类性能。另外, 该方法将海相标注和测井深度等与岩相密切相关的变量选择认定为重要变量, 使得这样得到的分类模型符合地质学知识, 具有可解释性。

## 6.5 地球物理学

土地干湿度是土地表面水分和能量动态交互过程的属性, 传统的干湿度度量大多使用气候条件变量的长期均值来进行计算, 难以反映短期的地表水分-能量交互。蒸散是表征短期地表的水汽-能量交互过程的关键变量, 包括土地和植物表面的水分散发, 传统上根据土壤湿度和能量供应将其分成水分驱动、能量驱动和过渡型三种概念框架。研究表明, 蒸散-土壤湿度关系也受其他因素影响, 如云层、风速和植被等, 考虑这些因素如何影响蒸散为开发新的土地干湿度分类框架提供了可能性。Shan 等 [39, 40] 通过考虑土地-大气的短期耦合效应, 提出了一种新的刻画土地干湿度的方法。该方法利用基于 CE 的条件互信息分别计算蒸散与土壤湿度和太阳辐射的因果关系强度, 再利用这两种因果关系的差值将土地干湿度分为 6 种类型, 分别对应到三种蒸散概念框架。他基于 1990-2020 年夏季中国大陆的气温、露点温度、土壤湿度、潜在热流、敏感热流、蒸散和地表太阳辐射等逐小时记录数据, 利用该方法得到了土地干湿度空间分布图, 并与联

合国环境规划署的干旱度指数进行了对比，发现该方法计算得到的条件互信息分布图与水分和能量的地理分布相符，由此得到的干旱度分布能够更精确地捕捉短期地表过程，因此提供了一种短期土地-大气交互过程的有价值的补充信息。该方法加深了对气候干旱特征的理解，提供了一种对极端热浪和骤发干旱等短期气候变化具有敏感度的表征工具方法。

陆气耦合 (Land-Atmosphere Coupling) 是指地表和大气边界层之间的湍流交换过程，导致了多时空尺度的能量和物质循环，也对极端天气的发生产生影响。引入陆气耦合可以提高数值天气预报和气候模型的预测能力。理解冬季稳定边界层的陆气耦合变弱现象，并在预报模型表示边界层网格尺度过程是一个值得关注的问题。低温天气下湍流变弱导致的解耦现象，使得预报模型中湍流参数化的传统相似度理论失效，进而使得近地温度建模不准确。2 米温度是近地大气的最常用预报变量，建立 2 米温度预测模型被认为是地表通量参数化开发的关键步骤。基于挪威芬瑟积雪覆盖山地地区的气象观测记录，Mack 等 [41] 提出了一种基于 Copula 贝叶斯网络的 2 米温度插值模型，用于替代数值天气预报系统中的传统模型。他们首先利用 CE 分析了芬瑟两个观测站的气象观测变量之间 CE 和解耦度量之间的关系，并对比了实际观测和传统 2 米温度模型分别对应的 2 米温度和 10 米温度之间的 CE 值与解耦度量之间的关系。分析发现，随着垂直解耦程度增加，两个站点之间的 CE 增加，称为“信息解耦”；同时，相较于观测，传统模型 2 米温度和 10 米温度之间的 CE 随着解耦度量的减小而增加的幅度更大，表明传统模型未能充分利用有效观测信息。他们利用藤 Copula 和 Copula 贝叶斯网络相结合的方法建立了地表和大气温度耦合的模型用以预测近地温度，获得了比传统模型更优良的性能，验证了 CE 分析的合理性。作者认为，利用 CE 理论进行耦合不确定量化是一个新的概念，同时利用 CE 计算模型和观测之间的信息损失也可以作为一个模型预测性能评估的新指标。

## 6.6 流体力学

离心泵是一个在多种工程场合广泛采用的机械设备，预测其性能是设计和使用过程中关心的重要问题。传统的基于流体力学数值模拟的离心泵性能预测方法需要大量的计算时间，过程相对复杂且预测误差大。利用神经网络预测离心泵的性能已经成为了一个重要的新方法，但此类方法需要大量的样本数据，而实际工程中往往数据有限。Chen 等 [42] 提出利用 SMOGN 过采样技术和 GAKHO 超参数调节技术相结合的方法来进行离心泵性能预测，其中 SMOGN 技术用于对小样本重采样预处理，GAKHO 方法用于调节神经网络的超参数。在此方法中，CE 等工具被用于验证 SMOGN 重采样技术的效果。他们利用 10 种离心泵不同速度下的数据验证了该方法，发现该方法能够在小样本的情况下对离心泵的扬程和效率等主要性能参数进行准确预测。

## 6.7 热学

微热管是一种高效的相变传热元器件，具有高导热率、尺寸小和传热功率大等特点，作为现代电子设备的散热单元得到广泛应用。微热管形状各异，核心部件吸液芯种类多，器件的几何构造和制造过程的工艺参数是影响其传热性能的两个主要因素。因此，研究微热管的几何构造参数和制造过程的工艺参数与其传热性能之间关系是一个重要的问题，可以用于指导微热管的结构

设计和制造工艺参数设定。传统的设计和制造方法依赖于人的经验和重复试验，成本高且效率较低。利用机器学习模型来建立设计需求与几何结构和工艺参数之间关系是一个新的解决路径。李勇等 [43, 44] 提出了一种基于 CE 和神经网络模型相结合的微热管设计和制造参数预测方法，利用 CE 度量微热管结构设计需求参数与其传热功率之间的关联强度，以选择关键设计参数作为模型输入，再用一维卷积残差网络建立选择的关键设计需求参数到结构参数和工艺参数的预测模型。他们基于实际生产制造的微热管样品的测试数据验证了该方法，结果表明，CE 选择的与传热功率相关的关键设计参数得到了实际实验结果的验证，基于 CE 选择参数的预测模型在实际制造的测试设计需求数据上的参数预测结果与真实值十分接近，最大误差仅为 2.6%。

循环流化床锅炉 (Circulating Fluidized Bed Boiler: CFBB) 是采用流态化燃烧的洁净煤燃烧技术，具有适应多种燃料、负荷调节能力好等优点，普遍用于热力发电厂等工业领域。但如何进一步提高 CFBB 的热效率并减少排放是行业关心的问题，但这两个目标之间存在冲突，无法同时达成。Ma 等 [45] 将此问题归结为多目标参数优化问题，提出了一种结合 CE、加权融合树模型和多目标树结构估计器的参数优化方法，其中 CE 用于分析燃烧过程参数与优化目标（氮氧化物排放和热效率）之间的函数关系，在此分析结果的基础上，利用加权融合树模型构建优化目标预测模型，最后，利用多目标树结构估计器来优化调节预测模型输入参数来调节优化目标。他们将该方法应用于一个 330MW 的 CFBB，CE 分析发现污染物排放和热效率都主要与煤和空气的供应有关，最后的优化结果也表明减少煤供应同时增加空气供应可以平衡调节锅炉的热效率和氮氧化物排放浓度。此技术方法可以很容易扩展到其他类型的工业锅炉上，从而为工业热力系统的燃烧控制提供了一致的解决方案。

## 6.8 理论化学

变构效应 (Allostery) 被认为“生命的第二秘密”，是普遍存在于几乎所有蛋白质的生命现象。它是指变构调节分子与蛋白质结合，诱导结合位点以外的远点发生变化的调节效应。最常见的变构系统模型是变构二状态模型，描述了变构过程的热力学循环。此类模型假设了受体活化是二状态过程，这与 NMR 实验揭示的多模态过程不相符合。深入理解配体诱导的受体活化的分子机制需要构建新的理论来理解配体结合点和激活点之间的热力学耦合关系。Cuendet 等 [46] 提出了一种新的理论，称为变构景观 (Allostery Landscape)，定义了热力学耦合函数来量化生物分子系统中的热力学耦合。他们指出新函数与 copula 密度函数和 CE 有密切联系，CE 定义了变构系统的信息传输属性，即配体结合点和激活点之间的信息传输。他们将新理论应用到丙氨酸二肽的 N 端和 C 端的热力学耦合分析中。

## 6.9 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 [47]

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 [592]，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。

## 6.10 材料学

耐热型含能材料是指具有高能量和高热稳定性的特殊材料，可以在高温的环境下保持稳定性，因此是国防、航空航天和地质勘探等重点领域关键性材料，如宇航和高超音速武器的推进燃料、深井钻探的炸药等。但此类材料数量稀少且实验研究具有极高危险性，因此设计此类材料是材料学家们一直努力攻克的挑战性难题。“从头设计”含能材料需要经历“设计-筛选-评估”的流程，其中采用机器学习的方法构建材料结构-性质预测模型对设计的分子性质进行预测是材料分子筛选的关键步骤 [593]。传统的含能分子性质预测模型构建过程只采用了与热稳定性线性相关的分子特征，没有考虑与含能材料热分解温度具有非线性关系的因素，如晶体结构和堆积方式等。田杰等 [48, 49] 提出了一种结合皮尔逊相关系数和 CE 的特征选择方法，从分子拓扑结构和量子化学计算特征中选择与热分解温度具有相关性的特征，并构建预测模型。其中，CE 方法的引入是为了筛选和热分解温度具有非线性关系的特征。他收集了 460 个含能化合物，并生成了包含 286 个特征的数据集，并应用该方法筛选得到了 87 个特征，再将筛选的特征做为随机森林和 SVM 等模型的输入以预测化合物的热分解温度，最终得到了较传统方法更好的预测效果，交叉验证实验的预测误差控制在了  $28.5^{\circ}\text{C}$ 。他们将方法应用于自己设计的分子生成器生成的分子，最终筛选出 16 个具有良好热稳定潜能且爆轰能力很强的含能分子，验证了方法的实用价值。

近  $\beta$  钛合金具有低密度、高强度、高韧性、耐腐蚀等特点，是航空航天领域器件制造的关键材料之一。合金组件属性决定于微观结构和结晶材质，而材质与加工过程的形变模式高度相关。研究形变加工对晶体材质的影响是钛合金研究的关键问题之一。计算仿真技术是实物实验以外材料加工研究的新手段，如晶体可塑性等数值方法。Zhang 等 [50] 研究了近  $\beta$  钛合金在  $\beta$  区形变机制和材质演化之间的关系，利用实验仿真了  $\beta$  区形变的 8 种钛合金主要成分在三种滑动形变模式下的演化，利用 CE 计算了每种形变模式与各个成分含量之间的非线性相关性强度。结果显示，三种模式与除  $\{001\}<100>$  成分外的其他成分高度相关，与已有研究结论相符合； $\gamma$  纤维和  $\{110\}<001>$  材质表现出较低的相关性；同时， $\{112\}$  和  $\{123\}$  模式较  $\{110\}$  模式具有更强的相关性，表明二者在高温形变过程中起决定性作用。

难熔多元合金（Refractory Multi-Principal Element Alloys: RMPEAs）是一类由多种难熔金属元素组成的近等摩尔混合物材料，因其表现出异常的高温强度和抗高温软化性能，被认为是极具潜力的高温应用候选材料之一。在特定应用中发挥 RMPEAs 的潜力需要准确预测其物理属性，但其巨大的组合空间给预测带来了挑战。传统的密度泛函理论方法应用到 RMPEAs 需要较高的计算量，增加了材料选择的难度，而 EMTO (Exact Muffin-Tin Orbitals) 方法显著提高了

计算效率。在传统密度泛函理论方法的基础上构建机器学习模型以预测合金的属性成为了一个广泛采用的技术路线，可以加速材料设计过程。但如何分析合金特性和物理属性之间的非线性关系，进而确定机器学习模型是一个关键的问题。Jin 等 [51] 提出将 EMTO 第一性原理计算与基于 CE 的特征选择方法相结合，构建 RMPEAs 物理属性预测模型。他们用 V-Nb-Ta 系统验证了该方法的可靠性，得到的机器学习模型预测与实验验证结果高度一致。其中，他们基于 CE 方法选择了 8 个特征作为模型输入，CE 方法分析得到的熔点和  $\Omega$  准则之间关系，以及这二者与 V、Nb 和 Ta 含量之间关系，都与理论关系一致，说明了 CE 方法的合理性。

## 6.11 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 [52-54] 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。Li 等 [55] 基于 CE 和机器学习方法研究了长江上游的月径流预报问题。他们利用 130 个全球环流指数、7 个气象因子和高场和寸滩两个水文站的月径流量数据，采用 CE 等 3 种变量选择方法和 5 种机器学习模型进行组合构建预测模型。结果表明，CE 和 LSTM 组合在高场站获得了最优预测性能，而随机森林和 CE 组合在寸滩站获得了满意性能。Mo 等 [56] 提出了一种长期径流预报模型框架，结合了 CE、LSTM 和 GARCH 三种方法，其中 CE 用于筛选与径流有关的预报因子。与传统方法相比，CE 更适合因子间具有交互关联的复杂情况。他们将方法应用于洪泽湖和骆马湖的径流预报研究，结果表明，与传统方法相比，该框架中的 CE 方法不仅成功辨别了因子间的交互效应，同时还量化了每个预报期内各个因子的贡献度，从而选出了与预报有关的关键驱动因子，最终该方法框架得到了较对比方法更准确、更稳定且更可靠的预报结果。陈佳雷等 [57] 提出了一种时空图卷积网络的径流预报方法，首先构建流域内站点的拓扑结构图，再利用邻接矩阵表示地理相邻站点之间的时空依赖性，并利用 CE 等工具分析相邻关系、周期性和气象要素与径流量之间的时空相关关系，最后构造相应的带有注意力机制的图卷积网络做为径流预报模型。他们以金沙江流域为对象，验证了方法的有效性。汪胤 [58] 提出了基于 CE 和小波神经网络相结合的洪水预报方法，利用 CE 选择预报模型的输入，再利用小波神经网络构建预报模型。他将该方法应用于江苏省苏州市盛泽镇丁家坝站日水位的预测，实验结果表明，该方法的 3 个预报性能指标均高于对比方法，达到实际正式预报发布的水平，为区域防洪系统开发提供了技术支撑。

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 [59] 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该地区干旱发生的影响最大。Huang 和 Zhang [60] 利用 CE 方法分析了兰州地区 1957-2010 年的气象数据，以构建该地区的干旱预测模型，发现该地区的

风速、气温、水气压和相对湿度是与干旱最相关的气象因子。黄春艳 [61] 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。牛犇 [62] 利用 CE 等工具研究了黄河流域 9 个分区干旱传播的时空特征。他基于 1961-2020 年各个分区的气象、土壤湿度和径流数据，利用 CE 计算不同类型非平稳干旱指数之间的非线性相关关系，进而得到干旱响应时间尺度、干旱传播强度和干旱传播率等指标，最终发现了各分区上气象干旱、农业干旱和水文干旱之间传播敏感度和传播强度的强弱特征。Ni 等 [63] 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。Kanthavel 等 [64] 利用 CE 和藤 copula 等理论工具，提出了一种综合干旱指数，整合了标准化降雨指数、干旱监测指数、标准化土壤湿度指数和标准化径流干旱指数等四种指数，可以更好地同时反映相关水文气象变量和不同类型的干旱。CE 理论被用来衡量新指数与原始指数之间的相关性。他们将该指数应用于印度中部的达布蒂 (Tapti) 河流域的单月和四个月尺度的干旱研究中，验证了该指数的有效性，并揭示了该地区干旱的时空分布特点。Mohammadi 等 [31] 利用基于 copula 和 CE 理论的三种相关性度量估计方法，在伊朗三座城市（扎黑丹、恩泽利和马什哈德）1950-2017 年的水文观测数据的基础上，分析了三地的干旱变量（干旱强度、时长和时间间隔）之间的依赖关系。徐袁 [65] 提出利用基于 CE 的 TE 方法分析干旱传播的方向和时间延迟。他利用该方法对美国科罗拉多河流域气象站和水文站 1972-2019 年的实际观测数据进行了分析，首先构建了两种非平稳干旱指数（非平稳标准化降水指数和非平稳标准化径流指数），根据水系结构将该流域划分为 7 个子区域，再利用 GC 对同一分区的气象干旱和水文干旱之间传播的因果关系和同一干旱类型在不同分区之间传播的因果关系进行检验，最后利用 TE 估计相应的干旱传播时间，成功地发现了该流域内干旱传播的方向和时间延迟特征。

河冰是中高纬度地区河流特有的水文现象，具有鲜明的季节性特征。河冰物候学就是研究河冰冻结和融化的季节性变化现象的学问，这种季节性变化具有空间异质性和年度差异。河冰物候现象受多种因素影响，如气候条件、水文条件和人类活动等，其中气候变化是首要驱动因素。在全球气候变化的背景下，过去一个时期的河冰现象发生了显著变化，理解和量化河冰动态性以及内在的驱动力对于制定河流管理策略至关重要。Xing 等 [66] 研究了黑龙江的河冰物候学时间变化对气候变化的响应，其中利用 CE 工具计算了河冰周期的 5 个关键日期（新冰日、冰封日、冻结日、开化日和冰销日）与气候因子的非线性相关性强度。结果发现，温度、气压和蒸发压是河冰物候最重要的气候驱动因子。特别地，新冰日和冰封日与前一、二个月的蒸发压强相关，冻结日与蒸发压的相关性具有超过 4 个月的滞后期，而这三个日期又与同月的平均最大温度、平均温度和最小极端温度强相关，表明了结冰日期对温度的反应要快于其他气候因素。同时，开化日与 2 个月的月平均温差、最小极端气温和最大极端气温相关，冰销日与 5 个月的月平均温差和月相对湿度相关，具有明显的滞后性特征。作者还基于 CE 研究了 15 年滑动平均的相关性，发现气压和温度等与河冰日期的时间变化规律。这些结论有助于将这些鉴别的因子加入到河流模型中，建立更准确的冰期预测模型，从而使河流管理更好地适应气候变化的影响。

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 [67–70] 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。作者将方法应用于黄河流域伊洛河水文观测网络和上海雨量观测网络的设计。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 [71, 72] 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雨量观测网的设计和优化，结果表明了方法可靠且有效。徐鹏程等 [68, 70, 73, 74] 提出利用藤 Copula 来构建站点关系网络，再基于估计的藤 Copula 来计算站点间的 CE 值，在此基础上提出了结合 CE 和克里金指标的站点优化目标，利用滑动窗口法选择优化站点。他们基于淮河流域 1992–2018 年的日降水量观测数据，利用该方法对该流域 43 个雨量观测站点网络进行了优化，结果表明该方法得到的网络能够较传统类似方法得到的网络更有效地获取降水相关信息。他们还将该方法应用到上海市雨量站网的优化设计中，得到的站网具有冗余信息小、信息总量大、站网内部估计误差最小的特点。杨惜岁 [75] 提出一个结合联合熵比、冗余度比和 NSE 效率系数的站网优化准则，并基于 CE 理论提出了新的 MI 计算方法，提高了计算的准确性。他将方法应用于美国查克托哈奇 (Choctawhatchee) 河流域的 14 个水文站点，进行站点优化研究，最终得到了只包含 5 个站点的网络，提高了站网的监测效率。

分析河流的干流和支流之间的相关性对水利工程设计、洪水预防和风险防控十分重要。三峡大坝作为长江上游河段的大型水利工程，其一个重要功能就是洪水控制，研究该河段的主要河流相关性对工程设计和安全运行具有重要参考价值。Chen 和 Guo [76] 提出利用 CE 来计算河流相关性的强度，他们将方法应用于包含了 5 条主要干支流的长江上游河段，基于干支流 1951–2007 年的洪水记录数据计算河流间的相关性。他们发现河流之间总的相关性并不高，这与该地区的气候特征相符；相关关系最强的是岷江和沱江，这是由于二者距离最近，且属于同一降水区域；金沙江和岷江、沱江之间具有一定的相关性，对三峡大坝的洪水控制构成了一定的威胁；金沙江、嘉陵江、岷江和沱江对长江盆地的洪水发生具有显著影响。

不同河流和区域的洪水事件叠加易于形成复合洪水事件，但不同洪水过程之间的空间关系很难利用现有相关性分析方法来准确地描述和评估。Wang 和 Shen [77] 提出了一个整合藤 copula 和相关性评估的方法框架，其中利用了 CE 理论从藤 Copula 来估计 MI、CMI 和 R 统计量等相关性强度。他们将方法用于评估长江上游已鉴别的 102 个复合洪水事件中两种极端径流序列变量（洪峰流量和洪水流量）之间的关系。结果表明，该框架的多维 R 藤 copula 模型能够更好地描绘复杂多样的水文相关关系，特别是藤结构表示了支流洪水汇入干流的顺序和水文站之间的空间位置关系；该框架估计的三种相关性强度比传统的相关性强度更好地反映了复杂时空水文系统的复合洪水事件中的非线性关系。

黄河水沙调控关系到黄河治理的策略制定，科学认知评估黄河的水沙通量变化特征是基础性的科学问题，对研判黄河泥沙情势具有重要意义。特别是近几十年来，受气候变化和人类活动的叠加影响，黄河水沙含量发生了显著变化，需要准确估计径流量和输沙量的分布变化情况。Copula 函数是分析这种分布的基本数学工具，但此类问题往往观测样本较少，难以准确估计 Copula 函数

数的参数。Qian 等 [78] 提出了一种基于 CE 和全相关 (Total Correlation) 关系的 Copula 参数估计方法, 用于解决在样本较少的情况下 Copula 参数估计问题。他们将方法应用于黄河西柳沟河流域 1960-2016 年度径流量和输沙量的数据的分析, 该流域在 1999 年前后水沙关系发生了显著变化, 但数据较少。分析结果发现, 对于 1999 年前后的两个时段, 新方法均得到比两种传统方法更准确的 Copula 参数估计, 对数据的拟合更好。

河流冲刷是指水流对河床泥沙不断侵蚀的过程, 其结果是导致河床降低或河道宽度变化。收缩冲刷是一种特殊类型的河流冲刷, 是指由于水流区域面积减小导致的河流冲刷, 通常发生在人工桥梁基台等对水流构成限制或者自然原因导致的河道变窄等情况下, 其会对桥梁等设施构成了安全风险。准确的估算收缩冲刷深度对桥梁等建筑的风险评估和结构设计至关重要。传统的经验公式方法往往预测准确度很低, 利用机器学习方法构建冲刷深度预测模型是一个新的技术途径。Wang 等 [79] 提出了一种主元分析 (PCA) 增强的支持向量回归 (SVR) 方法, 用于构建冲刷深度预测模型。他们基于收缩河道清水冲刷实验数据, 首先采用了 5 种降维方法 (包括 PCA、tSNE、NMF、LDA 和 KPCA 等) 对冲刷相关的流体参数、沉积特征和几何特征等原始变量进行变换得到新的输入变量, 再利用 CE 来选择与冲刷深度具有强相关的新变量, 结果发现 PCA 方法得到新主元向量与冲刷深度的相关度最强, 因此采用了 PCA 作为模型输入变量的生成方法。后续的 SVR 模型预测实验结果显示, 采用 PCA 方法的主元变量作为输入的预测模型能够给出比传统方法更高的预测准确度 ( $R^2=0.971, MAPE=7.54\%$ ), 验证了该方法的优越性。

流域分区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区是一种基本的研究路径。传统的流域分区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 [80] 提出采用基于 CE 的 R 统计量来衡量节点间的径流相似性, 再在此基础上利用社团检测算法对流域进行分区。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, 需要设计相应数据生成模型。Porto 等 [81, 82] 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 (Jaguaribe) - 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 [83] 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。

气候变化和人类活动等因素直接影响着水文系统循环，使得径流、降水和蒸发等水文因素发生了不同程度的时空变化。因此，从空间角度研究降水和径流等水文因素之间关系，进而分析这些关系时空变化背后的气候变化和人类活动原因是水文学领域的重要课题，受到了国内外学者的关注，对水资源规划管理等经济社会活动具有科学参考价值。蒋佩东 [84] 利用 CE 等工具分析了长江流域降水、蒸发、潜在蒸散发、径流和植被指数 NDVI 的流域栅格数据，从得到的空间相关性发现了这些因素的空间分布特征，并给出了定性的地理学解释。特别是，根据 CE 估计值判断，他发现实际蒸散发和降水对年径流的影响较高，而年径流与以上各因素的空间相关性具有空间异质性特征。

## 6.12 气候学

气候变化是气候学研究的课题之一，它不仅体现在水文气候变量幅度上的变化，也体现在变量的季节和周期变化的分布上。这种变化会对降水和气温的强度和频率造成影响，导致极端天气（如洪水、干旱和热浪等）的增加。降水和气温的相关性会加剧联合极端天气的发生和强度。研究气候变化对降水和气温相关结构的影响是一个重要的问题。Hao 和 Singh [85] 利用 CE 度量工具研究了气候变化对这种相关结构的影响。研究采用了美国德克萨斯州达拉斯市沃斯堡（Fort Worth）在 1948-2010 年的每日降水和气温数据，以每 5 年为期计算温度和降水之间的负 CE 值作为相关结构强度，发现该地区的温度和降水之间的相关结构强度（负 CE 值）从 1948-1980 年的 0.18 下降到了 1948-2005 年的 0.06，说明了气候变化对该地区水文气候变量之间关系造成了影响。

气候评估是科学应对气候变化的基础性工作，其目标是监测和分析全球和地区气候及其变化，特别关注于变化趋势和极端气候风险等。气候分类是指根据相似气候特征将地区分类，最常见的 Köppen 分类法采用的气候特征是温度模式和季节性降水。Condino [86] 提出了一种基于 Jensen-Shannon 距离的动态分类算法，其中基于 JS 距离的分类准则采用了基于 CE 理论的表示方法并进行估计。他将方法应用于欧洲气候评估问题，根据 1951-2008 年欧洲气象观测站每日温度和降水数据对欧洲 25 座主要城市的气候进行分类。结果表明，其提出的算法成功区分了分别属于欧洲南部和北部气候带的城市群，当进一步考虑南北气候过渡带时，算法也对欧洲中部城市给出了与实际气候情况相符的合理的分类结果。

## 6.13 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 [14] 利用其提出的基于 CE 的 TE 估计方法（见第2.4节），分析了北

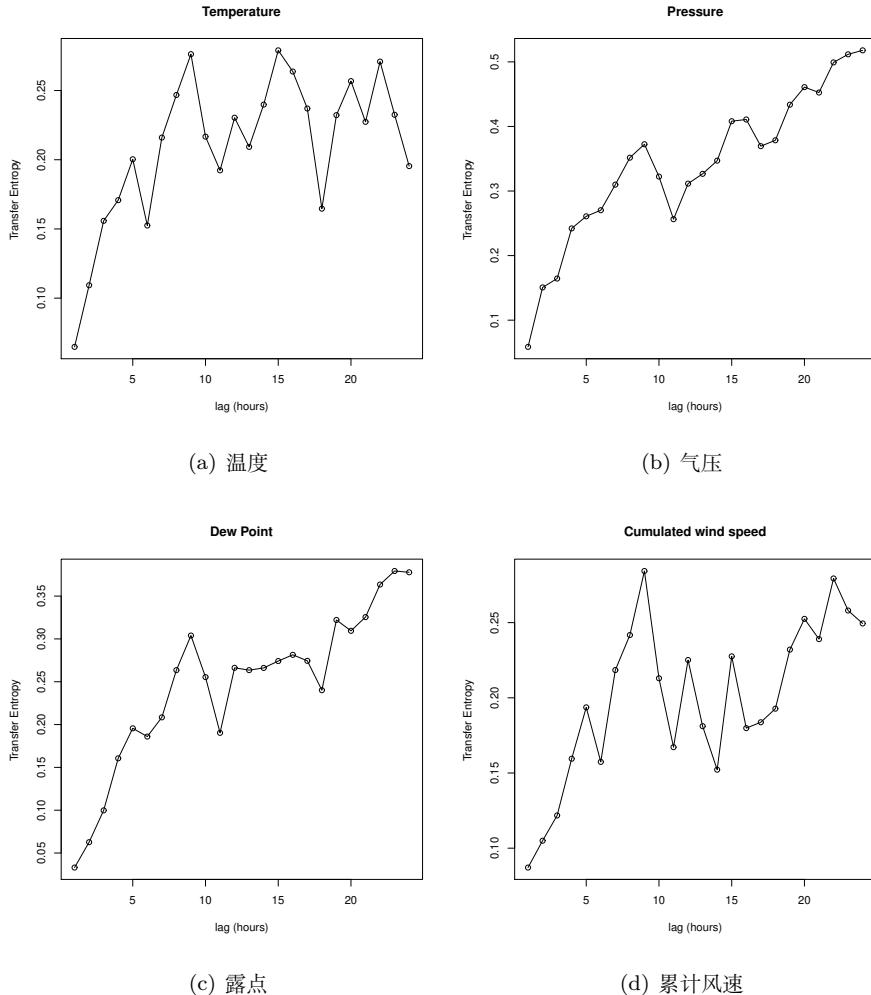


图 6.1: 四种气象因素到 PM2.5 浓度的 24 小时时滞内因果强度变化图.

京地区的气象和 PM2.5 连续观测数据 [374]，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6.1）。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。（更多内容见第2.4节）

有效的大气污染预测对于污染防控具有基础性作用，也利于保护居民健康。但当前的大气污染（如 PM2.5 浓度）预测在准确性和稳定性上还很难满足要求。开发性能更高的预测模型受到了广泛的关注。在综合考虑了传统方法的不足的基础上，Wang 等 [87] 提出了一种新的大气污染预测预警方法，使用了 CE 和多种机器学习模型的组合方法，CE 方法在其中被用来选择对

PM2.5 浓度波动有影响的因子，以用于构建最终模型。他们将开发的方法应用于上海和广州两地的实际大气污染预测预警系统，结果表明新方法能得到较其他对比方法更好的预测准确性和稳定性。Wu 等 [88] 提出了一种基于 CE 的 PM2.5 预测方法，利用 CE 计算气象因素与大气污染物浓度之间的相关性来选取模型输入特征，在基于 LSTM 和进化算法相结合的方法建立预测模型。该方法在北京地区 2016 年的历史数据上取得了良好的预测性能。Chen [89] 利用 CE 从多种因子中选出影响 PM2.5 的因子，再利用自注意力机制增强的时序卷积网络（TCNA）构建预测 PM2.5 浓度的模型，他将方法应用于北京市 12 个区域 2013-2017 年逐小时气象和污染观测数据，得到的预测模型具有高度的可解释性和预测准确度。Guo 等 [90] 提出了一种将经验分解模式组合与神经网络相结合的 PM2.5 浓度预测方法，其中利用 CE 等方法对不同方法得到的经验模式分解因子进行选择，他在北京市某观测站 2020-2022 年的 PM2.5 每小时观测数据上验证了该方法，证明了该方法的有效性和优越性。

全球气候变暖导致我国华南地区的台风强度越来越强，强台风给该地区造成了严重的损失。根据台风灾害的观测数据预测灾情程度，是台风灾害的研判和应对的重要参考。但台风灾害影响因子较多，且与灾情之间具有非线性关系，给预测模型构建造成困难。陈燕璇等 [91] 基于 CE 等工具，提出了一种台风灾情预测模型构造方法。他们基于 1985-2014 年间登陆或影响广西的 44 个台风灾害数据，以及同期与致灾、承灾和防灾减灾相关的灾情统计数据，构建了 21 个灾害影响因子，再利用 CE 筛选与灾情指数最相关的因子，发现最大风速、最低气压、暴雨时长和暴雨极值与灾情指数最相关，能够客观地反映实际情况。实验也表明，利用 CE 筛选的因子构建的模型的预测精度要高于同类对比方法构建的模型，可为广西台风灾情预测提供参考。

## 6.14 环境学

大气污染是现代城市面临的主要环境问题之一，严重影响城市运行和居民生活。对大气污染扩散规律的分析是环境科学的重要问题，对环境监管部门更好地了解污染规律并有效应对具有基础性的指导作用。大量的城市污染气象观测网点生成的数据，有助于分析扩散规律进而对污染扩散作出预测。吴京鹏 [92] 提出了一种节点无特征网络链路预测算法，并将其应用于城市大气污染传播路径建模和预测问题。他将方法应用于兰州市辖区环境监测站网 2017 年的 PM2.5 观测数据，利用基于 CE 的 TE 算法构建了传播网络，再应用提出的网络链路预测算法预测污染传播路径。实验结果表明，该方法可以准确发现污染传播路径，为兰州城市污染治理策略制定提供了理论支撑。

氮氧化物 (NO<sub>x</sub>) 是火力发电厂排放的主要污染物之一，需要通过实施监测来严格管控其排放浓度。电厂一般采用 SCR 脱硝反应器的中和方法控制 NO<sub>x</sub> 排放浓度，但控制过程存在大延迟，无法做到精确控制，一般采用软测量模型预测与 SCR 控制器配合来达成控制目标。金秀章等 [93] 提出了一种 VMD-Bayes-Lasso 相结合的软测量算法框架，以预测 NO<sub>x</sub> 排放浓度。该方法框架首先利用 CE 筛选与 NO<sub>x</sub> 浓度相关的系统变量，以预测分解后的 NO<sub>x</sub> 浓度模态变量，再叠加得到最终预测结果，最后设计了基于 Lasso 算法的模型误差预测模型来校正预测结果。他们在宁夏某 660MW 燃煤电厂的数据上验证了该算法框架，得到了较对比方法更好的预测精度，其中，通过 CE 方法分析了系统变量内部以及和目标变量之间的相关度，达到了精简模型和提高

预测精度的目的。杨媛 [94] 提出了一种电站锅炉 NO<sub>x</sub> 排放浓度的预测方法，首先对不同工况的系统延迟进行辨识，然后利用基于 CE 的自适应特征选择方法选择合适的系统变量，最后利用集成学习算法构建预测模型。作者在某电厂 1000MW 超临界锅炉系统上验证了该方法，结果表明，所提出的方法能够较对比方法更准确地预测 NO<sub>x</sub> 排放浓度。其中，基于 CE 的特征选择方法能够筛选出对 NO<sub>x</sub> 排放浓度影响显著的系统变量集合；同时，与同类方法相比，该特征选择方法能够给出更好的预测结果。

SO<sub>2</sub> 是火电厂燃煤机组排放的另一类主要污染物气体，对人类的生产生活环境构成严重损害，需要对其排放进行监管和控制。石灰石-石膏脱硫系统是火电厂广泛采用的湿法脱硫工艺，但由于数据采集问题和系统延迟等原因，系统出口 SO<sub>2</sub> 排放浓度难以准确测量，从而无法通过调整系统运行参数对排放浓度加以实时控制。乔鹏 [95] 提出了一种 SO<sub>2</sub> 提前预测的软测量技术，用于精准控制系统排放浓度。该技术首先利用特征选择、CE 和预测模型确定辅助变量及其时延和阶次，其中 CE 用于确定关键的变量时延；再利用 CatBoost-Bayes 误差补偿模型建立回归模型。他基于山西某 550MW 机组的历史运行数据验证了该方法，结果表明该方法建立的模型具有良好的软测量精度；同时，对比实验也表明，利用 CE 等方法对辅助变量的时延和阶次的优化显著提高了模型性能，验证了方法的合理性和有效性。

氨气 (NH<sub>3</sub>) 是大气中的重要碱性气体，在大气氮循环中发挥着重要作用，也因此与诸多环境问题密切相关。含氨气溶胶颗粒物是空气中 PM<sub>2.5</sub> 的重要来源；自然界中氨的含量变化也会导致土壤酸化、水域营养化和生物多样性降低等诸多问题。因此，研究氨气浓度的时空变化及其影响因素，具有重要的科学价值和现实意义。大气氨含量主要来自人类的农业、工业和城市交通活动，长三角地区作为我国的经济发达地区和人口密集区，氨气相关的环境问题尤其不容忽视。Xue 等 [96] 利用欧洲气象卫星红外大气干涉仪的氨气柱密度数据、NASA 的 NO<sub>2</sub> 柱密度数据和欧洲的 ERA5 气象再分析数据，研究了长三角地区在 2014-2020 年氨气柱密度的长期时空变化以及变化背后的驱动因素。其中，他们利用 CE 方法分析了导致氨气浓度空间变化的影响因素，通过计算氨气柱密度与气象因子、pH 值、人口密度和耕地占比等空间变量之间的相关性强度，发现氨浓度与地表气压、降水量、pH 值和耕地占比等因素密切相关，表明了长三角地区的氨分布受到自然和人类活动因素的双重影响。

## 6.15 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 [97] 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。

土壤是陆地生态系统的载体，气候环境的变化对其内部结构和功能演变具有直接影响，研究气候变化与土壤生态系统之间关系的规律是土壤生态学的主要问题之一。土壤碳氮是其生态系统的关键因素，土壤碳氮比是评估其系统功能的关键指标。土壤的碳氮转化主要由微生物过程来调解，而微生物对湿度的变化十分敏感。已有的研究主要关注土壤碳氮动态性对短期湿度变化的反应，对于长期的湿度变化，特别是气候变化导致的湿度变化下的反应缺乏了解。Li 等 [98] 利用微生物-酶分解 (Microbial-ENzyme Decomposition: MEND) 模型研究了土壤湿度变化对土壤碳氮动态性的长期影响。他们以广东鼎湖山国家自然保护区内的阔叶林生态系统为研究样地，采集了 2009-2012 年的气候环境观测数据和土壤实验数据。实验首先利用 CMIP6 中的 10 个全球气候模型 (GCMs) 生成了四种未来社会经济路径情景下的土壤水含量数据，再利用插值方法和偏差校正得到该研究样地的未来 (2021-2100 年) 土壤湿度变化数据，进而得到不同时间尺度 (1-72 个月) 的标准土壤湿度指数 (SSI)，然后利用实地采集的土壤分析数据对 MEND 模型进行参数校正，再利用 GCMs 生成的土壤湿度数据和校正的 MEND 模型仿真土壤碳氮动态性变化，最后它们利用基于 CE 的 MIR 统计量估计了不同时间尺度下土壤碳氮变量与 SSI 之间的相关性。研究发现，土壤碳氮动态性对湿度变化的反应呈现出滞后性和累积性，特别是土壤有机碳和全氮变量的相关性具有长期性 (72 个月)。其中，土壤有机碳倾向于在干旱条件下累积，而全氮也呈现类似的特点。实验将 MIR 与线性相关系数进行了对比，发现实验变量之间的关系具有非线性特征，因而基于 CE 的 MIR 更符合实验分析的需求。

自工业革命以来，人类活动对自然环境的影响显著增强，因此造成的生态环境变化反过来影响人类社会发展的可持续性。社会生态系统是人与环境构成的非线性、动态交互的层级系统，而系统韧性是指其能够对抗内外部的自然社会冲击，仍然保持系统核心功能的能力。生态脆弱区域的系统韧性对区域的经济社会可持续发展意义重大，理解生态脆弱区域的社会生态韧性演化对于提高区域生态保持能力至关重要。喀斯特地貌土壤形成慢，土层薄，破坏后难以恢复，是典型的生态脆弱系统，在我国西南地区分布广泛。Li 等 [99] 提出了一种利用基于 CE 的 TE 和网络分析等工具研究社会生态系统脆弱性演化的方法。他们以我国西南贵州和广西地区的喀斯特生态脆弱区域为研究对象，基于区域经济社会、气象观测和陆地遥感等数据，利用该方法研究了从 1990-2022 年期间西南喀斯特地区的生态系统特性演化。研究首先利用数据计算“驱动力—压力—状态—影响—响应” (DPSIR) 生态环境系统分析框架的 5 个维度指标，再基于指标变化将目标时间段分为 3 个特性不同的时期，然后利用 TE 分析每个时期内 DPSIR 五个维度之间的动态关系变化，最后利用网络分析影响系统功能的关键因素。基于 TE 的 DPSIR 因素关系分析结果表明，三个时期的驱动力到压力的 TE 呈长期下降趋势，表明该区域生态恶化对经济增长的影响降低，社会生态系统向发展与生态平衡的高质量发展模式转变；同时，三个时期的响应到驱动力的 TE 在逐渐增加，说明了人类对驱动力因素的正反馈增强，社会发展与环境保护的协同得到增强。作者认为分析结果展示了西南喀斯特地区社会生态环境脆弱性的演化过程，论证了这一演化过程与政府一贯而连续的生态保护政策密切相关。

## 6.16 动物形态学

动物形态学是动物学最古老的分支，研究动物体的形态和解剖结构以及其在发育和进化过程中的变化规律。作为动物学的基础学科，形态学的研究是动物分类的基础，比如鱼类的形态分类。由于鱼类的外形相似，对其种类进行鉴别往往会出现偏差，这就需要研究鱼类结构形态之间的相似性度量问题。Escolano 等 [100] 提出了一种图形形似度度量的估计方法，将图形转换为多维流形嵌入向量，再利用 CE 估计方法估计向量之间的 MI 作为图形相似度度量。他们将方法应用到 GatorBait 海洋鱼类图形数据库，该数据库包含了 30 个类别的 100 个鱼类外形三角网格图形。由于每类对应的是鱼类属而不是种，因此同一类别间具有形态差异，给分类造成困难。他们利用新度量方法对数据库中的鱼类图形进行分类，实验表明新的度量方法在数据集上得到了较传统方法更好的分类性能。

鲍是一类重要的海洋贝类，具有较高的营养价值和巨大的经济价值。鲍鱼的形态学研究是通过形态学变量的测量来研究其生长过程和种群分布等问题，对该类海洋资源的管理具有重要意义。Purkayastha 和 Song [101] 提出了一种新的因果关系度量概念，称为非对称 MI (AMI)，用于判断变量之间因果预测性的方向，并基于 CE 理论给出其快速且鲁棒的估计方法。他们将 AMI 方法应用于 UCI 鲍鱼数据集，分析了鲍鱼的长度、直径、身高和体重等形态学参数的测量数据，明晰了鲍鱼生长过程中年龄与这些变量之间的因果关系规律。

## 6.17 植物学

树干液流是指植物为了补充叶片蒸腾作用失水而从根本向叶片输送水分的过程，其可以作为评估树木蒸腾耗水量的关键指标。因其与植物内在生长机理和外在环境因子具有密切关系，因此可以利用这些因素对树干液流进行预测。传统的方法不能处理环境因子和液流之间的非线性关系，因此预测的效果不甚理想。王子祥 [102] 和 Li 等 [103] 提出了一种基于历史环境因子预测树干液流的方法，该方法首先利用 EMD 方法对树干液流数据进行分解，然后利用 CE 选择出最能体现液流数据特征的模态分量对液流进行重构，再利用 CE 选择出与液流最相关的环境因子，最后在所选因子和重构液流的基础上，利用传统机器学习方法和深度学习方法构建预测模型。他利用 SAPFLUXNET 项目公开数据集中 2012 年全年的贝壳杉树干液流数据以及 9 个环境因子数据验证了该方法，实验结果表明，CE 筛选了 5 个环境因子（包括饱和水汽压亏缺、风速、深层土壤含水量、大气相对湿度和净辐射等），由此得到的深度学习模型的预测效果要好于对比方法 ( $MSE=0.0229, MAPE=0.6759\%, R^2=0.9755$ )，说明引入 CE 方法后，得到的模型符合环境因子和树干液流之间的关联关系，从而使预测性能得到提高。

## 6.18 农学

全球变暖导致的环境变化会直接影响粮食产量，从而加重世界粮食安全问题。水稻是最重要的谷物作物之一，占我国谷物产量的四成左右，对我国的粮食安全至关重要。研究气候变化如何影响水稻产量并给出对策是关系到我国粮食安全的重要问题。Zhang 等 [104, 105] 利用作物模型

和大气环流模式研究了气候变化对我国南方（江南和华南）两季稻生长和产量的影响及对策。研究采用了 DSSAT 作物模型中的模拟水稻生长和产量的 CERES-rice 子模块和 CMIP6 中的四种大气环流模型（GCMs），并利用 CE 和随机森林分析各个月份的气象因子和作物产量之间的非线性关系。他们利用每个 GCM 的 27 组数据驱动南方 54 个地点的水稻作物模型以得到最终产量，同时研究了播种日期的影响。研究发现，气象因子的上升趋势会提前水稻成熟期并降低产量；如果再考虑  $CO_2$  作用，早稻产量则会增加，而晚稻仍会减产；根据 CE 计算结果，两季稻产量和  $CO_2$  浓度的关系也是气象因子中最强的；提前早稻播种和延后晚稻播种可能会增加一定的产量。该研究的结论为政府和农民应对未来的气候变化指明了路径，为采用相应的适应性对策提供了重要参考。

作为世界三大主要粮食作物之一，水稻在农业生产中具有重要地位。准确预测水稻产量有助于确保粮食安全和指导农业生产，是一个农业领域的重要问题。水稻作物的产量不仅与品种自身特性有关，也受到天气等环境因素的影响，这种影响具有非线性，给准确预测产量构成了挑战。张春磊等 [106] 提出了一种基于深度学习技术的水稻产量预测方法，其中利用 CE 来选择与产量具有非线性关系的环境因素变量，并利用 CNN 和 GRU 技术构建预测模型。他们在浙江省临安区真实数据的基础上验证了该方法，结果显示，CE 能够捕捉水稻产量与环境变量之间的非线性关系，CE 与 CGRU 相结合的方法给出了最好的预测结果。

## 6.19 神经病学

建立神经信号之间的因果关系对理解脑连接至关重要，因果关系连接反映了在脑认知过程中脑网络内部不同区域之间的信息传输方向，刻画了大脑认知过程的脑区之间动态关系特征。相比于传统的格兰杰因果检验，无模型假设的 TE 更适合此类因果分析任务。Redondo 等 [107] 基于 CE 理论提出了一种新的 TE 概念，称为 STE (Spectral Transfer Entropy)，用于计算频域滤波后的时域信号之间的 TE。与直接在原始信号上计算 TE 相比，在特定频域上计算的 STE 更具有神经学意义的可解释性。他们将方法应用于注意缺陷多动障碍 (ADHD) 患者 EEG 信号的分析，利用 STE 构建因果关系脑连接网络，发现了 ADHD 患者与健康人之间与注意力相关的脑连接网络连接的不同。实验结果表明，健康人在与注意力和受控记忆存取相关的  $\theta$  和  $\alpha$  频段表现出明显的因果联系，而 ADHD 患者的脑网络连接则主要在  $\delta$  振荡上，可解释为与注意力缺陷有关。

大脑神经系统在执行任务期间，各个脑区的信息处理活动是作为一个整体来协同进行的。相位同步 (Phase Synchronization) 是指大脑活动期间各个脑区神经信号在时序结构上的关系。传统的相位同步研究关注于成对脑区之间的关系描述，而全局相位同步 (Global Phase Synchronization) 则关注于多个脑区神经信号之间的同步强度。Li 等 [108] 提出利用 CE 计算多变量 MI 来估计 GPS 并检测脑网络结构。他们首先利用 Rössler 模型将该方法与其他同类方法对比，证明了其能够更好地估计网络中的 GPS 强度。然后，他们将该方法应用于颞叶癫痫症病人的 SEEG 信号数据的分析，发现癫痫的终止伴随着 GPS 强度和脑区整合度的增加，与已有的研究结论相符合。

脑卒中（俗称中风）是全球高致死病因之一，准确的诊断有助于有效的治疗和病情管理。核

磁共振成像 (MRI) 是脑卒中诊断的有力工具，但其复杂性给 MRI 数据分析构成了挑战。脑卒中对神经系统及其功能连接造成局部伤害，但是通过 MRI 功能连接变化的分析无法揭示病情对大脑动态功能的改变。有效连接 (Effective Connectivity: EC) 反映了大脑脑区之间动态的因果关系，为分析和诊断脑卒中对大脑动态功能伤害提供了有效路径。Ciezborka 等 [109] 提出了一种结合 EC 有向图分析和图神经网络分类的脑卒中分类诊断方法，利用储备池计算 (Reservoir Computing: RC)、格兰杰因果分析 (GC) 和 TE 三种方法生成 EC 有向图，再利用 GNN 等方法对 EC 图进行分类，以进行病情诊断。该方法利用了基于 CE 的 TE 估计算法生成 EC 图。他们在圣路易斯华盛顿大学收集的脑卒中病人和健康人的 MRI 数据上对上述方法进行了对比，实验结果表明，基于 RC、GC 和 TE 方法生成的 EC 图在小样本异质性的情况下得到的分类性能大致相当。该诊断方法可以进行 EC 图可视化和解释，有利于将研究发现转化为临床实践。

## 6.20 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 [111] 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 [110] 和听觉语音刺激任务的 MEG 数据 [112]。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。

在 GCMI 算法的基础上，Combrisson 等 [113] 提出了基于信息论的群体层面分析大脑认知网络的方法，将非参数的排列操作与信息度量相结合，用于分析固定效果或随机效果模型，以适应多人间和多次任务间的变化。他们将方法应用于两个已有研究的数据：第一个研究分析人执行认知行为映射任务时的 MEG 数据中的高 Gamma 行为 (High Gamma Activity)，发现了任务相关的大脑网络，涉及多个运动区、体感区和视觉皮层区域等；第二个研究分析奖惩学习任务的前脑岛 (anterior Insula) SEEG 数据，发现了奖惩任务的响应时延，以及奖和惩响应的显著差别。汪方毅等 [114] 提出了一种老年人认知水平分类方法，首先采用 GCMI 构建脑认知网络，再利用 GCMI 进行特征选择，最后利用 SVM 从选择的脑网络连接进行认知水平分类。他们将方法应用于 98 名葡萄牙老人的静息态 fMRI 数据，发现提出的方法能够捕捉数据中脑区间的非线性关系，并能够最终得到较同类方法更高的分类准确率。

语音理解是人脑的主要认知功能，研究人脑的神经活动对语音信息的编码和解析是认知神经学的重要问题。语音包络 (speech envelope) 包含了语音信号中的低频时序信息，研究表明其可

以解释大部分神经响应的变化过程，语音包络跟踪就是通过脑电图等手段研究语音包络及其神经响应之间关系的问题。由于大脑的非线性特征，常用的线性模型不能很好的表示这种关系。MI 作为非线性关系度量工具，被认为能够捕捉语音包络和神经响应之间的非线性关系。De Clercq 等 [115] 利用根据 CE 理论构建的 GCMI 工具，基于两组故事讲述语音和相应采集的 EEG 数据，对比了线性模型和 MI 分析对大脑非线性成分的刻画能力。实验结果表明 MI 分析检测到了线性模型以外的显著的非线性成分，证明了 GCMI 是比线性模型更适合于研究神经包络跟踪问题的工具。作者也实验证了与传统的 MI 估计方法相比，基于 CE 原理的 GCMI 方法具有鲁棒、无偏和适合多变量分析等诸多优点。

神经元特化 (neuron specification) 是指其具有执行特定功能的属性，可以通过研究外部环境刺激和神经响应信号之间的关系来鉴别。MI 作为一种非线性关联关系度量，是研究此问题理想的工具。Pospelov 等 [116] 利用 GCMI 方法计算钙荧光信号和环境变量、动物行为之间的相关性强度，对小鼠大脑海马的 CA1 区记录的钙信号进行了分析，揭示了与动物外部环境相关的特化神经元，如位置神经元，以及与其行为活动相关的特化神经元，如在跑、直立和静止时活动的神经元。研究也发现了一些对离散变量进行响应的神经元，如动物的场地位置（中央，靠墙和角落）和其速度（休息、慢和快）。他在四组实验中一共检测到 472 个神经元的 781 种特化。

神经学研究表明，大脑的意识水平与脑活动复杂度紧密相关，大脑活动复杂度的变化会导致意识清醒水平的变化。因此，通过度量大脑复杂度来衡量大脑意识水平是一个重要的研究方向。信息论为大脑复杂度的分析提供了多种度量工具，其中  $\Omega$  信息度量在评估大脑内部交互水平上具有独特性，可以不仅是度量整体的相关性，而是衡量协同作用或冗余作用的显著性，但其估计算法面临变量组合爆炸的问题。Belloli 等 [117, 305] 提出了一种基于 GCMI 方法来估计  $\Omega$  信息的方法，显著提高了其估计算法的计算效率。他们将该估计算法应用于人类清醒和深度麻醉的 fMRI 数据集，用以分析不同意识水平下的大脑复杂度。实验将数据中的 55 个脑区活动分为 11 个不同的脑区网络进行分析，其中每个网络包含 5 个脑区。实验结果表明，深度麻醉导致了脑区交互的  $\Omega$  信息的最大值和最小值都有所减小，使不同脑区网络间的交互协同作用减弱，同时也降低了同一脑区网络内的冗余作用。

瞳孔扩张是认知和行为过程的反映，由与脑干瞳孔神经支配系统连接的皮层结构所控制，瞳孔测量为理解认知活动的脑机制提供了一个有力工具。现代记录技术已经可以记录神经元活动的脉冲数据，使得分析认知行为和神经元活动之间的关系成为了领域重要问题，而如何从大量的神经元活动实验数据中分析认知-神经元相关性成为了一个瓶颈问题。Walden [118, 119] 基于 CE 理论提出了 Copula-GP 与高斯过程因子分析 (GPFA) 相结合的方法，用来研究活体小鼠的视觉皮层神经营过程与瞳孔扩张之间的相互关系。他将该方法应用于小鼠对漂移光栅刺激过程中记录的神经元脉冲信号数据，首先利用 GPFA 对数据进行降维，再利用 Copula-GP 估计藤 Copula，最后估计 CE 来衡量认知和神经元活动之间的相关性。实验结果表明，Copula-GPFA 方法能够有效地从数据中估计 CE 值，证实了神经元活动轨迹和瞳孔扩张之间的信息交互过程，与已有的研究发现相符合。

视觉诱发电位 (Visually Evoked Potentials: VEP) 是通过大脑视觉皮层的视觉刺激的电生理响应，研究大脑视觉认知功能的神经学分析技术。利用信息论方法分析大脑活动期间的信息流是一个重要的神经学数据分析方法，可以促进对大脑动态响应特征的理解。基于 GC 理念的定向

信息（Directed Information）和 TE 是两个主要的信息流度量工具。Kaufmann [120] 提出利用这两种工具分析 VEP 的 EEG 记录数据，他给出了定向信息和 TE 基于两种不同因果关系定义得到的入流（inflow）和出流（outflow）的两种分解形式，再基于 CE 理论给出了高斯 Copula 假设下的两种信息流分解形式的估计方法。他将该定向信息和 TE 的信息流估计方法应用于一组健康人视觉刺激实验 VEP 的 EEG 数据，成功发现了从枕中枢到额中枢电极之间的信息流。

## 6.21 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 [121, 122] 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 [123, 124] 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。Zhu 等 [125] 提出了基于 CE 的表示 TE，再利用 R 藤 copula 估计 CE 进而估计 TE。他们将该方法应用于上肢肌肉间耦合网络的研究，基于疲劳/非疲劳状态下上肢肌肉运动的 sEMG 数据构建了肌肉耦合网络，发现疲劳状态下的肌肉群间耦合关系较非疲劳状态逐渐加深。金国美等 [126] 提出利用小波分析和 CE 估计相结合的方法，分析健康人自主运动下的肌肉疲劳状态的 sEMG 信号数据，发现在肘关节屈曲运动中，肌间耦合强度在 Beta 与 gamma 频段最为显著，协同肌肉对耦合强度比拮抗肌肉对耦合强度大；疲劳后的耦合强度相对于疲劳前有所增强。

## 6.22 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 [127, 128] 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 [129] 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。Coroian 等 [130] 基于 CE 理论

提出了一种估计连续分布的协同的方法，以提高现有协同估计算法工具的计算效率，并应用于大脑神经活动数据，以分析脑功能连接图谱和脑区活动模式变化过程中的协同涌现现象。

### 6.23 心理学

大脑是一个分布式的网络系统。它不仅控制身体，改变内部生理状态，也影响多个高级过程。同时，内脏信息也时刻受到大脑的监控，也就意味着内脏过程也会反映到皮层活动中。内脏事件相关的大脑活动研究是一个重要的话题。植物神经系统中的过程之间相互关联，而信息论则提供了研究它们之间关系的工具。Ravijts [131] 研究了四种情绪刺激特征（效价、唤醒、支配和喜欢）下心跳诱发脑电位（HEP）的时间交互近似估计问题。他采用了用于情绪分析的生理信号 DEAP 数据集，利用基于 CE 的 GCMI 方法估计了 MI、协同和冗余等统计量，用于度量不同情绪刺激下 HEP 上的时间交互。实验发现了支配和喜欢情绪刺激下 HEP 上的时间交互现象，第一次揭示了情绪感知调制的 HEP 的时序特性。

### 6.24 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 [132] 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 [133] 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。

### 6.25 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 [134] 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 [135,136] 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。Shang 等 [137–139] 提出了一个基于 CE 的特征选择方法，称为 CEFS+，用于高维基因数据分析工作。他们将该方法应用于 3 个公开的高维癌症基因数据集，将所提出的方法与同类方法进行了对比，结果表明，该方法在实验中给出了最好的预测结果，证明了该方法的优越性。Pan 等 [140] 提出了一种基于 CE 改进的灰狼优化算法，用于对高维基因组数据进行特征选择。该方法利用基因对应的 CE 值来初始化灰狼优化算法的种群，然后再进行特征选择优化。他们在 10 组高维小样本基因表达数据集上验证了该改进算法，结果表明，相对于其他特征选择方法，该改进算法在选择更少特征的情况下改善了分类性能，基于 CE 的改进算法在 10 组数据集中的 7 组上取得了最好的分类结果，证明了算法的有效性和优越性。

在生命过程中，复杂的疾病往往由多种基因共同导致，反映到基因组数据中表现为疾病类型数据与多个基因特征同时相关。因此，在建立生物信息学模型时，特征选择过程中必须考虑这种多个基因之间的交互作用，才能得到性能更好的疾病分类模型。钟琦 [141] 提出了一种基于 Copula 和标准化度量的特征选择方法，称为 IFSMRMR，同时考虑特征与疾病标签之间的相关性和特征之间的冗余度，其中 CE 被用于度量特征选择过程中的交互作用强度。他在 8 个公共的基因表达数据集上将该方法与 6 种常见特征选择方法进行了对比，结果表明，该方法通过考虑特征交互作用提升了分类性能，同时也得到了更好的特征聚类效果，性能优于同类对比方法，验证了方法的有效性和优越性。

基于基因测序数据推理构建基因调控网络是生物信息学的主要问题之一，目的是理解基因功能和识别基因表达的动态过程。单细胞测序技术能够同时测量大量单个细胞的全基因组表达情况，时序单细胞测序数据则反映了细胞中基因调控动态过程。因此，可以利用 TE 等非线性时序因果分析工具发现基因调控网络。竺政彤 [142] 提出了一种基于 TE 因果关系分析的基因调控网

络构建方法，称为 GRN-PAGATE，其中采用了基于 CE 的 TE 估计方法。他分别在 DREAM3 挑战中的 Ecoli 数据和小鼠胚胎早期血液发育的单细胞测序数据上验证了该方法，并与同类方法进行了对比。实验结果表明，该方法在 Ecoli 数据上具有与 GRNTSTE 同等的性能，略高于 DynGENI3 和 SCRIBE 等同类方法；在小鼠胚胎数据上，该方法能够有效发现其他方法未能发现的关键基因调控关系，性能优于同类对比方法。

贝叶斯网络 (Bayesian Network: BN) 作为一种因果关系分析工具，在系统生物学的组学数据分析中越来越被广泛使用，但其具有线性假设，使其难以胜任复杂调控网络动态的分析。CE 作为一种模型无关的非线性关系度量，提供了分析处理复杂高维基因数据的高效算法工具。Li [143] 提出了一种将 BN 和 CE 相结合的新方法，用于构建基因调控网络，其中 CE 被用于对 BN 生成的子网络结构做进一步过滤识别，以得到最终的基因调控网络。基于 6 个公开的肝细胞癌测序数据集，作者将该方法应用于肝细胞癌 (Hepatocellular Carcinoma: HCC) 的基因调控子网络识别问题，并将该方法与 5 种同类方法进行了对比。基于该方法的实验给出了一个包含 12 个基因和 28 条边的基因子网络，是所有方法中最小的子网络。作者将每种方法得到的子网络应用于三个 GEO 数据集，进行 HCC 肿瘤分类实验，以检验每种方法得到的子网络的有效性。实验结果显示，该方法得到子网络在 3 组分类实验中均表现最佳，平均 AUC 都显著高于 90%。随后的差异基因表达分析进一步证实所得到的子网络基因在 mRNA 和蛋白质水平上被显著下调，是潜在的疾病诊断生物标志物。该方法得到的结果具有可解释性，能够增进对疾病潜在分子机制的理解，有助于识别潜在的治疗目标并开发出相应的治疗方法，有望在其他疾病的同类分析中得到应用。

人类基因组测序数据为理解疾病的生物学机理打开了一扇大门。然而基因组数据的维度远超过了测序人数，使得基因组数据分析面临维度灾难问题。理解数据并将其转化为生物医学技术需要更多新的高维统计分析工具，信息论为解决此问题提供了理论工具。传统信息论的 MI 和部分信息分解理论的独特信息、冗余和协同等概念可以帮助我们分析高维数据，但如何从数据估计它们是一个实际面临的难题。Lacalamita [144] 提出利用基于 CE 的 GCMI 方法来估计 MI 和协同概念，并在此基础上提出了面向疾病预测的基因组数据分析流程方法。该方法首先利用复杂网络社区检测技术对基因进行聚类分析得到基因类簇，然后对每个类簇利用特征选择技术构建疾病预测模型，从而得到预测性能较好的基因类簇，再利用基于 CE 的 MI 和协同度量对得到的基因类簇做进一步聚类分析，得到基因子类簇，最后利用子类簇的基因进行疾病预测。他利用 NCBI 关于肝细胞癌 (Hepatocellular Carcinoma) 和自闭症谱系障碍 (Autism Spectrum Disorder) 的公开基因组数据验证了该方法，发现该方法可以对高维基因进行有效的筛选，同时利用协同概念聚类选择的基因要好于利用 MI 聚类选择的基因，表现为前者所选的基因更符合生物医学意义上的疾病机理研究结论，且得到的疾病预测性能更好。

## 6.26 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 [328]，马健 [13] 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。（更多内容见第 2.3 节）

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 [594]，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 1999-2008 年的 10 年间 101,721 名住院患者的数据，Mesiar 和 Sheikhi [145] 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。

癌症预后是指基于癌症病情的临床表现和诊断结果，对病情的未来发展进行评估，以帮助进一步的临床决策。临床评估考虑的预后因子在评估中至关重要，但又常常数量众多，需要进行分析选择。比如，肺癌的预后因子就多达百种。预后模型是在预后因子的基础上建立的病人风险预测模型，是癌症治疗中重要的临床工具。马健 [146] 提出了一种基于 CE 的生存分析变量选择方法，并将其应用于预后因子的选择问题，以建立预测病人生存时间的预后模型。他基于两个公开的肺癌数据验证了该方法，发现其能选择符合临床标准的预后因子，并获得较同类方法更好的预测模型，在保证模型可解释性的同时具有更好的预测性能。

乳腺癌是女性最常见的恶性肿瘤之一，在我国的发病率和死亡率都有逐年上升的趋势，严重威胁着女性的身体健康和家庭幸福。利用统计方法分析临床数据并构建诊断模型来辅助临床诊断决策，可以提高医生工作效率并降低误诊率，从而促进患者健康改善。付金露 [147] 提出采用特征选择的方法构建乳腺癌患者预后模型，采用了 Lasso、CE 和 RFREF 三种特征选择方法，分析了 SEER 数据库中 2010-2014 年的乳腺癌患者临床诊断数据，利用三种方法选择的特征分别构建了逻辑回归、随机森林、XGBoost 和 Stacking 四种模型，用以预测患者 5 年生存状态。结果表明，利用 CE 选择的特征构建的逻辑回归模型给出了最高的预测准确率（96.84%）。

白内障是眼科的常见疾病，是导致患者失明的主要病因。白内障超声乳化手术（Phacoemulsification）是世界各国治疗白内障的首选手术治疗方式。尽管该手术已十分成熟，但临床仍然可能会导致术后角膜水肿等并发症，从而影响视力恢复并造成患者不适。构建基于风险因素的角膜水肿风险预测模型在临床十分必要。Luo 等 [148, 149] 提出利用 CE 方法构建术后角膜水肿风险预测模型，将方法应用于临床 178 名患者的数据，从数据的 17 个变量中筛选预测变量，最终将临床预测模型使用的四种变量（糖尿病、最佳矫正视力、晶状体厚度和累积耗散能量）减少为两种（最佳矫正视力和累积耗散能量），且不影响预测精度。结果分析表明，利用 CE 得到的预测模型具有临床应用价值，可以在保证预测性能的情况下减少预测需要收集的临床信息。

主动脉瓣反流（Aortic Regurgitation）是一种常见的心脏瓣膜疾病，主要症状是在心脏舒张期，血液从主动脉回流到左心室。主动脉瓣膜置换手术是主动脉瓣反流的传统治疗方法之一。左心室射血分数（Left Ventricle Ejection Fraction：LVEF）是一项衡量心脏功能的重要指标，研究其在手术前后的改善关系可以为瓣膜置换手术时机选择和效果预测提供参考证据。Sunoj 和

Nair [25] 利用 survival copula 扩展了 CE 概念，提出了一种称为 Survival Copula Entropy (SCE) 的新概念，用于衡量生存函数相关变量之间的依赖关系。他们将 SCE 应用于主动脉瓣置换手术临床数据，发现了手术前后 LVEF 之间的正相关关系。

脑肿瘤是一种高致死率肿瘤，约占全身肿瘤的 5%，近年来在我国发病率呈上升趋势。脑肿瘤病变具有形态多样、位置不定的特点，诊断难度大，基于无侵式医学影像的分类识别是主要的临床诊断方式。利用深度学习方法，从肿瘤医学影像提取定量特征并构建诊断模型，可以辅助医师的临床诊断，因此得到了大量的研究。如何提取和选择图像的定量特征是构建辅助诊断模型的关键问题。潘红宇 [150] 提出了一种此类特征选择方法，首先利用 CE 等相关性度量初始化特征集合，再利用灰狼优化算法以分类性能为目标优化特征集合。他利用来自重庆医科大学附属第一医院、西南医院和四川省肿瘤医院的 102 例具有 ATRX 突变的低级别脑胶质瘤患者影像数据，提取了五类共 5530 个影像组学特征，结果表明，相较对比方法，提出的方法在选择使用了最少（13 个）的特征的情况下得到了最优的分类性能，且所选特征与 ATRX 突变特征状态相关，具有作为生物标志物的潜力。

脉搏波是传统中医的主要问诊方式，因其携带了复杂多样的病理信息，在一定程度上反映了心血管系统的生理状态。传统中医的诊脉主要依靠名医的个人经验，研究脉搏波数据的分析算法，对糖尿病和高血压等常见疾病的无创诊断具有重要意义，有助于传统中医的科学化发展。汤宇飞 [151] 提出了一种基于图卷积神经网络的多模态脉搏波诊断算法，通过将脉搏波转换为包含互补的病理信息的三通道图像，再利用 ResNet 提取图像特征，最后利用 CE 等相关度量得到反映脉搏波信号间时间相关性的邻接矩阵构建图卷积神经网络，从而进行疾病分类诊断。他在实际腕部和指尖的脉搏波数据的基础上，对高血压和糖尿病患者的健康状态进行分类，结果表明算法可以得到 99% 以上的预测准确率。

## 6.27 老年医学

阿尔兹海默病（Alzheimer's disease，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 [152] 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。

帕金森病（Parkinson's disease: PD）是另一种常见的神经退行性疾病，临床表现为动作迟缓和运动功能障碍等症状。重复经颅磁刺激（repetitive transcranial magnetic stimulation: rTMS）是利用脉冲磁场作用于中枢神经系统，以改善生理功能的临床治疗技术，广泛应用于神经、精神类疾病的治疗，并在近年应用于 PD 康复治疗的研究中，以期缓解患者症状并改善运动功能。李润泽等 [153] 研究了 rTMS 对 PD 患者运动症状辅助治疗的神经调控机制，利用基于 CE 的 GCMI 等方法分析了 rTMS 治疗前后的 EEG 数据，构建了脑功能网络连接矩阵并得到 3 种网

络特征参数。实验结果表明 rTMS 主要改变 PD 患者的 beta 和 gamma 振荡，其中运动皮层的相应变化可能与运动功能改善有关。

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 [154] 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 [155] 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。

## 6.28 精神病学

抑郁症是一种常见的情绪相关的心理精神障碍，全世界约有 3.5 亿名患者为此病所困扰，对其进行研究对人类健康具有重要意义。脑电图 (EEG) 是一种非侵入式的大脑活动电信号测量手段，广泛应用于大脑疾病的研究中。脑功能网络是在 EEG 信号基础上构建的反映大脑活动的功能性指标，可采用 MI、相干性等多种方法构建此类网络。张婷婷等 [156,157] 提出基于相干性虚部 (Imaginary part of Coherency) 构建的脑网络连通性指标来研究抑郁症患者识别问题。他们利用 CE、Relief 过滤等特征选择方法对脑电网络连通特征进行选取，发现利用 CE 和 Relief 过滤联合得到的相干性在线反馈指标特征集合能够有效区分抑郁症患者和健康人群。

精神疾病是由心理、生理和社会环境等因素导致的大脑功能紊乱，如精神分裂症、双相型障碍、注意缺陷多动障碍等。大量研究表明，这些疾病可能具有一个共同的精神病理学模式，因此跨疾病的病理学分析就变得十分重要。功能核磁共振成像 (fMRI) 作为非侵入的脑活动检测方法可以提供大脑脑区多频段的大脑活动信号，进而推断脑功能连接网络，用于疾病机理和诊断方法研究。然而，传统的脑功能网络大多基于脑区间的线性关系，而大脑神经活动则呈现非线性动态性特征，需要设计能够具有非线性关系检测能力的脑功能网络生成方法。Han 等 [158] 提出了一个基于变分模式分解 (VMD) 和 CE 的多频段分解熵 (Multi-frequency Decomposition Entropy: MDE) 方法，用以从 fMRI 时序数据推断非线性脑功能连接网络。其中，VMD 用于对脑区的 fMRI 时序信号进行频率模态分解，而 CE 则用于计算不同模态上脑区之间的非线性关联强度矩阵，最后基于此矩阵选出具有强关联的成对脑区关系。他们将该方法应用于 openfMRI 精神疾病脑成像公开数据集，其中包含了精神分裂症、双相型障碍、注意缺陷多动障碍和健康人的样本数据。研究发现，该方法得到的三种疾病的病人和健康人的网络关系具有明显的不同，且每一种疾病的网络都具有自己独特的特征，因而具有作为疾病生物标记物的潜力。作者将 MDE 方法生成脑功能连接网络与线性方法生成的网络进行了对比，发现基于 CE 非线性度量的 MDE 方法能够

检测到病人和健康人之间网络上的差异，而传统线性方法则不能在网络上检测到差异，验证了该方法的优越性。

## 6.29 法医学

法医 DNA 分析是法医医学领域的热点前沿技术之一，是指基于 DNA 遗传信息和人体表型特征之间的对应关系分析，从生物样本的 DNA 检验信息对 DNA 来源人表型特征（如身高、年龄等）进行刻画预测，为案情分析和身份鉴别提供线索和证据。身高是法医学关心的重要生物特征，基于 DNA 分析推断人身高是法医学关心的重要问题之一。DNA 甲基化（DNA methylation）是重要的表观遗传学现象，很多研究表明了 DNA 甲基化与身高之间存在联系。因此，利用 DNA 甲基化位点预测身高特征是一个值得研究的问题。王中华 [159] 研究了基于 DNA 甲基化位点预测身高的问题。他首先利用公开数据证实了 DNA 甲基化与身高之间的联系，并筛选了相关位点集合，再基于招募人群的测序数据利用机器学习方法构建了基于所选位点的身高预测模型，得到了较为精确的预测性能（预测误差约为 4.5cm）。其中，他利用 CE 等方法分析估计了不同性别对应的所选位点与身高之间的相关性，发现甲基化特征与身高的相关性存在性别差异，这种差异符合人体生长发育过程的性别异质性，同时也在模型预测实验结果中得到了体现，从而表明了基于性别构建不同身高预测模型的必要性。与之相对照的是，基于线性 Pearson 相关系数的分析没有发现这种性别差异性，说明了 CE 作为非线性相关性度量的优越性。

## 6.30 药学

药物-靶点相互作用（DTIs）预测是指对药物小分子与蛋白质靶标之间的结合进行计算预测，是药物发现流程中的最重要环节之一，可以提高传统实验候选药物筛选的效率。利用机器学习方法从药物和蛋白特征集合预测 DTIs 已经得到了广泛的研究，取得了大量的成果。其中，如何选择合适的特征集合以建立 DTIs 预测模型是该领域的关键问题，选择结果对最终的模型预测性能至关重要。高浩田等 [160] 提出了一种交互式多特征融合的 DTIs 预测算法，其中利用 CE 等工具设计了一种多特征融合算法，称为 RCI，可以在不损失预测精度的前提下，选择具有高交互性和低冗余度的最优特征集合。他们利用四类基准数据集（Enzyme、IC、GPCR 和 NR）将该方法与同类 DTIs 基线算法进行了对比。实验结果表明，RCI 方法性能优于同类特征选择方法，在所选特征集合最精简的情况下模型预测精度结果最优；同时，基于 RCI 的预测算法的预测性能指标要好于同类 DTIs 基线算法，从而验证了算法的优越性。

## 6.31 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 [145] 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%。

高血压是全球首要致死病因，对人群健康构成严重威胁。全基因组关联研究表明多个基因与高血压密切相关。已有多个研究报道 I 型细胞膜钙离子转运酶基因 (ATP2B1) 与收缩压和舒张压相关联。该基因有 21 个 CpG 位点。研究该基因及其 CpG 位点与高血压的关系是一个新的重要问题。Purkayastha 和 Song [101] 提出了一种新的非对称可预测性概念，称为非对称 MI (AMI)，并利用 CE 理论给出了其估计方法。他们将该方法应用于 ELEMENT 数据集，分析 525 个年龄在 10-18 岁之间的儿童的数据，发现 ATP2B1 与舒张压相关联，证实了已有的发现；同时发现该基因的 CpG 位点 CG17564205 与舒张压相关联，且根据 AMI 判断，舒张压对该位点具有预测性，这一新发现表明血压可以改变位点。

## 6.32 经济学

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 [161, 162] 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。Zhang 等 [163] 将同样的方法应用于上述调查数据中河南省的数据，得出了基本相同的结论。

经济学的核心目的是发现因果关系。传统的经济学依靠推理建模以及基于此的实验设计。因果发现是从数据中发现因果关系的方法，将其与经济学理论模型相结合是设计经济学实验的新路径。Bossemeyer [164] 基于 CE 和 MI 的关系提出了一种条件独立性测试算法，并将其应用于因果结构发现的 PC 算法中。作者利用新 PC 算法研究了经济学中的议价理论，研究讨价还价行为中互惠关系的作用，以及响应时间在这个过程中的作用。作者将算法应用于 eBay 的 Best Offer 平台数据，发现交易双方让价行为之间存在关联，印证了互惠理论；同时，发现了对手还价响应时间对下一次要价存在因果效应。

产业链是指产业部门之间基于经济关系形成的链条式关联关系形态。产业链基于资源要素分配和专业化分工等多种因素构成上下游关系，来进行价值互换，上游企业向下游企业提供产品和服务，同时接受下游企业的反馈信息，从而构成关联互动关系。产业链各环节之间的相关性分析，对产业布局管理和投资组合设计具有重要参考意义。韦颖璐 [165] 基于 CE 概念，提出了 pair-copula 熵的概念，用于度量多变量内部的成对相关关系。她将该概念应用于国内畜禽养殖产业链各环节之间的相关性研究，基于该领域内 9 家上中下游主要上市企业的股票价格数据，运用 pair-copula 熵度量了产业链内上中下游之间的相关性，发现该产业链上游相关性较强，下游

相关性较弱；无条件相关性强，条件相关性弱；上中之间相关性强等现象。

投资者情绪对财经市场有着广泛而多面的影响，投资者情绪分析是经济学研究的重要问题之一。由于社交媒体和市场关系整合，投资者情绪会在人群和国家间传播，进而形成传播网络，使得局部情绪波动得以迅速扩散，造成系统性影响。Han 和 Zhou [166] 提出了一个基于小波分析、传递熵和网络分析组合的方法，研究公司间投资者情绪传播的模式，其中采用了基于 CE 的传递熵估计方法。他们采用 2015-2021 年的中国 137 家新能源汽车上市公司的百度搜索索引数据来代表投资者情绪，将其用小波分析分解为多尺度信息，再用传递熵构建情绪传播网络，最后用网络分析的方法分析短期和长期传播特征。他们发现，投资者情绪表现为短期局部活跃，并具有连续且逐渐增长的进化模式。

通胀预期直接影响市场主体的经济行为，是通胀的成因之一。研究通胀与预期的关系是一个重要的课题，特别对中央银行决策者具有重要价值。Ardakani [167] 提出利用 CE 分析预期对通胀的信息量，证明了负费舍尔信息 (Fisher Information) 是 CE 的下界，可以作为通胀和预期关系的最小度量值。他利用 CE 等工具分析了美国 1982-2022 年逐月通胀指数 (CPI 和 PPI) 和通胀预期指数 (密歇根大学调查指数、克利夫兰联邦储备银行 2 年、10 年和 30 年预期指数) 数据，发现 30 年预期与通胀之间 CE 最小，说明其提供了更多可以预测通胀的信息。此研究为中央银行管控预期以达到通胀目标提供了一个有力工具，能够帮助理解不同预期对通胀的预测能力，从而更有力地调控通胀。

### 6.33 管理学

准确预测农产品期货价格有助于为政府相关部门的科学决策提供参考，因而对保障国家粮食安全具有重要意义。然而价格预测受多种复杂因素的影响，如国际形势、市场情绪博弈等。因此，识别价格的影响因素对构建准确的价格预测模型至关重要。An 等 [168] 提出了一个基于历史数据和文本数据的融合多种方法的混合预测框架，其中经验模态分解 (Empirical Mode Decomposition: EMD) 用于预处理历史数据，动态主题模型 (Dynamic Topic Model: DTM) 和情感分析用于提取微博文本信息，再利用 CE 等方法对提取的因子进行筛选，用于构建预测模型。作者在两个实际数据上验证了该方法框架：国家统计局的猪肉价格数据和大连商品交易所的大豆期货价格数据，并收集了相应时间内的微博文本数据。在实验中，作者将 CE 方法与同类的 dCor 和 HSIC 方法进行了对比，结果表明，在两个数据上，基于 CE 的预测模型都给出了最好的预测性能。

巴西是全球第一大食糖生产和出口国，其甘蔗种植历史可以追溯到 1532 年，种植范围几乎遍布整个巴西领土，中南部地区和东北部是主要的产区，其中圣保罗州在种植面积和其衍生物产量方面都处于领先地位。同时，巴西又将过半甘蔗用于生产无水和含水乙醇作为车辆燃料，旨在减少该国对进口石油的依赖，其为全球第二大燃料乙醇生产和第三大燃料乙醇消费国。因此，巴西的汽油和乙醇组成的燃料市场与甘蔗生产的农业市场密切相关，并受多种自然、经济和社会因素影响，这三种大宗商品的价格之间关系复杂，分析它们之间的关系对巴西市场管理者和参与者具有重要参考价值。Flores [169] 利用基于 CE 的 TE 方法分析了 2004 年 5 月至 2023 年 11 月期间三种商品价格、收益和波动率的时序数据，分析新冠疫情 (COVID-19) 前、疫情期间和疫情后三个时段三种商品之间的动态关系变化。分析发现，三个时段呈现不同的动态关系：疫情前的

市场稳定且可预测，商品价格之间温和互动；疫情期间的商品价格波动性和相互关联明显增强；疫情后的市场稳定在一个新的平衡态，三者关系动态性受地缘政治和能源政策的影响。研究还表明，汽油和乙醇相互影响，且汽油对乙醇的影响要大些，符合乙醇作为汽油替代品的地位；同时，乙醇对糖的影响要强于汽油对糖的影响，因为汽油是通过乙醇间接影响糖。基于 CE 的 TE 方法成功揭示了巴西糖、乙醇和汽油三种大宗商品价格之间关系，以及疫情对三者关系的影响，证明其是一种动态经济关系分析的有力工具。

库存管理是企业运营管理过程中的关键环节，也是管理学的重要问题之一。报童问题是典型的单周期库存管理模型，一直是本领域研究的焦点。近年来，利用数据驱动模型和方法的报童问题研究展现出比传统方法的优越性，进而成为了热门话题。Tian 和 Zhang [170] 提出了一种端到端的算法框架，利用深度学习模型从在线商品评论等特征数据中预测订单数量，其中采用了包括 CE 在内的方法来选择模型的输入特征。他们将方法应用于汽车库存管理问题，基于 2016-2022 年的大众朗逸汽车的历史销售量、某网站的评论、某搜索引擎指数、和宏观经济指数等数据构建了模型。结果显示，本方法能够大幅减少超额成本和短缺成本之和，与同类方法相比减少了 31.8% 的成本。

中国企业海外并购面临着时代的机遇和挑战。探究影响中国企业海外并购的国内外各种因素，分析并购的短期和中长期绩效，具有重大的理论和现实意义。王琳君等 [171, 172] 提出利用 Copula VECM 模型，分析与海外并购数量强关联的经济变量对并购的影响，特别考虑了被其他研究者忽视的宏观经济变量的动态影响。由于此类经济变量较多，容易使构建的 VAR 模型复杂度增加，导致估计模型的不准确性。因此，他提出利用 CE 对经济变量进行选择后再建立模型。他在 Wind 数据库中选取了海外并购数量和其他 7 个与并购数量可能关联的宏观经济变量的季度数据，通过 CE 关联度分析后，得出结论认为宏观经济杠杆率、GDP、货币供给增长率和汇率四个宏观经济因素是影响我国海外企业并购活动不可忽视的重要因素。他进一步分析论述了所选变量对并购数量影响的内在经济逻辑，增强了模型的合理性。

## 6.34 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。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 [15] 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。

### 6.35 教育学

高中教育各学科之间具有内在的联系，教学大纲中强调了数学对物理、化学和生物等学科的基础性地位，数学知识、数学思维和思想方法深刻地渗透影响着其他学科的教学。因此，数学成绩被认为与其他学科成绩具有相关性。利用实证的方法研究数学与其他学科的关系，分析数学成绩与其他成绩之间的相关性是一个重要的基本问题，对于教学改革和学习方式的选择具有普遍参考意义。柳琼 [173] 基于某市 2013 级理科学生高一、高二期末考试成绩和高三两次模拟考试成绩，研究了数学成绩与其他学科成绩之间的相关性。作者比较了经典线性相关系数、秩相关系数和 MI 三种相关性度量方法，从 CE 和 MI 理论关系的角度分析论证了 MI 度量的优越性，并实验证明了 MI 度量能够更好地刻画揭示数学对其他不同学科（语文、英语、物理、化学和生物等）的影响力机制。

### 6.36 计算语言学

城市服务热线是政府公共管理系统的重要组成部分，促进了政府和市民的沟通，改善了政府的公共服务。但传统的人工派单方式无法满足日益增长的热线诉求，如何高效快速的处理大量的市民热线诉求是城市服务热线提高服务质量面临的重要课题。大量的热线文本数据积累为快速筛选和处理热线诉求提供了可能，可以利用自然语言处理方法处理热线文本数据，进而构建智能派单系统。陈作海等 [174] 提出了一种基于知识图谱技术的城市热线派单方法，基于城市热线数据构建热线知识图谱，再对待派单诉求根据构建的知识图谱检索结果进行派单，大大改善了热线服务的工作效率。在此智能派单系统中，CE 作为特征选择方法被用来对城市热线数据进行预处理，以构建和更新知识图谱。结果表明，CE 表现优于其他同类方法。作者将该方法应用在济南市民服务热线的系统上，通过不断更新知识图谱，最终获得了 90% 以上的派单准确率。

### 6.37 新闻传播学

公共卫生事件发生过程如何影响公众情绪是一个重要的问题，具有理论和现实意义，对政府的信息发布和舆情管控具有参考价值。特别是新媒体环境中，公众情绪的传播和演化过程受多种因素影响，因而更趋复杂。新冠疫情的发生给研究这类问题提供了条件。Zhang 等 [175] 研究了上海新冠疫情发生期间，疫情过程对公众情绪的影响特点和机理。他们以微博平台上“上海疫情”主题的数据为基础，研究了公众情绪的影响因素、时间演化以及疫情与公众情绪之间的因果关系。研究利用了基于 CE 的传递熵方法分析了疫情和公众情绪之间的因果关系，实证地发现了疫情过程对公众负面情绪的因果效应大于正面情绪，且正面情绪对负面情绪具有抑制效应。

### 6.38 法学

社区是基本的社会生活单元，社区治安管理与每个人的生活息息相关。社区属性与社区犯罪之间具有内在联系，分析社区经济、社会和人口等属性与各类犯罪之间的关系，可以加深对犯罪

行为发生理解，对执法部门合理安排部署资源力量具有重要参考意义。Wieser [176] 基于 CE 与 MI 的等价关系，提出了一种新的信息瓶颈（Information Bottleneck）估计方法。由于利用了 CE 的变换不变性，该方法较传统同类方法具有更好的估计性能。他将该方法应用于美国社区与犯罪数据集，分析 125 种经济社会因素与 18 种犯罪属性（包括 8 种犯罪行为，人均犯罪率和人均（非）暴力犯罪率）之间的关系，学习得到了可以表示这种关系的潜变量模型，为构建犯罪预测模型提供了参考。

## 6.39 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 [177] 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。

## 6.40 军事学

目标意图及时准确识别是战场态势感知的一项重要内容，是指挥决策的基础和前提。空中飞行目标意图识别会面临多种不确定性的挑战，如行为特性与物理特性的不确定性、飞行规则的不确定性和行动能力的不确定性等，使得及时准确的意图识别十分困难。张可等 [178] 提出了一种基于动态贝叶斯网络的目标意图识别方法，用于从复杂态势中目标的时序数据中完成意图识别，方法利用基于 CE 的 MI 估计算法从目标属性和目标意图数据来生成贝叶斯网络结构，再利用自适应遗传算法迭代优化网络结构，利用最终优化得到的网络来进行未知目标的意图识别。他们将该方法应用于空中目标的处理过程，利用空中目标的位置信息、飞行信息，以及雷达和通讯系统信息来识别其 6 种不同意图（巡逻、预警/指挥、电子侦察、电子干扰、攻击和打击等）。该方法可不限于空中飞行目标，可以很方便地推广到其他类型目标上。

## 6.41 情报学

颠覆性技术是具有原始创新性的技术，会对现有主流技术和产业产生变革性作用，推动经济社会发生突变式进步。开展颠覆式技术的前瞻识别及研判研究是科技情报分析领域的重要问题，对科技政策制订、科技产业布局和科技创新生态培育具有指导意义。基于知识网络分析的科学、技术和产业互动模式研究是解决识别研判问题的路径之一。许海云等 [179] 提出了一个颠覆性技术研究流程框架，以渐进式技术为参照获取科技、专利和产业文献资料的文本数据，利用自然语言处理技术分别构建三者的知识网络，再利用知识网络的三种整体网络属性和网络社区相似度属性将知识网络互动模式划分为预设的五种模式，包括科学-技术-产业联动模式。其中，CE 被用来度量三种知识网络的整体网络属性之间的关联度，以表征互动模式。他们以再生医学（干细

胞) 领域作为颠覆性技术对象, 以白血病治疗领域为渐进性技术参照开展实证研究, 获取了截至 2020 年底的权威数据库相关文本数据, 利用该流程框架研究了两个对比领域科学-技术-产业互动模式的共性和差异, 加深了对颠覆性技术创新生态要素的知识流动和扩散规律的认识。

## 6.42 能源电力

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 [180] 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。

光伏发电技术受天气等环境因素影响, 具有较大的不确定性, 给电网的安全稳定运行构成影响。根据气象条件等因素对光伏发电站有功功率进行预报, 有助于电网调度人员更好地制定调度策略, 应对光伏发电的不确定性给电网的冲击威胁。朱正林和张冕 [181] 提出了一种结合优化算法、模态分解、CE 和深度学习模型的方法, 用于提高发电功率的预测精度。他们在澳大利亚 Yulara 地区光伏电站数据上将方法与多种同类方法进行了对比, 表明该方法得到的模型能够更好地适应天气变化的影响, 取得最好的预测效果。

在天气明显变化的情况下, 光伏功率输出会发生剧烈变化, 从而导致目前预测精度下降。针对这种情况, 杨秀等 [182] 提出了一种基于天气类型划分的光伏功率日前预测方法, 首先将天气分为转折天气日和多种不同的平稳天气日类型, 再利用基于 CE 的 TE 分析得到不同天气类型下与光伏功率存在因果关系的气象因子, 最后利用多种神经网络融合模型进行点预测和区间预测。他们基于上海松江地区某光伏电场 2021 年的功率和气象数据, 利用该方法进行预测实验。实验首先将天气划分为转折天气日和四种平稳天气日(分别对应冬季晴朗天气、冬季阴雨天气、夏季晴朗天气和夏季阴雨天气), 再利用 TE 选择不同类型天气日相关的气象因子, 最后利用该方法的预测模型与 3 种同类模型进行了对比。实验结果表明, 基于 TE 分析发现, 转折天气日与光伏功率存在因果关系的气象因子有风级、降水、地表太阳辐射强度等, 而晴朗天气有关的气象因子为云量和地表太阳辐照强度, 阴雨天气有关的气象因子为风级、降水、地表太阳辐照强度和气压; 在此因果关系基础上构建的模型在不同天气类型下性能均好于对比模型。

光伏组件是太阳能发电站的核心设备, 其运行状态直接关系着电站的运行效率和稳定性, 如发电量、洁净度和故障状态等。因此, 预测光伏组件状态是光伏电站管理的核心问题之一。传统的状态预测方法大多只基于单个组件信息, 忽略了组件之间的联结关系, 导致预测准确度不高。王士涛和邹晨鑫 [183] 提出了一种光伏组件状态方法, 在考虑单个组件状态的影响因素的同时, 还考虑组件之间的联结关系, 用以构建状态图神经网络预测模型。其中, 他们利用 CE 估计不同

影响因素之间以及状态和影响因素之间的相关性，用于构建预测模型中变量之间的拓扑关系。他们基于一套光伏组件的样本数据验证了方法的有效性。

风能作为一种主要的清洁能源，具有间歇性和不确定性的特点，导致风电机组的功率预测和控制十分复杂。基于风电机组的监测数据，分析机组内各变量之间的相关性特征，有助于机组的健康状态监测和风电功率预测，从而更好地利用风能资源。崔双双和孙单勋 [184] 提出利用 CE 来分析风电机组状态变量之间的相关性，再基于 CE 相关性进行聚类以得到机组工况的划分。他们将方法应用于广东某海上风电场数据采集与监控（SCADA）系统的数据，发现 CE 方法较传统方法能更好地描述数据中的相关性，并利用 K-means 方法得到了能精确地反映风电机组运行特性和状态的工况划分，具有重要的现实意义。

电力负荷预测是根据历史数据来预报未来一段时间的用电量，对智能电网调度和规划电力输送具有重要意义。电力负荷受多种因素影响，具有周期性和季节性等特点，特别是受天气因素的影响明显。因此，构建准确的电力负荷预测模型需要考虑天气等多种因素，并对天气对负荷的影响特点进行分析。Ma [17] 提出利用基于 CE 的 TE 方法来分析动态系统的时延特性，并将方法应用于摩洛哥缔头万 (Tétouan) 城的电力消费数据，从时延的角度分析了五种天气因素对该城三个电力供应网络的负荷的影响，发现了影响的每日时延变化特征。Yan 等 [185] 提出了一种结合聚类算法、预测算法和集成学习方法的综合能源负荷短期预测方法，首先根据负荷数据特性对数据进行聚类，再对每类数据利用基于 CE 的 TE 算法分析选择对负荷有影响的外部因素（包括天气和时间两类），最后利用集成学习算法对负荷进行预测。他们将方法应用于 2018 年美国亚利桑那居民建筑综合能源负荷数据，以预测电力、燃气、制冷和供热四种负荷。实验结果表明，利用基于 CE 的 TE 算法选择的外部因素可以在预测模型上得到最好的预测性能，效果明显好于其他相关性变量选择对比方法，原因是 TE 可以准确度量外部因素和负荷之间的时序非线性关系。阚超 [186] 提出了一种基于深度学习的综合能源多元负荷短期预测方法，首先利用 VMD 对多元负荷进行分解，再利用 CE 计算分解得到的 IMF 分量与负荷影响因素之间的连接强度，作为图卷积网络的邻接矩阵权重，再将如此得到的时序耦合特征输入到 LSTM 模型，将由此得到的模型输出与另一个 Transformer 模型的输出进行点运算作为最终预测结果。他在美国亚利桑那州立大学坦佩小区的数据上验证了方法的有效性，发现 CE 能够很好地计算出气象和时间等因素与冷、热、电负荷各分量之间的耦合强度关系，增加了模型的可解释性。胡程林 [187] 提出了一种基于多任务学习的多节点符合预测方法。他首先分析了节点负荷间的相关性和因果性关系，特别是节点负荷与总负荷之间的关系。其中，他利用基于 CE 的 TE 方法分析探索了节点负荷与总负荷之间的非线性因果关系。他基于新西兰配电系统九个地区变电站负荷数据集进行了 24 小时因果分析，发现节点负荷对总负荷的因果关系是一个波动的因果关系过程，说明总负荷中含有节点负荷的变化信息。吴迪等 [188] 提出了一个基于混合机器学习的配电网负荷预测方法，该方法结合了 CE 和图神经网络等多种方法，其中 CE 用于计算电网负荷及其影响因素之间的相关性强度，以构建图神经网络结构。王伊佳 [189] 提出了一种并行 CNN-GRU 注意力模型，用于基于多元时间序列数据的电力负荷预测，方法中采用 CE 等多种方法筛选模型的输入因素。他在国内某地区的电网数据和相应的气象数据的基础上验证了该方法，得到了比对比方法更优的预测性能。其中，CE 值不仅度量了输入变量和负荷之间的统计相关性，还反映了底层系统中的信息传递和能量交换，具有比传统相关系数更丰富的信息。

可再生的风光能源越来越成为电力能源的重要组成部分，如何保证风光电力接入的经济效益和安全可靠是可再生能源利用的主要关切。合理的规划对于解决此关切十分关键，可保证建设投资回报和系统合理运行，防止风光能源被弃用的发生。储能系统可以平抑风光能源的不稳定波动性，是风光系统规划的组成部分。董海燕等 [190] 提出了一种考虑源荷时序相似性的风光储协同规划配置方法，其中利用 CE 衡量风光能源与负荷之间的相似性，以提高系统风光能源的利用效率。他们将方法应用于某工业园区的风光火储联合发电系统的规划配置，结果表明，该方法能有效降低储能系统的装机容量，提高新能源的消纳能力，经济效益和减排效益明显。

频率是电力系统最重要的物理量指标之一，频率稳定性是保障电力供应稳定性的一个基本要求。可再生能源由于具有不可预测性，其大量接入电网给电网频率稳定性带来了挑战。为了稳定和控制新能源带来的频率波动，需要准确快速地预测系统的频率稳定性，以帮助系统操作员提前制定控制策略。传统的频率稳定性预测是模型驱动的，由于求解耗时从而无法做到在线预测。基于机器学习的模型方法，通过简化模型以提高计算效率，可以满足在线预测的需求。Liu 等 [191, 192] 提出了一种结合深度学习和 CE 的频率稳定性预测方法，CE 被用来选择模型输入变量，减少冗余信息以提高计算效率。作者将方法应用于两个系统：一个是新英格兰 39 节点系统，集成了美国西部电力调度委员会的动态风场模型；另一个是基于南加州西部的电网系统建立的 ACTIVSg500 系统。实验表明该方法建立的模型相较同类模型取得了最好成绩，达到了实用的要求。CE 方法不仅简化了模型、大幅降低了计算时间，且分析发现了与频率稳定性相关的电网变量，使得模型具有了可解释性。

电力系统宽频振荡由电力电子设备的动态交互作用引发，在电网中的传播会造成连锁反应，严重危害电网安全运行。宽频振荡激发机理复杂，具有显著的时变、非线性和广域传播等特征，难以有效地进行建模分析。冯双等 [193–195] 利用 CE 的模型无关特性，提出了一种宽频振荡影响因素和传播路径分析方法。该方法以系统运行的状态参数为随机变量，通过计算其与各个频率区间的振荡阻尼之间的 CE 来选取影响振荡的关键因素；同时，利用系统发生振荡时的数据，计算系统变量之间的 copula 传递熵网络，用于分析振荡的传播过程和振源定位。该分析方法是数据驱动的方法，可以在系统模型未知的情况下得到相应的分析结果。作者仿真了直驱风机并网系统和含风电场的四机两区系统，对控制器内部各环节和复杂系统各母线之间的振荡因果关系进行分析。仿真结果表明，该方法能够从设备级和网络级两个层面准确确定宽频振荡的传播路径和振源位置，为研究振荡传播机理提供了支撑，为进一步采取振荡抑制措施提供了参考。孙文涛等 [196, 197] 也提出了一个利用 CE 识别交直流混联系统宽频振荡风险识别方法，通过分析计算振荡影响因素变量与各个子频率区间内振荡模态的阻尼变量之间的 CE 来进行风险识别。他们利用该方法分析了某省份电网系统在小扰动下的振荡风险，采用 LCC 模型发现了整流器控制参数和直流传输功率等关键影响因素，为后续设计抑制振荡的针对性调整方案提供了准确且可靠的依据。

新能源的接入给电网带来了动态性和不确定性，增加了电压协同控制的复杂性，给电网运营带来了新的挑战。电压控制要求多个变压器的分布协同控制，但通信网络存在时延、抖动和丢包等信息传输问题，给这种分布式协同控制带来了不利影响，导致性能下降甚至构成系统不稳定风险。因此，研究通信不确定性下的电压协同控制是一个重要问题。由于网络信道特性和拥堵的原因，时延、抖动和丢包等问题之间存在非线性的相关性，这种相关性在以往的研究中没有被充分

考虑，导致可能的次优控制甚至系统不稳定。Yang 等 [198] 提出了一个多变量相关通信不确定性事件条件下事件触发的滑模控制策略，用于电网的电压协同控制。作者提出利用 CE 来度量时延、抖动和丢包三种事件之间的不确定相关性，进而得到联合熵用于衡量总的通信事件不确定性，再将这样得到的总不确定性用于滑模控制的状态观测器设计，进而得到滑模控制器的控制律。他们在江苏电网的一个 7 总线配电系统和一个 IEEE 69 总线配电系统上进行了仿真和实物实验，将提出的滑模控制策略与 3 种同类方法进行了对比，结果表明，所提出的算法展现出了良好的控制性能和鲁棒性，各种指标均优于对比方法，使其更适合复杂通信环境下电网的协同电压控制。

线损率是电力能源企业的一项重要经济技术指标，衡量其经济效益水平的高低。因此，线损管理和异常线损稽查是电力部门的一项重要工作。线损分析是利用科学的计算手段分析线损在电网中的分布规律，能为管理提供高效、准确的决策支持。Hu 等 [199] 提出了一种基于 TE 的线损分析方法，通过 CE 估计计算每个用户对区域总线损的 TE 值来判断其对总线损的贡献。他们基于每日电力供应和线损数据的计算分析，将用户根据线损贡献度排序，以应用于实际线损管理工作，从而减少总线损率。

配电网拓扑辨识是电网系统分析的重要问题，为潮流计算、电网状态估计、无功优化调节和网络重构等配电网管理功能提供基础。随着分布式能源大规模接入配电网，其波动性和不确定性导致系统拓扑重构更加多变，给拓扑辨识带来了新的难题。秦超和潘毓笙 [200] 提出了一种新的配电网拓扑辨识方法，基于时空相关性将辨识问题转化为多个开关节点状态识别的子问题。该方法首先利用 CE 和马尔科夫链分别提取节点电压序列之间的空间和时间非线性相关性特征，在此基础上得到能够识别单个开关状态变化序列的模型，最后结合多个此类开关状态识别结果完成一定时间内的网络拓扑结构辨识。他们模拟了接入风机和光伏的拓扑结构动态变化的配电网，为其仿真生成了为期 120 天的配电网家庭负荷，在此网络节点量测数据的基础上检验所提出的方法，结果表明 CE 能够有效分析节点电压之间的相关性，导致该方法能够在短时间内有效辨识网络拓扑结构。

电价预测问题在电力市场参与者决策中至关重要，可以帮助其开发交易策略并合理分配资源。但新能源的广泛使用使电力供应具有不确定性，从而使电价预测变得更加复杂，造成预测模型构建较为困难。Xiong 和 Qing [201] 提出了一种基于时序数据的混合电价预测框架，将基于 CE 的特征选择方法与信号分解、贝叶斯优化和 LSTM 模型相结合，以构建预测模型。他们将方法应用于 2017 年美国宾夕法尼亚州-新泽西州-马里兰州互联网络（PJM）电力市场数据上，证明了该方法的有效性和实用性。

锂电池是使用最广泛的绿色清洁能源。但锂电池的电池容量会随着使用次数而退化，因此电池健康状态监测是电池管理系统中的主要问题之一。传统的健康状态监测模型大多在单一负载状况假设下得到，无法适用于真实场景下的多种状况，导致在原始数据上得到的模型无法适应新的情况。针对此问题，Hu 和 Wu [202] 提出了一种基于迁移学习思想的电池容量估计方法，结合了因果分析、注意力机制和 LSTM 等工具，其中基于 CE 的 TE 被用于选择与容量退化相关的健康状态指标，以保证构建模型在不同状况下的可迁移性。作者将方法应用于 NASA 的 3 种负载状况下的锂电池退化数据，结果表明，基于因果分析构建的模型比基于两种传统方法的模型的跨工况预测准确度分别提高了 8.6% 和 12.4%，增强了模型的鲁棒性。赵长胜 [203] 提出了一种电池健康状态预测的方法，该方法包括基于 CE 等多种方法相结合的多维特征选择框架、结合深

度学习时序预测模型和集成学习方法的模型构建和调优方法，以及适应不同电池和不同工况的模型迁移学习方法。他基于公开的 CALCE 和 NASA 锂电池退化数据集验证了该方法，结果表明，该方法得到的模型预测误差在 2% 以内。他特别对所提出的特征选择方法的冗余性和充分性进行了实验分析，结果表明，该特征选择框架给出的模型具有适应不同工况和不同电池的良好泛化性和鲁棒性，在数据少的情况下依然具有优异预测性能。

能源效率是工业 4.0 的主要目标之一，生产系统的数字化给提高工业设备的能源效率提供了巨大的机会。能效异常是改善能源效率的突破口，发现异常并给出其原因是改善能效的有效途径。然而工业系统大都具有复杂的结构和运行机理，难以通过传统建模方法分析能效异常的根本原因。马健 [204] 提出利用 TE 对能效异常进行根因分析，针对工业系统的非平稳性，给出了一个称为 TE 流的能效异常原因诊断方法，其中采用了基于 CE 的 TE 估计方法。由于 TE 是模型无关的，该方法也就可以在无设备机理的条件下对各种设备进行能效异常根因分析。他将该方法应用于一个空气压缩机系统，成功地对系统运行的因果关系进行了描述，从而找到了导致系统能效异常状态的空压机子系统。

## 6.43 纺织工程

棉纱布作为基础纺织产品，也是一种常用医疗器械，广泛应用于如手术过程的无菌隔离和止血等医疗服务过程中，其生产品质直接关系到医疗服务效果和患者安全。因此，纱布生产过程的质量控制都受到政府行业部门的严格监管。密封力是指纱布生产打包过程中的密封强度，对于保持纱布开包使用前的无菌和免受污染至关重要，因此是纱布生产过程中的重要质量控制参数。Mortezaejad 等 [205] 提出了一种基于 CE 的非参数多变量质量控制图方法，面向多维非正态分布变量的质量控制情况。该方法首先利用最大 CE 法估计多维分布函数，再利用 Hotelling  $T^2$  统计量生成控制图。他们将该方法应用于 Mega 公司吸水棉纱布生产过程的密封力数据，生成了基于 Copula 的质量控制图，能够检测传统质量控制图方法难以检测到的微小质量变化，从而证明该方法能够对棉纱布生产质量参数进行控制。

## 6.44 食品工程

葡萄酒作为一种奢侈农产品，越来越走进广大普通消费者。葡萄酒质量的品鉴对其生产和销售都至关重要，葡萄酒酿造业大量投入在质量评价环节，以改善酿造工艺并促进消费。传统的质量品鉴主要依靠理化测试和专家感受，但专家的味觉感受主观性较强，其内在机理难以理解。因此，有必要研究酒的成分和专家评价之间的内在联系，以增进对葡萄酒质量的理解，提高质量评价的客观性。Lasserre 等 [206, 207] 利用基于 CE 的（条件）独立性度量估计，提出了一种因果关系网络学习算法，称为 CMIIC，并将其应用于著名的葡萄牙绿酒的质量评价数据上，分析发现了分别与红葡萄酒和白葡萄酒的质量相关的理化成分。

## 6.45 土木建筑

建筑能源消耗占全部能源消耗的四成左右，建筑节能技术是重要的绿色能源技术，对实现联合国的碳中和目标意义重大。供暖、通风和空调（HVAC）系统贡献了商业楼宇四成以上的能耗，是建筑节能的主要研究对象之一。HVAC 系统的运行具有时延的特性，来自于媒介传导的滞后和热惯性。理解并运用这种特性，有利于设计适当的控制策略，从而达到节能的目的。Li 等 [208] 将基于 CE 的 TE 理念方法引入到 HVAC 领域，开发了一种基于信息论框架的无模型时延估计方法，用于 HVAC 系统的时序预测。他们改进了 kNN 的多变量 TE 估计器，结合优化方法设计了时延估计算法。他们将算法应用于大连某四层教学楼的供热监控系统，分析室内温度与天气参数（如室外温度、相对湿度、太阳辐射、风速等）和供热参数（如热水供应和回流温度等）的数据，辨识时延参数，进而利用后两组参数预测下一段时间的室温。结果表明，TE 方法能够辨识参数之间的时延关系特性，进而提高室温预测性能。

钢筋混凝土结构是广泛应用的建筑结构形式，而地震会对建筑物的钢筋混凝土结构造成显著破坏，因此评估这种结构的抗震性能关系着人民的生命和财产安全。塑性转角和挠度一般被用于钢筋混凝土梁的地震破坏评估和抗震性能等级分类，因此如何评价钢筋混凝土梁的地震性能等级极限就是一个重要的问题。然而，不同国家有着各自不同的评价方法，使得其不能用于描述地震带来的结构渗漏，特别是地下结构渗漏。Ma 等 [209, 210] 基于机器学习方法提出了一种考虑了裂缝发育的钢筋混凝土梁地震性能等级预测方法。他们采集了 452 个钢筋混凝土梁的试验测试结果，再利用皮尔逊相关法分析得到预测的力学参数集合，同时利用基于 CE 的 MI 方法选择得到与性能极限具有非线性关系的力学参数集合，最后利用 7 种机器学习方法建立性能等级极限的预测模型。实验结果显示，皮尔逊相关法选择了 27 个中的 22 个与性能极限具有线性关系的力学和维度参数，而 MI 法选择 5 个与性能极限具有非线性关系的参数；基于以上选择的参数，作者构建了 4 个性能极限的极限预测和等级分类模型，得到了良好的预测和分类结果，验证了该方法的有效性。

受剪承载力是结构设计中的重要参数，指结构对剪切力作用的承载能力，其预测对于评估结构的安全性和稳定性至关重要。现有的钢筋混凝土柱的受剪承载力预测大多是基于机理与经验相结合的模型，而实际受剪性能机理复杂，非线性影响因素多，导致已有模型的预测结果准确性不高。利用机器学习方法构建钢筋混凝土构件受剪性能预测模型是一个解决问题的有效途径。常旭等 [211] 提出了一种利用基于 CE 特征选择的受剪承载力预测模型构建方法，其中 CE 被用于对受剪力及其影响因素的非线性关系进行度量，以便为机器学习预测模型选择合适的影响因素输入变量。他们收集了包含 441 组样例的钢筋混凝土柱受剪试验数据，基于此对所提出的预测模型构建方法进行了验证。实验结果表明，CE 方法对受剪承载力影响因素进行了合理的选择，基于此得到的预测模型的预测精度要好于 5 种传统的半经验半理论公式，验证了该方法的有效性和优越性。

房间作为建筑基本单元，承担着实现建筑功能的角色，这来自于房间内设备和设施的协同发挥作用。理解地震损坏机理，特别是地震中由于房间内元素协同作用导致的损坏后果，对于震后建筑和房间功能评估具有至关重要的作用。Copula 理论作为变量间相关性的表示方法，特别是 Vine Copula 方法，是房间内元素相关性建模的有力工具。但传统的 Vine Copula 构建方法依靠

相关系数等工具，导致结构构建过程中不确定性的累加，同时这样隐含作出的高斯性假设也不符合实际问题的情况。Liu 等 [212] 利用 CE 理论，提出了基于 CE 和模型选择的 Vine Copula 构建方法，用于对房间内元素相关性建模，从而评估房间的震后功能性。其中，CE 用于构建 Vine Copula 的基本结构，将结构学习和函数估计分开进行，增加了 Vine Copula 模型的可靠性和准确性。他们以医院手术室为研究对象，构建了实物仿真系统，进行了不同振动强度下四种室内情景设置的振动台试验，采集了试验中手术室医疗器械和设施的视觉、加速度和位移等数据。他们利用试验数据建立了房间元素损坏状态模型，进而又基于 Vine Copula 得到了房间功能失效状态模型，发现利用此方法估计的模型给出的房间系统脆弱性仿真结果与振动台测试结果基本一致，证明了利用 Copula 函数对房间元素协同作用建模进行房间系统脆弱性评估的有效性。该方法对于医院建筑功能性评估具有重要意义，也可以扩展到具有更复杂功能性设置的房间系统。

隧道是构建在地下、水下或山体中的交通建筑物，盾构施工是隧道建设的关键技术。但盾构施工会受到地质条件和机械设备等因素影响，造成掘进路径和设计线路之间的偏差，从而影响隧道建设的进度和安全。目前的盾构控制主要靠人的经验，难以有效应对复杂的施工环境，因此有必要研究设计方法对盾构轴线偏差进行提前预测。林超 [213] 提出了一种基于 CE 的盾构轴线偏差预测方法，利用 CE 选择对预测有效的盾构掘进参数。他基于南昌市地铁 3 号线轨道交通线路施工某区段的掘进数据验证了该方法，利用 CE 等方法对 40 个盾构掘进参数进行选择，构建了切口水平偏差、切口高程偏差、盾尾水平偏差和盾尾高程偏差四个预测模型。分析结果表明，与同类方法相比，CE 选择的特征集合更精简，且与盾构偏差所处的方位相符合。在训练数据集、测试集以及验证集上的结果都表明，基于 CE 的方法在选择最少特征的情况下给出了最优的预测性能。

## 6.46 交通运输

大件货物运输是指通过多种运输方式对具有不可拆解属性的大型物件的专业运输作业活动，在国民经济中占有重要地位，对国计民生重点行业的基础设施建设起着重要的支撑和保障作用，也关系着国防军事和国家安全。大件货物运输大都需要铁路、航运等多式联运的方式才能完成，需要制定各个局部运输环节模块联动的整体方案。随着交通系统的数字化，大量的相关方案数据得到积累，基于数据的大件货物运输方案制定成为了一个重要的问题，其研究有助于提高方案制定的科学性和适用性。黄达 [214] 利用 CE 等多种数学工具提出了一种基于模块链构建的大件货物多式联运方案制定方法。该方法先将运输方案分解为多个局部环节模块，再利用 CE 等相关性度量工具筛选一组模块属性用于计算方案之间的相似度，最后在已有运输案例库中检索与目标运输任务相似度高的案例作为初步运输方案。由于大件运输方案的多样性，一些案例模块属性会具有非高斯性，使得传统的相关系数工具不再适用于计算属性间相关性，而 CE 由于具有普适性则依然适用。作者在 600 多个实际案例的数据上验证了该方法，并构建了方案制定原型系统。

航空和高速铁路是我国最主要的两种旅客运输方式。相较于航空，高铁票价的市场化水平处于落后的水平，欠缺灵活性和动态性。因此，研究影响票价的因素以期改进高铁票价的定价机制是学界十分关心的问题。许罗豪等 [215, 216] 基于京沪航空和高铁票价的数据，利用 CE 和决策树等工具研究了出行需求、旅客选择、出行效率和出行路线四类因素对航空和高铁票价的影响。

他们发现购票提前期对两种票价的影响程度不同，但旅行时间对二者的影响程度较为相似。这些研究结论对高铁定价具有一定的参考价值。

城市轨道交通已经成为我国各大城市的主要交通出行方式之一，提升城市轨道交通系统的管理水平和运营效率是交通系统面临的重要问题之一。城市交通客流分析与预测可以为正常客流引导、异常客流疏导和轨道列车调度提供依据。基于出行记录数据分析轨道交通和公交、出租车等其他交通方式客流之间的互动关系，有助于提升轨道交通客流预测效果。王升 [217] 提出利用相关分析和因果分析等方法对客流时序数据进行分析，以增进对不同交通方式客流之间关系的理解。其中，基于 CE 的 TE 方法被用于客流间因果关系分析。他将方法应用于苏州市轨道交通系统四个站点 2018 年 8 月 6 日至 12 日期间的轨道交通、公交和出租车客流时序数据，因果分析结果表明，三元坊和东环路站的出租车客流到轨道交通进站客流的影响有 1 小时的滞后效应，而东方之门站的这种滞后效应则有 5 小时。这一分析结果对轨道交通站点的客流预测具有重要指导意义。

铁路客流量预测是铁路客运服务管理的基础，准确的预测可以改善铁路运力的统一调度、协调路网资源和提高经济效益。但客流受自然和社会因素共同影响，准确预测具有一定的难度。作为一个典型的时间序列预测问题，一般采用时序模型来完成预测，这其中一个关键的问题就是如何处理客流和其外部影响因素之间的非线性关系。Chang 和 Song [218, 219] 提出了一种改进的 Prophet 客流预测模型，其中利用 CE 来分析天气因素和节假日因素与客流之间的非线性关系。他们利用 2015 年 1 月至 2016 年 3 月期间的真实铁路客流数据进行了实验研究，利用 CE 相关性分析发现天气因素对客流的影响可以忽略不计。他们又基于 CE 工具构造选择了新的节假日时序特征，用于提高预测性能。实验结果表明，利用如此改进的 Prophet 模型可以提高客流预测的准确性。

疲劳驾驶是导致交通安全事故的主要原因之一。疲劳检测技术通过对驾驶员疲劳状态的监测和预警，可以有效提升驾驶安全，具有重要的现实意义和社会价值。通过脑电信号进行驾驶员疲劳状态检测是一个主要的技术方向，得到了大量研究。但脑电信号具有动态性和非线性等特点，如何提高疲劳检测的准确度是一个本领域面临的主要难题。周泳江 [220] 提出了一种基于拓扑脑电特征选择和融合的疲劳检测方法，首先通过经验模态分解得到脑电信号的模态分量，再在此基础上计算脑功能性网络并提取拓扑特征，然后利用 CE 方法进一步对提取的特征进行选择，再将所选择的特征进行融合并输入机器学习疲劳检测模型。他基于公开的驾驶员疲劳脑电数据验证了该方法，证明了该方法的有效性，得到的疲劳检测识别率为  $93.98 \pm 3.36\%$ 。实验中，CE 特征选择方法提升了所有实验模型的识别准确率，证明了方法的有效性和普适性。

## 6.47 机械制造

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 [221] 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。Cai 和 Rong [222] 提出了一种鉴别影响质量的关键因子的方法，首先利用 CE 建立因子间相关矩阵，再用网络反卷积方法消除因子之间的间接影响，从而鉴别出影响质量的关键因子。他们将方法应用于 UCI 机器学习库的三个数据集，结果表明该方法能够较同类方法更高效地鉴别关键因子并取得最高的预测准确率。他们又将方法应用于一个薄膜晶体管液晶显示器生产的数据，结果显示，该方法从 1540 个因子中选出 154 个因子，并得到了最好的质量预测精度。

复杂机械产品的整机制造包括设计、制造和装配三个环节。作为产品生产的最后一个环节，装配过程在零部件的制造过程基础上组装高精度产品，装配质量控制在零部件制造质量的基础上保障整机产品质量。复杂机械产品零部件数量种类繁多、相互关联，装配环节错综复杂，上游环节的装配质量误差会对下游环节质量构成影响。王小巧 [223] 在装配质量控制中考虑了上下游工序和质量控制点之间的相关性，利用 Copula 对控制点间相关关系建模，并用 CE 度量这种相关性，进而提出了一种装配质量控制点控制阀优化方法。她将方法应用于江淮汽车某型汽油发动机关键零部件缸盖的装配工序过程，验证了方法的有效性。

现代工业系统变得越来越高度复杂和自动化，使得工业过程监测变得愈加困难。如何监测系统异常并发现异常原因是一个具有广泛应用的重要问题。利用因果分析得到工业系统内部复杂的因果关系图，有助于准确发现异常的传播路径，进而及时进行干预。Dong 等 [224] 提出了一个结合动态 PCA、TE 和 LSTM 的故障分析框架，其中基于 CE 的 TE 被用分析系统内的因果关系。作者将该方法应用于辽宁鞍钢的热轧带钢工艺过程数据的分析，成功地对过程中的两个故障及其原因进行了分析。作者还将基于 TE 的因果图分析方法与同类格兰杰因果分析方法进行了对比，表明 TE 方法能够更准确地对故障进行根因分析。刘鹏阳等 [225, 226] 提出了一种动态过程分布式监控的 CE-DR-SVDD 方法，首先利用基于 CE 的 Louvain 算法对系统变量分组，再利用动态递归支持向量数据描述算法构建局部监控模块，最后利用贝叶斯推理融合局部监控结果来得到全局监控结果。他将方法应用到田纳西伊斯曼过程的实验数据上，并与同类方法进行了对比，结果发现该方法在仿真的 21 个故障中的 19 个上获得了最好的检测结果。

烧结过程 (Sintering Process: SP) 在钢铁工业中至关重要，同时也会消耗大量的能源。动态预测 SP 的碳消耗有助于节约能源和减少碳排放。传统的 SP 建模基于一定的假设，无法适应 SP 的系统动态特性，基于数据的机器学习模型可以克服传统模型的不足。Hu 等 [227] 提出了一种动态建模方法框架，可以自动识别过程工况状态，从而进行碳消耗预测。该方法框架结合了 AKFCM 聚类算法、基于 CE 的模型选择和宽度学习模型方法。作者在一家钢铁企业的实际数据上验证了方法的有效性，证明了 CE 可以快速地捕捉不同工况下 SP 中复杂的相关关系模式，从而使该方法能够比传统方法更准确地预测烧结碳消耗。

航空发动机是航空飞行器的核心组件，其制造工艺是航空制造业技术水平的集中体现。涡轮盘是航空发动机的关键核心部件之一，对发动机性能起着决定性作用，因其在高压、高温和高速运转的极端条件下工作，容易产生疲劳退化，要求其具有高可靠性，因此对其制造过程提出了极高的要求。通过制造工艺参数优化提高涡轮盘的制造质量是得到高可靠性部件的有效途径。模锻是涡轮盘制造的核心工艺过程，决定着涡轮盘的质量和性能，因此对模锻工艺参数进行预测和优化就成为了一个重要问题。李家宝等 [228, 229] 提出了一整套涡轮盘制造工艺参数的预测和优

化方法，用于分析制造过程数据、预测制造质量，从而优化改善制造工艺。其中，CE 被用于计算工艺参数之间的相关性，对工艺参数进行聚类分组，以利于后面的预测建模和参数优化。他们在 GH4169 合金的涡轮盘制造过程中的模锻工艺上验证了该方法的有效性，基于 CE 的聚类方法成功将工艺参数进行了分类约简。

## 6.48 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 [230] 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。

砂轮是数控磨床的关键核心部件，用于对工件表面进行磨削加工作业，其物理磨损程度直接影响加工质量和效率。因此，砂轮的维修和保养十分重要，如何对其进行预测性维护是一个关键的问题。程毅 [231] 提出了一种基于 CE 和最大相关最小冗余的特征选择方法，用于构建砂轮剩余寿命预测模型。他基于威孚高科 CPM2.2 凸轮轴生产线上 5 个磨床上 55 个参数的 SCADA 数据，对比了多种相关性特征选择方法，发现基于 CE 的方法能够有效地计算出传统相关性方法不能发现的非线性特征关系，得到的 15 个参数与砂轮剩余寿命密切相关，符合磨床运行机理。

机械设备是交通、制造等领域的核心功能模块，滚动轴承则是各种机械设备的基础性关键部件。轴承的老化会导致其精度急剧下降，从而造成机械设备的当机、重大损失，甚至严重事故。因此，预测滚动轴承的剩余使用寿命对机械设备的安全和维护至关重要，利用机器学习构建此类预测模型是一个重要的方法。但传统的预测方法存在无法处理轴承特征与剩余寿命之间非线性关系、未考虑测量信号中的时序信息、以及不能处理数据分布迁移等问题。Meng 等 [232] 提出了一种将 RSA-BAFT 模型与 CE 相结合的滚动轴承剩余使用寿命预测方法，其中 CE 被用于度量时频域测量特征与剩余寿命之间的关联强度以进行特征选择。他们在 XJTU-SY 滚动轴承数据集上验证了该方法，结果表明，CE 方法能够选择与剩余寿命具有非线性关系的特征，基于 CE 选择的特征构建的模型比基于同类特征选择方法构建的模型具有更好的预测性能；在 CE 特征选择的基础上得到的 RSA-BAFT 模型的预测性能也要好于同类对比方法。

风电机组是开发利用风能的关键核心设备，但由于运行环境恶劣、运转负荷大且工况多变等原因，导致其发生故障的概率高，维修的成本高、难度大。因此，对机组的状态监测和故障预警成为解决问题的主要技术手段。风电设备的 SCADA 系统采集了机组运行的历史数据，基于此类数据开发故障预警技术是当前的主要研究方向。耿妍竹 [233] 提出了一种结合聚类算法、CE 和机器学习技术的风电机组故障预警方法，首先采用 DBSCAN 聚类算法进行数据清洗，再利用 CE 等方法选择与齿轮箱油池温度相关度大的 SCADA 运行参数，最后利用 9 种机器学习算法建

立齿轮箱油池温度预测模型。他们基于河北某风电场机组 2018-2019 年历史数据验证了该方法，实验结果表明，该方法能够得到较高的模型预测精度，能够提前 2-3 小时发出准确故障预警，证明了该方法的有效性。

## 6.49 石油工程

煤层气是蕴藏在煤层中的一种天然气，通常通过向煤层钻井来开采，在我国已经有 30 多年的开发历史。煤层气产量受开采过程中的多种地质和工艺因素影响，鉴别出与其密切相关的关键因素对作出煤层压裂开采决策至关重要。但煤层压裂有效性和影响因素之间关系具有非线性特征，且因素之间相互影响，给关键因素分析和产量预测构成了挑战。Luo 和 Xi [234] 提出了一种将基于 CE 的特征选择方法、混合优化算法和神经网络方法相结合的方法，用于构建煤层气井产量预测模型。他们在鄂尔多斯盆地某地块的气井数据上验证了该方法，利用基于 CE 的方法从 36 个备选因素中得到了 4 个关键因素，包括与气层相关的含气量和气饱和度，以及与压裂工艺相关的前置液量和含沙液量，由此得到的预测模型能够在实验数据上预测最高 83% 的产量。作者在同一地块的一口典型钻井上应用该方法选择地层中的高产气量区域，使得该井的日产量从每天 322 立方米大幅提高到 950 立方米。

陆地原油生产包括陆地勘探、原油开采和加工、原油存储和运输等过程。为了保证原油的质量，生产过程会使用大量的能源，从而造成大量温室气体排放。因此，石油生产部门需要准确地量化生产过程的碳排放，进而施行全过程减排降碳。Yuan 等 [235] 基于胜利油田的实际生产数据，分析了原油生产全过程的碳足迹，明确了各个环节的碳排放贡献和排放类型，并利用 XGBoost 和 CE 相结合的方法分析了过程中电能消耗和燃料消耗的主要影响因素，进而针对重要影响因素给出了减排降碳的建议。研究将该分析方法与其他 3 种同类分析方法进行了对比，发现该方法对影响因素重要性的排序最为准确。

## 6.50 矿业工程

煤炭是我国的第一主要能源，到 2030 年我们煤炭消费占比将达 55% 左右。我国的煤炭年产量接近 40 亿吨，其中 90% 为地下开采，面临地质灾害事故多发，大采深、采掘难等突出难题。无人化智能开采是煤矿开采的重要技术方向，可以降低安全事故并提高开采效率。煤-岩识别作为煤炭开采领域的国际难题，其解决对于无人化智能开采具有重要意义。但煤矿井下开采工作面条件复杂，地质条件多样，现场信号精度和质量低，给煤-岩识别问题增加了难度。已有的研究从多种信号出发，以期准确地感知工作面煤-岩介质属性。高峰 [236] 提出了一个将可见光、近红外光谱和采煤机截割物理信号多模态融合的煤-岩识别技术解决方案。其中，他通过模拟掘进试验采集了振动、扭矩和压力等采煤机截割信号数据，提取了 689 个时序信号特征，再利用 XGBoost 和 CE 相结合的特征选择方法，选择了其中的 45 个特征，以在截割和进刀工况下进行煤-岩识别。他开发了面向截割轨迹规划的工作面煤-岩识别解决方案，其中利用上述基于截割信号的识别方法进行实时岩性感知。他在陕西彬长矿业集团大佛寺煤矿巷道掘进工作面进行了现场

技术验证试验，基于截割信号的煤-岩识别方法的验证正确率为 91.14%，且识别错误均发生在煤岩交界附近，取得了良好的工程应用效果。

重介质选煤是煤炭生产过程的重要工序，主要是利用一定密度的介质对精煤和矸石进行分选。由于整个流程包含多道工序，且具有非线性、动态性和多变量耦合等特点，特别是过程变量间的时延的存在，使得对工序过程进行实时监测、分析和控制具有很大难度，而传统的相关系数等方法不能准确地估计这种时延。建中华和代伟 [237] 提出了一种将进化算法和 CE 相结合的时延估计方法，称为 EVO-CE，用于估计重介质选煤过程中的时延参数。他们将该方法应用于山西某选煤厂 2024 年期间的重介选煤生产过程数据，对过程中的四个关键过程变量相对于精煤灰分的时延参数进行估计，并将该方法与包括 CE 在内的 4 个同类方法进行了对比。分析结果表明，EVO-CE 和 CE 方法能够准确地估计出符合生产实际和过程先验知识的时延参数值，而同类对比方法不能做到这一点。同时，EVO-CE 的计算时间要远小于 CE 的计算时间，因而更能够满足生产的实际需求。

## 6.51 治金工程

高纯金属材料是具有很高纯度的特殊材料，具有高电导率和稳定性、良好的光学性能等物理特性，是制造各种精密科学仪器和高科技产品的必备材料。制备高纯金属需要精密的工艺来保证高纯度，但传统工艺方法普遍存在制备纯度低的问题。真空蒸馏法则可以绿色高效地提纯金属，但其工艺参数需要手动调节，依赖于人的经验。田庆华等 [238] 提出了一种真空蒸馏制备高纯金属的优化方法，利用 CE 等机器学习技术筛选出能够保证高纯度和低杂质的工艺参数集合，建立以纯度和杂质含量为目标变量的预测模型，再基于此模型利用参数寻优方法得到最佳工艺参数，用于高纯金属制备。他利用该方法进行了真空蒸馏制备高纯金属硒和碲的工艺参数优化实验，基于 CE 等特征选择方法发现蒸馏温度、保温时间、冷凝温度和真空中度对制备纯度具有重要性，保温时间、蒸馏温度、升温速度和冷凝温度对杂质含量具有重要性。经过不断的迭代循环实验，该方法所得工艺参数能够获得良好的制备效果，可以根据不同产品需求对工艺参数进行自动控制优化。

透气性指数是一个反映高炉冶炼过程中炉况的重要参数，衡量了炉子接受风量的能力。对高炉透气性指数进行预测，可以更好地控制冶炼过程，避免炉况的异常情况发生。Lin 等 [239] 提出了一个基于小波去噪、非线性 Transformer 模型的透气性指数预测方法，充分地考虑了高炉冶炼过程数据的多尺度、非线性和数据量大的特点。其中，该方法采用了皮尔逊相关系数和 CE 来选择影响透气性指数的关键变量作为模型输入。实验结果表明，该方法能够给出高预测准确度和快速的推理速度，为控制透气性指数提供了理论支撑，对提高冶炼过程的稳定性和自动化程度具有现实意义。

## 6.52 化学工程

故障诊断对化学过程的安全、高效运行至关重要，数据驱动的故障诊断方法是实际生产运行中的主要方法之一。为了构建诊断模型，构建合理的正常和故障状态的过程表示是问题的关键环

节。Yin 等 [240] 提出了一种基于 CE 的灰度相关空间的故障诊断方法，通过变量之间的 CE 相关性矩阵来刻画过程的正常和故障状态，再将矩阵作为卷积神经网络的输入来构建故障分类模型。他们将方法应用于田纳西伊斯曼 (Tennessee Eastman) 过程的故障诊断数据，结果表明该方法取得了 95% 以上的诊断准确率，验证了方法的有效性。主元分析法 (PCA) 是一种常用的多变量过程检测方法，原理是基于最大方差准则从一组过程变量构建过程检测统计量，但其仅适用于线性的情况。Wei 和 Wang [241, 242] 提出了一种基于 CE 的非线性 PCA 方法 (CEPCA)，从具有非线性特征的 CE 矩阵得到过程检测统计量。他们将方法应用于田纳西伊斯曼过程数据，并与 PCA 方法进行了对比，结果表明，CEPCA 方法获得了更好的故障检测率结果。Pan 等 [243, 244] 提出了一个基于关联故障因果图构建的故障传播和根因分析方法，称为 DTMTE，通过基于 TE 的时延估计来构建传感器变量之间的因果关系图，用于识别故障的根源和传播路径，其中基于 CE 的 TE 被用于分析因果关系。他们分别在田纳西伊斯曼过程和一个实际的碳酸二甲酯化工生产过程上验证了该方法，结果表明 DTMTE 方法能够准确地找到故障原因和传播路径，性能好于传统对比方法。

理解化工过程变量之间的因果关系对于过程控制十分重要，有助于更好的过程监测和故障诊断。利用因果发现方法构建化工过程因果关系图，可以对故障进行根因分析，是故障诊断的重要方法之一。Bi 等 [245] 提出了一种基于深度学习进行因果发现的 CGTST 方法，并与基于 CE 的 TE 等多种方法进行了对比。实验结果表明，在一个 5 变量的连续搅拌槽式反应器数据上，TE 方法获得的反应图结果非常接近于真实情况；在田纳西伊斯曼过程数据上，TE 方法也取得了接近于真实情况的估计结果，体现出了较强的实用性。

软测量技术是化工过程建模的重要方法之一，指通过易测量的过程变量来估计推断难以直接测量的过程变量。然而，受实际生产过程中设备故障、环境干扰和信号传输等多种因素的影响，过程变量数据往往包含大量的缺失值，因此需要进行缺失值补全。生成对抗补全网络 (Generative Adversarial Imputation Nets: GAIN) 是一种以生成对抗网络算法框架为基础的数据补全方法，但当缺失值数量较大时，算法的性能难以满足实际需求。武昊 [246] 提出了一种改进设计的 GAIN 算法框架，称为信息增强 GAIN (IEGAIN)，其中 CE 被用于计算权重矩阵以作为新算法中生成器的输入。他分别在 UCI 的 Spam 和 Letter 数据集、公开的火电厂数据集和脱丁烷塔过程数据集和实际的聚丙烯生产过程数据上，将 IEGAIN 与 GAIN 等其它经典算法进行了对比，结果表明 IEGAIN 能够以最低的误差补全数据缺失值。

生物发酵过程是一种绿色、节能的化工生产方式，已经成为许多化工行业（如食品加工、制药等）的首选，具有显著的经济效益和广阔的市场前景。补料分批发酵是一种典型的间歇生产方式，其微生物的代谢过程对工艺参数控制的精度要求较高，因此需要采取智能优化调控等技术手段，这就要求研究发酵过程关键参数的实时测量方法。软测量技术具有经济性、易于实现和维护的特点，克服了复杂发酵过程难以准确机理建模的难题，是发酵过程参数监测的有效解决方案，而基于数据建立参数预测模型是软测量技术的关键核心问题。针对间歇发酵生产过程的特点，时和畅 [247] 提出了一种基于时序差分神经网络的发酵过程软测量建模方法，利用基于 CE 的 MI 来进行模型输入变量选择，再利用带有差分算子的时序神经网络建立发酵质量参数预测的软测量模型。他将该方法应用于模拟工业规模青霉素发酵过程的 IndPensim 数据集，建立青霉素浓度预测模型，获得了较高的预测精度和鲁棒性，验证了该方法的有效性。

## 6.53 医学工程

脑机接口 (Brain-Computer Interface: BCI) 是通过分析处理大脑信号来产生控制指令的系统。基于运动成像的 BCI 可以帮助人通过大脑来产生运动指令，具有很多重要用途。脑磁图 (MEG) 具有高信噪比和时空分辨率的优点，在 BCI 领域具有应用潜力，但如何提高基于 MEG 的 BCI 系统的准确率是一个难题。Tang 等 [248, 249] 提出了一种基于 MEG 的运动成像 BCI 系统方法，其中利用基于 CE 的 TE 算法来选择 BCI 系统的信道集合。他们基于公开的 MEG 数据集测试了该方法，发现该方法能够在保证预测准确率的同时，大幅减少选择的信道数量，得到的预测性能优于同类基于随机森林的信道选择方法，为基于 MEG 的 BCI 的实际应用提供了技术保证。

## 6.54 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 [250] 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 [251, 252] 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 [253] 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。

涡扇发动机是喷气式飞机最常用的发动机，具有高效、可靠和节能的特点，是现代航空业的关键设备之一。涡扇发动机结构复杂，且长期在极端环境下运行，导致其容易出现磨损和老化，因而监测其健康状态，进而开展故障预测和维修保养，对于保障航空安全、提高涡扇发动机的可靠性和使用寿命至关重要。因而，如何评估发动机的健康状态是一个基础性的关键问题。贾如侠 [254] 提出了一种涡扇发动机的健康指标，采用证据推理方法融合发动机传感器监测数据度量发动机健康状态，其中 CE 被用于推理过程中计算发动机传感器变量的可靠度。他将方法应用于 NASA 格林中心提供的引擎性能退化模拟数据集，并与两种传统方法进行了对比，结果表明新方法对发动机健康状态的评估效果更好，这得益于方法融合了基于 CE 度量的传感器变量间非线性相关性信息。他进一步利用得到的一维复合健康指标建立了发动机故障预测模型和剩余

寿命预测模型，都获得了较对比方法更精确的预测效果。孙秀慧 [255] 提出了一种基于 CE 特征选择的发动机剩余寿命预测方法，首先利用 CE 选择与发动机健康因子具有非线性相关关系的传感器变量，再利用这些变量重构健康因子进而建立发动机失效的指数退化模型，然后利用相似性距离从发动机历史数据库中选择与指数退化模型预测相似度高的一组发动机，最后将这些发动机的健康因子中位数作为剩余寿命预测值。她基于 C-MAPSS 数据验证了该方法，结果表明，该方法基于 CE 的非线性度量能力，从 21 个传感器中选择了 8 个用于构建指数退化模型，在 50%、70%、90% 运行周期下，该方法的预测误差与传统方法相比分别降低了 39.25%、41.69%、50.53%，显著提高了剩余寿命预测精度，同时降低了模型复杂度从而提高了计算效率。她还提出了基于 Copula 相似性的剩余寿命预测方法，首先利用 CE 选择成对或多个相似传感组合，进而构建 Copula 函数或藤 Copula 结构，再在此基础上计算这些组合之间的 Copula 相似度量矩阵，然后基于这些相似矩阵选择历史数据库中相似的一组发动机，最后将这些发动机的剩余寿命平均值作为预测值。基于 C-MAPSS 数据集的实验结果表明，与传统方法相比，该方法在 50%、70%、90% 三个运行周期上，预测误差均有显著提高。她提出的这些方法具有较大的工程应用价值，有望应用于军用和民用航空发动机的健康管理。

航班延误是影响国际民航业正常有效运行的主要问题之一，不仅给旅客造成出行不便，也给航空业带来巨大经济损失。航空系统是一个有机的整体，运行中存在航班资源的上下游共享，带来了系统耦合，导致上游航班的到港延误会向下游传播，因此航班延误管控首先需要对这种延误因果关系进行分析。吴格等 [256] 提出利用一种基于 CE 的 TE 估计器来分析机场的航班延误时间序列之间的因果关系强度的方法，使民航信息系统具有了分析两个航班之间是否具有延误因果关系的能力，从而能够深入理解和利用航空系统节点间航班延误的内在关系。

机场协同决策 (A-CDM) 系统是国际民用航空组织支持的标准运营框架，用于通过航空系统各个单元之间实时共享数据来优化航班支持决策过程，是增强机场运营效率、提高可预测性和准点率的核心工具。机场和航线支持部门在实际中是基于预计上轮挡时刻 (EIBT) 调配资源，EIBT 包括了航班滑入时间和实际着陆时间 (ALDT) 两部分。当前的 A-CDM 系统主要关注 ALDT，对于 EIBT 考虑的不多。像北京首都机场和上海浦东机场等对于滑入时间的估计还主要基于空管员的经验，缺乏基于实时数据的精准预测，或者只是在航班着陆后才给出 EIBT，影响了服务的及时性和效率。Tang 等 [257] 提出了一种两阶段的航班到达时间预测方法，将飞机进入终点机动区后的时间分为空中飞行和地面滑入两个阶段，构建了基于机器学习的到达时间预测模型。他们基于浦东机场 2022 年 10 月 A-CDM 系统的数据，提取了 5 类共 18 个特征（包括飞行器和航班特征、机场地面运行特征、机场终点机动区特征、到达/出发流特征和天气特征等），再利用 CE 分别选择了与空中飞行时间、地面滑入时间，以及两阶段总时间相关联的特征，最后训练 LightGBM 模型进行预测。实验结果表明，CE 方法选择出了合理的到达时间相关特征，其中与两个阶段最相关的特征分别是飞行距离和滑入距离，与飞行时间相关的特征还包括飞行高度、速度和角度，与滑入时间相关的特征是运行热点数目。这些发现与已有的研究结论一致。预测实验表明，基于 CE 选择特征的两阶段到达时间预测模型的性能要好于采用全部特征的模型，3 分钟以内预测准确度约为 70%，5 分钟以内预测准确度约为 90%，可以为实际机场运行的灵活选择提供支持。实验也表明，CE 方法选择出了符合浦东机场运行模式的特征，表明了该方法针对不同特点机场的泛化能力，从而保证所构建模型的预测性能。

## 6.55 兵器工程

武器装备效能评估是指对某一武器的技术指标和作战性能进行全面、系统、科学的分析和评价。由于武器装备系统及其运用的复杂性，评估需要考虑多方面因素，因此就需要一套综合的指标体系来完成评估。效能指标体系往往包含大量不同类型的指标，从而造成指标之间具有相关性，导致指标体系维数大，需要对其进行约简，以利于后续的评估流程。传统的约简方法一般采用相关系数等数学工具，但其线性假设在实际问题通常得不到满足。陈爱真等 [258] 提出了一种指标体系约简方法，利用 CE 度量指标之间相关性，通过比较每个指标与其它指标之间的平均 CE 来约简指标。他利用评估对象的仿真数据实施验证了该方法，证明了该方法具有可处理指标间非线性相关性关系的优点，较传统方法更为科学和准确。

## 6.56 车辆工程

现代汽车的电子设备系统由车载网络连接集成，提高了乘坐的舒适性、安全性和多功能特性。但随着智能车辆技术的发展，车内设备也成为了黑客攻击的对象，对车辆安全构成了威胁。CAN 总线是一种智能车辆内连接控制各个车辆电子组件的数据通信协议，已在汽车领域成为事实上的主流标准，但由于缺乏加密、认证等机制，其在网络攻击面前非常脆弱。因此，研究 CAN 总线的入侵检测技术成为了提高其安全性的主要技术手段之一。Gao 等 [259] 提出了一种轻量级神经网络设计方法，用于检测 CAN 总线入侵事件，其首先分析异常 CAN 数据帧的属性集合，再利用 CE 选择出众多属性中与入侵攻击有关的少数属性，再利用这些属性构建一种 CanNet 神经网络检测器以检测入侵。他们利用现代汽车索纳塔 YF 的 CAN 总线数据验证了 CanNet 方法，结果表明该方法与同类方法相比具有高检测率、高实时性和低内存占用的优点。

高速动车组是我国高铁运行的主要交通运载工具，随着其速度的不断提高，在线预警异常运行状态变得越来越重要。滚动轴承是动车组的关键机构部件，具有运行转速高和载荷大的特点，必须保证其健康可靠，故障监测预警技术是保障其安全运行的关键技术之一。基于振动信号的机器学习健康监测是传统的旋转机构健康评估方法，但该方法在电动车组的应用面临诸多难题：1) 直接采集的振动时变信号具有高噪声扰动和非平稳性的特点；2) 动车组运行在多个不同工况，无法直接采用单个离线训练的模型进行监测；3) 实际故障情况样本数据少，给模型训练造成困难；4) 传统的基于云计算和深度学习的监测预警技术框架，由于计算耗时长且时延大等原因，在端侧计算资源有限的条件下无法满足高效性和实时性的技术要求。Xu 等 [260] 提出了一个自适应实时滚动轴承故障诊断技术框架，利用 CE、迁移学习、宽度学习算法和端云协同等技术成功解决了上述难题。其中，他们利用 CE 方法发现了在故障情况下振动信号与力信号之间的相关性，使得在端侧通过非直接信号进行故障诊断成为可能。他们基于正常高铁运行场景下的真实数据验证了该方法，发现基于力信号的诊断模型可以准确地检测到损坏故障，该技术框架与传统方法相比准确性和实时性都大幅提高，准确率达到 99.98% 以上，可以有效地对动车组滚动轴承进行实时故障诊断。

## 6.57 控制工程

航站楼是航空系统交通枢纽，负责旅客在天地之间的传输转换，在提升城市可持续、促进环境友好航空方面发挥重要作用。HVAC 系统消耗了航站楼 60% 以上的能源，在改善航站楼能效方面具有提升空间。航班计划导致航站楼客流具有某种动态模式特征，如何考虑客流动态特征改善 HVAC 系统控制的能源效率是一个重要的问题，而问题的关键在于室内温度预测以满足系统控制期望达成的舒适度和能效双重目标。Li 等 [261] 提出了一种以旅客为中心的模型预测控制 (MPC) 框架，以进行高能效的航站楼室内温度控制。该框架包括一个室内温度目标动态设定器、一个室内温度预测器和一个控制指令优化器。其中，MPC 的预测器由一个因果关系约束的神经网络模型构成，模型的输入由基于 CE 的 TE 方法筛选的与室内温度有因果关系的因素构成。他们以广州白云机场 1 号航站楼为验证对象，选择了 4 个国际和国内候机区域作为目标验证该方法框架。基于现场实际采集数据的 TE 分析实验发现，历史室温、邻近区域室温、室外温度、太阳辐射强度、旅客密度、供应气流温度和风扇频率是室内温度的因变量，分析同时辨识了因果关系的时延参数，而 GC 等传统线性因果分析方法并未给出同样合理的分析结果。在此基础上，基于 TE 分析结果得到的神经网络预测模型给出了比对比方法构建的模型更准确的预测性能。基于 CE 的 TE 方法在此展现了比传统线性因果分析方法更强大的非线性因果关系分析能力，赋予了预测模型以可解释性。在航站楼目标区域的现场试验表明，基于此 MPC 框架的能耗比对比方法低 17% 到 46%。

## 6.58 电子工程

半导体芯片的集成度的不断提高，对微电子封装的要求也越来越高。微电子封装起着隔绝外部环境、散发内部热量的功能，对集成电路的稳定运行具有至关重要的保护作用。这就要求封装材料具有良好稳定性、高强度，同时还要满足其他物理性质。刘勃 [262] 以 Cu 基材料为主体，建立 CuNi 二元合金体系，利用第一性原理与机器学习相结合的方法，基于团簇相关函数特征，预测分别与材料强度和稳定性相关的构型能和杨氏模量。作者利用 CE 分析了预测模型的合理性，通过计算特征之间的相关性，以及特征与构型能和杨氏模量之间的相关性，发现模型特征与杨氏模量之间的相关性更高，同时构型能与杨氏模量之间的相关度较低，增进了模型的可解释性，有助于设计更合理的材料性质预测模型。

## 6.59 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 [263, 264] 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。

第 6 代（6G）通信网络技术的研发需要面对的主要挑战之一就是要达到更高的数据传输率，以满足更极致的体验、3D 视觉、工业智能等场景需求。传统的通信理论没有考虑传输信息中的语义信息，而 6G 技术可以利用基于 AI 的语义通信来达到更高的网络传输性能。傅宇舟等 [265] 提出了一种面向 6G 网络的基于语义通信的端到端服务框架，将语义通信与 AI 的语义分析能力相融合，利用基于 Transformer 的编解码器来压缩语义信息。其中，语义编码器的损失函数由基于欧式距离的语义损失函数和基于 CE 的信息量损失函数组成。他们利用图像数据验证了该服务框架，使用 ImageNet-1K 数据集训练框架，再使用 VOC2012 数据集进行仿真验证。结果表明，与传统通信方案相比，该服务框架在目标检测和图像语义重建上均取得了最优性能，且取得了与全语义特征传输方案相近的性能，有望成为 6G 网络的技术内容。

## 6.60 高性能计算

提高能源效率是高性能计算研究的一个重要目标。通过配置程序的最优能效设置，如处理器频率等，可以降低程序执行时的能耗。但决定最优配置是一个费时的过程，程序一旦修改就需要重新配置。利用机器学习方法通过性能事件来自动决定最优配置是一个新的研究方向，但需要确定哪些事件是能效相关的以决定最优配置。Gocht-Zech [266] 提出利用特征选择的方法来选择能效相关事件，他选择了 6 种特征选择方法，并基于 CE 理论给出了相应的估计方法。实际数据实验表明该基于 copula 的方法能够鉴别出能效相关的性能事件，从而提高程序执行时的能效，在增加 7% 运行时的成本下节省了 24% 的能源消耗。

## 6.61 信息安全

对抗性攻击和防御是信息安全领域的热点问题，是指攻击者利用对系统和算法的特性的了解发动的攻击以及相应的防御手段。深度神经网络是机器学习领域的一类重要算法，应用领域十分广泛，研究其攻击和防御算法对该类人工智能系统的安全具有重要意义。Liu 等 [267] 提出了一个基于 CE 的 MI 估计算法，称为  $CE^2$ ，并利用此算法提出了一个神经网络对抗训练算法。该算法充分利用了基于 CE 的 MI 估计对对抗攻击的可靠性，设计网络训练算法以引导神经网络预测模型最小化对抗样本的攻击。作者首先通过仿真实验证明了  $CE^2$  相对于传统 MI 估计算法的性能优势，然后在 CIFAR-10 和 CIFAR-100 数据集上验证了基于  $CE^2$  的神经网络防御算法在典型深度神经网络对抗性攻击的防御中相对于其他同类经典防御算法的优越性。

物联网（IoT）作为一种万物互联的信息技术，已经广泛地应用到如工业生产、医疗保健、智能家居等社会生活的各个领域，成为一类重要的信息基础设施。IoT 安全是保障设施正常运行的前提，因此，IoT 入侵检测作为一个重要的安全问题被广泛研究。然而，由于数据量所限，现有的入侵检测方法无法很好地跨域应用。王倩等 [268,269] 提出了一种新的跨域 IoT 入侵检测方法，利用增量聚类算法从入侵数据集生成图结构，再利用图神经网络提出域特征，然后利用基于 CE 定义的距离将特征进行跨域对齐，最后利用条件域对抗神经网络进行入侵检测。他们在 4 个公开

的 IoT 入侵检测数据集上验证了方法的有效性。其中，基于 CE 的数据跨域对齐方法能够有效地减小提取特征的域间差异，从而提高检测分类性能。

云计算的广泛采用使得云计算环境安全成为了技术应用者面临的挑战之一。特别是，恶意软件通过利用云计算基础设施的弱点进行攻击，会造成数据泄漏、非法系统访问和身份窃取等不良后果。利用机器学习算法对网络流量数据进行分析进而检测恶意软件攻击是当前广泛采用的云安全防护技术手段之一，可以有效提高安全防护的自动化程度，达到快速、准确和自适应响应的目的。为了改善传统机器学习恶意软件检测算法的不足，Baawi 等 [270] 提出了一种新的恶意软件检测分类器算法，以提高检测性能。他们在公开的恶意软件检测数据集 Meraz'18 的基础上，对所提出的新算法进行了验证并与其他机器学习算法进行了对比，对比实验中设计了有特征选择和无特征选择两种分类检测情况，其中特征选择采用基于 CE 的方法。实验结果表明，该新算法检测性能优于同类传统机器学习算法，在采用 CE 进行特征选择后，该新算法仅用 20 个选择特征就取得了使用全部 53 个特征几乎同等的检测性能。

## 6.62 测绘遥感

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 [271] 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。

变形是指变形体在外部因素作用下产生的位置、尺寸及形态特征的变化，其往往是一个极其缓慢的物理过程，如山体或建筑物的形变。变形达到一定程度就会引发安全风险，如不有效预防就可能在地震、滑坡等自然灾害中造成严重危害。变形监测就是对变形体进行的监控测量，通过利用监测数据构建形变模型，以对变形趋势进行预测。工程变形监测是工程测量领域的重要问题之一，需要保证监测精度和可靠性，对大型工程的施工运营安全具有重要意义。常见的变形监测分析方法一般只是针对单个监测点的建模和预测，但变形体内部监测点间不是孤立的，而是具有内在的相关性，因而可以利用这种相关性提高单点监测的预测精度。张旭辉等 [272, 273] 提出了一种考虑邻近变形点间相关性的机器学习变形分析预测方法，其中利用 CE 来选择与预测点有相关性的邻近点。他基于太湖隧道施工中测量机器人在第八仓围堰上采集的 2020 年 12 月 10 日至 2021 年 10 月 8 日间 12 个监测点的数据验证了该方法，将 CE 与两种传统的相关性度量进行了对比，结果表明基于 CE 选取的邻近监测点的模型得到的预测精度更高，证明了 CE 更适合变形测量的非线性时间序列预测问题。该方法对实际工程围堰预警等长期变形预测问题具有良好的应用价值。

## 6.63 海洋工程

人类对海洋空间的探索是海洋工程建设、海洋资源开发和管理以及海洋军事行动等活动的基础，海洋底质信息探测是诸多活动的前提，因而是海洋测绘学研究的重要问题之一。多波束声呐系统是海洋测绘领域的主要调查设备之一，可以用来通过声学探测获取海底的底质信息并对其进行分类。赵廷 [274] 提出了一整套多波束声呐海底底质分类技术，在多波束反向散射图像的基础上提取一组空间、频率和尺度特征，再利用 CE 等相关性工具去除其中的冗余特征，最后利用筛选后的特征构建底质分类模型。他在比利时 Oostende Harbor 数据集上对提出的特征选择和模型构建方法进行了实验验证，结果表明，利用 CE 等工具可以发现特征之间的非线性相关关系，在此基础上去除冗余特征后，模型的分类性能得到了显著提升。

## 6.64 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 [275] 和 ArbitrageLab [276] 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 [277] 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)。廖轶楠 [278] 研究了投资标的筛选的问题，他基于净资产收益率、净利润三年复合增长率和市盈率三项指标从 A 股 4000 多家上市公司中初步筛选了 10 家 A 股上市公司，再利用 CE 等工具对标的股票的价格数据进行了统计分析，以判断投资组合的抗风险能力。

股票市场的投资者总是希望投资发展良好的上市公司，因此甄别一只股票的好坏对投资者十分重要。ST 股票制度是在我国 A 股市场实施的股票风险警示机制，有助于投资者选择投资组合并规避风险。股票分类是股票分析领域的一类重要问题，对金融市场投资者具有参考价值。朱仲儿 [279, 280] 提出了一种基于机器学习方法的 ST 股票分类方法，采用 Boruta 算法和 CE 方法进行特征选择，再利用 6 种回归模型进行预测，利用 Optuna 框架对模型的超参数寻优。他选取了 tushare 数据库中上交所和深交所的 2076 只股票（含 351 只 ST 股票）自 2016 年以来的数据，含有 139 个股票特征变量，最终利用 Boruta 和 CE 方法筛选了 7 个可解释的变量。模型预测结果表明该方法在筛选特征和 XGBoost 模型组合上获得了最好的预测精度。

“一带一路”是由我国倡导的针对丝绸之路沿线国家的国际合作倡议，对我国和相关国家的经济社会发展具有重要的推动作用。“一带一路”指数 (Belt&Road Index: BRI) 是与此发展倡议相关的行业和公司的金融市场指标，反映了倡议涉及国家和地区的发展趋势和变化，对政府和

投资者的决策具有重要参考价值，因此指数的预测分析是本领域的一个重要问题。徐泽晖 [281] 提出了一个结合了 GAS 模型、CE 和 lightGBM 的 BRI 收益率预测方法，其中 CE 被用于对预测模型的输入特征进行选择。他利用 2020-2023 年的 58 支 BRI 成分股的数据验证了该方法，结果表明，与同类对比方法相比，GAS-CE-LGBM 方法在所有四个预测性能评估指标上表现最优。特别是，利用 CE 进行特征选择显著地提高了模型的预测性能，表明 CE 能够捕捉到问题中变量之间的非线性动态关系。

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 [282, 283] 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 (Information Excess) 的概念。作者将建模方法应用于 1990-2008 年标普 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。

R 藤 Copula 是一种灵活的构建多元 copula 分布的工具，确定藤的结构是建立此类模型的关键步骤。Alanazi [284] 基于 CE 和 MI、CMI 之间的关系，提出了一种 R 藤 copula 的构建方法，基于 MI 建立最小生成树，再计算前一子树每对边上的 CMI，根据 CMI 建立新的子树并决定藤 copula 的层级结构。他将该 R 藤 copula 构建方法应用于股票间相关结构的建模问题，基于德国 DAX 指数 15 种主要股票数据（2005 年 1 月至 2009 年 8 月）构建了资产间关系结构的 R 藤 copula 模型，与传统方法相比，该方法建立的 copula 相关结构模型能够更好地拟合数据。王念鸽 [285] 基于 CE 与 MI、CMI 之间的关系提出了一个类似的藤 copula 结构选择算法。作者利用该算法分析了中证五大行业指数之间的相关结构，利用 2019 年 3 月 1 日至 2022 年 3 月 1 日之间的数据，构建了基于 Kendall 相关系数的藤 copula 结构和基于 MI 的藤 copula 结构，结果表明，从拟合优度指标看，后者的结果优于前者的结果；从可解释性角度看，后者的结果刻画的五大行业资产之间的依赖关系更合理。

金融危机的发生使金融系统的系统性风险问题受到各国监管部门的关注。我国股票市场放开管制加深了经济金融的一体化程度，造成了各个行业之间的耦合，从而加大了系统性风险的程度，因此需要对跨行业的风险溢出效应加以研究，以期进行防范和化解。熵作为量化不确定性的数学工具，十分适合度量金融风险组合。熊靖宇 [286] 采用 CE 等工具对 2005 年 1 月 5 日至 2020 年 7 月 3 日我国股票市场 11 个行业的日对数收益率数据进行了分析，研究行业个体风险和跨行业风险溢出特征的动态演变过程，特别针对 2008 年金融危机、2013 年钱荒和 2015 年股灾三个时期的风险特征进行研究。研究发现，行业联合 CE 动态变化滞后于累加独立熵发生，说明了行业间联动导致了系统性风险增强；2008 年金融危机的市场内部传染性更强，破坏程度更大；近期 11 个行业内部关联水平较强。丁永辉 [287] 利用 CE 对 116 家上市金融机构 2006 年 10 月 27 日至 2023 年 12 月 31 日之间的日收益率数据进行了分析，以研究金融系统风险联动的特征。研究发现，系统性冲击会导致金融系统的风险联动程度急剧上升，银行业的部门内风险联动要强于其他部门，部门间风险联动的程度要高于部门内，多元金融部门会传播放大系统性冲击造成的风险。

金融脆弱性是由金融部门自身高负债经营带来的内在不稳定性。金融脆弱性度量工具可以使国家及时地对危机进行响应和干预，因此得到了大量的研究。日益成熟的网络分析理论为从金融

网络的角度度量金融脆弱性提供了方法工具，但传统的网络构建方法只是基于线性关系度量工具，如皮尔逊相关系数等，不能够反映金融系统中的非线性关系特性。Chen 等 [288] 提出了一种利用 CE 改进的网络曲率（Network Curvature）金融脆弱性度量方法，该方法先利用 CE 构建金融网络，再计算网络的四种离散 Ricci 曲率作为市场脆弱性度量。他们将该度量方法应用于 2006 年 4 月至 2022 年 4 月间沪深 300 指数的股票数据，分析金融危机前后的市场脆弱性。结果表明，该度量方法比基于皮尔逊相关的方法更清晰地描述了金融危机后市场的脆弱性，且具有传统风险度量同样的风险度量能力。

近年的研究表明，金融资产组合在极端事件的金融动荡中会受到冲击，存在巨大的金融风险。家庭和金融组合管理者迫切需要了解金融冲击和极端事件对投资的影响。传统的风险度量工具难以检测到这种尾部相关性，而基于 Copula 模型的方法越来越显示出在此问题上的优越性。Ardakani 和 Ajina [289] 提出利用基于 CE 的 MI 来度量极端事件区域的尾部风险，来明确多样化策略对于应对尾部风险的意义。他们在 2022 年消费者金融调查的数据上应用了此风险度量工具，发现某些资产组合表现出了强相关，因而加大了尾部风险。这一发现对家庭理财应对尾部风险具有重要参考价值。

信用风险是金融银行业面对的主要基本风险之一，保障金融安全需要有效地管理信用风险。信用评分卡模型是一种对客户进行信用风险评价的模型方法，是管控金融风险的决策工具。该类模型根据客户的信用历史数据为其划分信用等级，来决定其金融权限。传统的建立信用评分卡模型方法依靠专家经验，效率低且生成的模型不够完善。孔祥永等 [290] 提出一种基于 CE 的自动化信用风险模型构建方法，能够显著提高建模效率，可以同时保证模型具有高预测性能和可解释性。作者将该方法在真实信用卡数据上与专家建模进行了对比，实验结果表明方法大大缩短了建模时间，且能够得到媲美专家模型的预测性能和可解释的客户信用特征。

P2P (Peer-to-Peer) 借贷是一种通过互联网进行集资和放贷的金融模式，该类金融模式的信用风险主要是由借贷人未能履行还款义务造成的，对集资债权人的资金安全构成了巨大风险。因此，如何准确地评估借贷方的信用风险是一个重要问题，通过借贷数据构建个人信用风险模型是一个主要的解决方法。彭翊庭 [291] 提出利用 CE 度量风险变量和个人数据高维特征之间的非线性相关性，用以选择个人信用风险预测模型的输入特征。他利用美国 P2P 借贷平台 Lending Club 的贷款数据展开实证研究，对比了 CE 和皮尔逊相关系数两种常用特征选择方法，发现 CE 选择的非线性特征在 XGBoost 模型上获得了更好的预测结果。

绿色信贷是金融机构提供的一种以生态环境保护为目标的上市公司融资工具，研究绿色信贷风险评估能够提升金融机构对工具使用中的风险把控。王钊颖 [292] 提出了一种基于 CE 和机器学习模型相结合的绿色信贷风险评估方法。她以 2021 年 A 股上市公司作为实际案例，选取了公司状况、创新发展能力和绿色评价三个方面的 67 个指标，基于 CE 选择了其中的 18 个指标构成了风险评估指标体系，再利用四种机器学习方法构建评估模型。实验结果表明，所得模型的准确率达到 95.01%，为绿色信贷风险提供了可靠的评估工具。

准确地预测金融产品价格可以帮助投资者管理风险并进行投资决策，因而建立相关预测模型是研究者关心的重要问题之一。由于金融产品之间存在内在的市场逻辑，它们的价格也会产生相应的因果联动效应。因此，可以利用这种价格间的因果关系建立比传统方法更准确的价格预测模型。Zhang 等 [293] 提出了一种基于价格间因果关系的迁移学习框架，利用基于 CE 的 TE 方法

计算不同金融产品价格之间的因果关系，以选择因变量价格用于预测果变量价格，再在选择的基础上，提出了用于训练深度学习模型的学习算法以得到预测模型。他们将算法分别应用于国际主要的财经指数、能源期货价格和农产品期货价格 2010 年至 2021 年的每日价格数据上，结果发现，利用基于 CE 的 TE 方法发现了同类价格间的因果关系，在此基础上，利用该迁移学习框架得到的模型在三类价格数据上均给出了较同类对比算法更好的预测结果。

流行病疫情对人群健康构成严重威胁，促使社会和个体采取应对措施。这些疫情应对进而会产生巨大的经济社会影响，特别是对金融市场的影响。研究疫情对金融市场的影响是一个重要课题，对市场利益主体具有现实意义。Gurgul 和 Syrek [294] 利用 CE 方法研究了波兰股票市场指数在 2019 新冠疫情期间的相关性特征，特别研究了 2020 年 3 月 13 日波兰疫情发生当天 WIG 指数和其 14 个板块指数之间的相关性，发现宣布疫情后这种相关性明显增强。作者还利用同样的方法研究了四个国家（法、德、英、美）的股市，利用 CE 计算了各国股市板块收盘价和股市指数之间的相关性，发现了疫情后这种相关性也明显增强 [295]。这一发现与 2008 年金融危机得到的经验是一致的。他们还发现 CE 方法得到的结论与经验相符，而传统皮尔逊相关得到的结论则不符合过去的经验，如后者低估了德国股市疫情后的相关性。作者认为这是因为 CE 可以度量金融市场变量之间的非线性相关关系而不做任何假设，从而验证了 CE 的优越性。

中国保险行业经历了 40 余年的快速发展后，正在经历数字化进程，保险科技的应用正在深刻影响行业企业，面向行业痛点的解决方案正在加快落地。因此，研究我国保险企业对科技的应用程度，以及这些应用对企业经营造成哪些影响是一个重要课题。栗嵩林 [296] 从理论层面提出了保险科技对保险公司绩效的四个方面（包括发展能力、盈利能力、营运能力和风险管理能力等）产生影响的理论假设，并以 2018 至 2020 年全国 114 家保险公司的相关数据为基础进行了实证研究。其中，他利用回归基准模型分析了保险科技水平对企业的影响，再利用 CE 计算保险科技与模型变量之间的非线性关联强度加以验证，二者均表明保险科技水平对保险公司业务费用率、临时分保比率、总资产收益率、综合投资收益率等产生了显著影响，实证研究结论与理论分析相符。他基于该理论和实证研究，对保险公司和行业监管提出了具有重要价值的建议。

保险费的制定是保险公司提供保险服务的重要环节，关系着保险业务经营管理的效益和安全。保险公司一般根据客户的基本社会信息进行风险评估，来确定其保险费的额度。传统的保险费制定根据经验来完成，缺乏科学性和合理性。基于数据分析模型进行保险费预测是一种新的业务模式。然而，客户的信息项目往往较多，建立预测模型需要先对这些信息进行筛选，才能获得准确的模型预测能力。Uddin [297] 提出利用基于 CE 的变量选择方法建立保险费预测模型。他将此方法应用于公开的汽车保险业务数据，从 16 个客户信息项中选择了其中的 5 项建立预测模型，同时将方法与同类建模方法进行了对比。实验结果表明，CE 方法建立的模型给出了最优的预测性能。

近年来，机器学习方法在金融市场预测领域的研究兴趣正在增加，主要得益于其非线性分析能力和较高的资产预测准确性，但是在加密货币市场预测的实际部署却很少，因为传统的机器学习方法不能够在动态市场环境和极端市场条件下选择出与目标金融资产有关联的预测变量，根本原因在于方法背后不合理的有效市场假设。CE 方法能够在无分布假设的条件下分析非线性、非高斯性和非对称性的相关性，为解决问题提供了工具。基于自适应市场假设，Mahmudovic [298] 提出了一种在真实市场动态性条件下进行有效且可解释预测的方法，方法采用基于 CE 的方法

选择时变且尾部相关的指标变量，同时采用 Copula 散度混合误差函数来指导预测模型学习。他基于四种加密货币（比特币、以太币、瑞波币和狗币）真实多年历史数据验证了方法，结果发现，基于 CE 的方法在提高了预测准确度的同时还增加了模型的可解释性，而 Copula 散度误差函数也减小了累积误差。方法的成功说明了自适应市场假设在加密货币市场上的合理性。



## 附录 A 软件实现

本书所述的 CE 估计算法、TE 估计算法、正态性检验和双样本检验的统计量估计算法和变点检测算法已在 R 和 Python 语言的 `copent` 算法包中实现 [314]，分别在 CRAN 和 PyPI 上共享：

- CRAN <https://cran.r-project.org/package=copent>；
- PyPI <https://pypi.org/project/copent/>。

相关源码见作者的 GitHub: <https://github.com/majianthu/>。

另，基于 CE 理论的 CE/MI/TE 估计等算法的第三方软件实现包括：

- R 语言的 `cylcop` 包 [97, 595]；
- Python 语言的 `MLFinLab` 包 [275]、`ArbitrageLab` 包 [276]、`gcmi` 包 [111, 596]、`pytorch-mighty` 包 [597]、`HOI` 包 [598]、`THOI` 包 [305]、`Frites` 包 [599]、`Tensorpac` 包 [600]、`driada` 包 [601]、`CopulaGP` 包 [602, 603]、`Polars-ds` 包 [604] 和 `effconnp` 包 [109, 605]；
- Julia 语言的 `CopEnt.jl` 包 [606]、`CausalityTools.jl` 包 [607] 和 `Copulas.jl` 包 [608]；
- Matlab 语言的 `gcmi` 包 [111, 596] 和 `FieldTrip` 包 [609]；以及
- C++ 语言的 `NPStat` 包<sup>1</sup> [610]。

---

<sup>1</sup>该包的 CE 非参数估计算法实现被分解为经验 Copula 密度估计和熵估计两个单独的函数。



## 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] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987.
- [3] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
- [4] Roger B Nelsen. *An Introduction to Copulas*. Springer New York, NY, 2006.
- [5] Harry Joe. *Dependence Modeling with Copulas*. CRC press, 2014.
- [6] Abe Sklar. Fonctions de repartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.
- [7] Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- [8] Thomas M Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2005.
- [9] 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.
- [10] 马健. *Copula 结构学习*. 博士学位论文, 清华大学, 2009.
- [11] Jian Ma and Zengqi Sun. Dependence structure estimation via copula. *arXiv preprint arXiv:0804.4451*, 2008.
- [12] Jian Ma. Discovering association with copula entropy. *arXiv preprint arXiv:1907.12268*, 2019.
- [13] 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.

- [14] Jian Ma. Estimating transfer entropy via copula entropy. *arXiv preprint arXiv:1910.04375*, 2019.
- [15] Jian Ma. Causal domain adaptation with copula entropy based conditional independence test. *arXiv preprint arXiv:2202.13482*, 2022.
- [16] Jian Ma. Multivariate normality test with copula entropy. *arXiv preprint arXiv:2206.05956*, 2022.
- [17] Jian Ma. Identifying Time Lag in Dynamical Systems with Copula Entropy based Transfer Entropy. *arXiv preprint arXiv:2301.06037*, 2023.
- [18] Jian Ma. System identification with copula entropy. *arXiv preprint arXiv:2304.12922*, 2023.
- [19] Jian Ma. Two-sample test with copula entropy. *arXiv preprint arXiv:2307.07247*, 2023.
- [20] Jian Ma. Change point detection with copula entropy based two-sample test. *arXiv preprint arXiv:2403.07892*, 2024.
- [21] Jian Ma. Testing symmetry with copula entropy based two-sample test. *ChinaXiv preprint ChinaXiv:202505.00167*, 2025.
- [22] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.
- [23] Jian Ma. Evaluating independence and conditional independence measures. *arXiv preprint arXiv:2205.07253*, 2022.
- [24] Seyedeh Azadeh Fallah Mortezanejad, Gholamreza Mohtashami Borzadaran, and Bahram sadeghpour Gildeh. Joint dependence distribution of data set using optimizing Tsallis copula entropy. *Physica A: Statistical Mechanics and its Applications*, 533:121897, 2019.
- [25] S.M. Sunoj and N. Unnikrishnan Nair. Survival copula entropy and dependence in bivariate distributions. *REVSTAT-Statistical Journal*, 23(1):101–115, 2025.
- [26] Shital Saha and Suchandan Kayal. Copula-based extropy measures, properties and dependence in bivariate distributions. *Communications in Statistics - Theory and Methods*, 2025. See also arXiv preprint arXiv:2311.08061.
- [27] Mohd. Arshad, Swaroop Georgy Zachariah, and Ashok Kumar Pathak. Multivariate information measures: A copula-based approach. *arXiv preprint arXiv:2408.02028*, 2024.
- [28] Swaroop Georgy Zachariah, Mohd. Arshad, and Ashok Kumar Pathak. Dependence and uncertainty: Information measures using Tsallis entropy. *arXiv preprint arXiv:2502.12779*, 2025.

- [29] Shital Saha and Suchandan Kayal. Multivariate Rényi inaccuracy measures based on copulas: properties and application. *arXiv preprint arXiv:2502.17215*, 2025.
- [30] Morteza Mohammadi and Mahdi Emadi. Nonparametric tests of independence using copula-based Renyi and Tsallis divergence measures. *Statistics, Optimization & Information Computing*, 11(4):949–962, 2023.
- [31] Morteza Mohammadi, Mahdi Emadi, and Mohammad Amini. Bivariate dependency analysis using Jeffrey and Hellinger divergence measures based on copula density estimation by improved probit transformation. *Journal of Statistical Sciences*, 15(1):233–254, 2021.
- [32] Aman Pandey and Chanchal Kundu. Copula-based modeling of fractional inaccuracy: A unified framework. *arXiv preprint arXiv:2506.19748*, 2025.
- [33] José M. Amigó, Sámuél G. Balogh, and Sergio Hernández. A brief review of generalized entropies. *Entropy*, 20(11):813, 2018.
- [34] Nader Ebrahimi, Ehsan S. Soofi, and Refik Soyer. Information measures in perspective. *International Statistical Review*, 78(3):383–412, 2010.
- [35] Jian Ma. On thermodynamic interpretation of copula entropy. *arXiv preprint arXiv:2111.14042*, 2021.
- [36] Jian Ma. Photometric redshifts with copula entropy. *arXiv preprint arXiv:2310.16633*, 2023.
- [37] Aderonke Akerele, Babatunde Rabiu, Samuel Ogunjo, Daniel Okoh, Anton Kascheyev, Bruno Nava, Olawale Bolaji, Ibiyinka Fuwape, Elijah Oyeyemi, Busola Olugbon, Jacob Akinpelu, and Olumide Ajani. Complexity and nonlinear dependence of ionospheric electron content and Doppler frequency shifts in propagating HF radio signals within equatorial regions. *Atmosphere*, 15(6):654, 2024.
- [38] Jian Ma. Facies classification with copula entropy. *arXiv preprint arXiv:2501.14351*, 2025.
- [39] Yufeng Shan, Jiangfeng Wei, and Beilei Zan. Improving estimates of land–atmosphere coupling through a novel framework of land aridity classification. *Geophysical Research Letters*, 51(2):e2023GL106598, 2024.
- [40] 单昱峰. 不同土壤湿度与蒸散发耦合状态下的陆气耦合特征分析及其对干旱传播的影响. 硕士学位论文, 南京信息工程大学, 2025.
- [41] Laura Mack, Marvin Kähnert, and Norbert Pirk. Probabilistic modelling of atmosphere-surface coupling with a copula bayesian network. *arXiv preprint arXiv:2509.11975*, 2025.

- [42] Yuhao Chen, Fan Zhang, Ke Chen, Lufeng Zhu, and Shouqi Yuan. Research on performance prediction of a small sample centrifugal pump based on a pre-processing approach for imbalanced regression and key hyperparameters optimization. *Physics of Fluids*, 37(6):067148, 2025.
- [43] 李勇, 张靖昊, 刘苑喆, 高昂, 陈昕宇, 和 张泽华. 基于一维卷积残差网络的微热管几何结构及制造工艺参数预测. 中国科学: 技术科学, 55(2):281–294, 2025.
- [44] 张靖昊. 基于机器学习的微热管结构及工艺参数设计系统开发. 硕士学位论文, 华南理工大学, 2024.
- [45] Yunpeng Ma, Meng Li, Mengqian Wang, and Chenheng Xu. Improve thermal efficiency and reduce NO<sub>x</sub> emission of circulating fluidized bed boiler based on multi-objective optimization framework. *Thermal Science and Engineering Progress*, page 103753, 2025.
- [46] Michel A. Cuendet, Harel Weinstein, and Michael V. LeVine. The allosteric landscape: Quantifying thermodynamic couplings in biomolecular systems. *Journal of Chemical Theory and Computation*, 12(12):5758–5767, 2016.
- [47] 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.
- [48] 田杰. 基于机器学习的耐热型含能材料设计方法研究. 硕士学位论文, 西南科技大学, 2023.
- [49] Jian Liu, Jie Tian, Rui Liu, Yuechuan Tang, Chunming Yang, Junhong Zhou, and Chaoyang Zhang. Screening heat-resistant energetic molecules via deep learning and high-throughput computation. *Chemical Engineering Journal*, page 160218, 2025.
- [50] Song Zhang, Y.C. Lin, Dao-Guang He, Yu-Qiang Jiang, Hui-Jie Zhang, Ning-Fu Zeng, Gui-Cheng Wu, and Majid Naseri. Correlation between plastic deformation mechanism and texture evolution of a near  $\beta$ -Ti alloy deformed in  $\beta$  region. *Intermetallics*, 170:108333, 2024.
- [51] Chengchen Jin, Kai Xiong, Zhongqian Lv, Congtao Luo, Hui Fang, Jiankang Zhang, Aimin Zhang, Shunmeng Zhang, Yong Mao, and Yingwu Wang. High-throughput calculation and machine learning-assisted prediction of the mechanical properties of refractory multi-principal element alloys. *Advanced Theory and Simulations*, page e00784, 2025.
- [52] 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.

- [53] 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.
- [54] 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.
- [55] Xiao Li, Liping Zhang, Sidong Zeng, Zhenyu Tang, Lina Liu, Qin Zhang, Zhengyang Tang, and Xiaojun Hua. Predicting monthly runoff of the upper Yangtze river based on multiple machine learning models. *Sustainability*, 14(18):11149, 2022.
- [56] Ran Mo, Bin Xu, Ping-An Zhong, Yuanheng Dong, Han Wang, Hao Yue, Jian Zhu, Huili Wang, Guoqing Wang, and Jianyun Zhang. Long-term probabilistic streamflow forecast model with “inputs–structure–parameters” hierarchical optimization framework. *Journal of Hydrology*, 622:129736, 2023.
- [57] 陈佳雷, 彭甜, 葛宜达, 王熠炜, 张楚, 孙娜, 王政, 李茜, 钱诗婕, 和 李燕妮. 一种基于多要素注意力时空图卷积网络的径流预报方法, 2023. CN117151285A.
- [58] 汪胤. 基于 Copula 熵和神经网络的降雨径流预报及村镇防洪系统研发. 硕士学位论文, 同济大学, 2018.
- [59] 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.
- [60] C.Y. Huang and Y.P. Zhang. Prediction based on copula entropy and general regression neural network. *Applied Ecology and Environmental Research*, 17(6):14415–14424, 2019.
- [61] 黄春艳. 黄河流域的干旱驱动及评估预测研究. 博士学位论文, 西安理工大学, 2021.
- [62] 牛犇. 黄河流域气象-农业-水文干旱的时空传递特征. 硕士学位论文, 西北农林科技大学, 2023.
- [63] Lingling Ni, Dong Wang, Jianfeng Wu, Yuankun Wang, Yuwei Tao, Jianyun Zhang, Jifu Liu, and Fei Xie. Vine copula selection using mutual information for hydrological dependence modeling. *Environmental Research*, 186:109604, 2020.
- [64] P. Kanthavel, C.K. Saxena, and R.K. Singh. Integrated Drought Index based on Vine Copula Modelling. *International Journal of Climatology*, 42(16):9510–9529, 2022.
- [65] 徐袁. 基于 GAMLSS 框架下非平稳气象干旱和水文干旱指数的构建及干旱传播分析. 硕士学位论文, 东北农业大学, 2025.

- [66] Ruofei Xing, Zefeng Chen, Jie Hao, Wenbin Liu, Qin Ju, Dawei Zhang, Shiqin Xu, and Huimin Wang. Temporal variation in river ice phenology of the Heilongjiang river in response to climate change. *Journal of Hydrology: Regional Studies*, 54:101868, 2024.
- [67] 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.
- [68] 徐鹏程. *CEM 模型和 KCEM 模型在水文站网优化中的应用*. 博士学位论文, 南京大学, 2018.
- [69] 王栋, 徐鹏程, 王远坤, and 吴吉春. 一种基于 Copula 熵的水文站网优化模型的优化方法, 2019. CN106897530B.
- [70] 王栋, 吴吉春, 吴剑锋, and 王远坤. 不确定性分析方法在水文站网优化中的研究与应用. 中国水利水电出版社, 2022.
- [71] 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.
- [72] 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.
- [73] 徐鹏程, 仇建春, 李帆, 刘赛艳, and 蒋新跃. 基于高维 Copula 熵和克里金的站网优化方法, 2022. CN114595556A.
- [74] 徐鹏程, 李帆, 张昌盛, and 仇建春. 基于 C-Vine Copula 熵多目标优化模型的水文气象站网优化研究. 中国农村水利水电, 2:16–21, 2022.
- [75] 杨惜岁. 多目标准则下流域水文站网的优化与评价. 硕士学位论文, 武汉理工大学, 2019.
- [76] Lu Chen and Shenglian Guo. *Copulas and its Application in Hydrology and Water Resources*. Springer Singapore, 2018.
- [77] Xu Wang and Yong-Ming Shen. A framework of dependence modeling and evaluation system for compound flood events. *Water Resources Research*, 59(8):e2023WR034718, 2023.
- [78] Longxia Qian, Yong Zhao, Jianhong Yang, Hanlin Li, Hongrui Wang, and ChengZu Bai. A new estimation method for copula parameters for multivariate hydrological frequency analysis with small sample sizes. *Water Resources Management*, 36(4):1141–1157, 2022.

- [79] Yibo Wang, Liu Yakun, Ze Cao, and Di Zhang. Prediction of contraction channel scour depth: based on interpretability analysis and PCA-enhanced SVR. *Journal of Hydroinformatics*, 26(12):3287–3305, 2024.
- [80] 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, 33(3):442–451, 2022.
- [81] 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.
- [82] Victor Costa Porto. *Advancements in streamflow modeling and forecasting in Brazil*. PhD thesis, Universidade Federal do Ceará, 2023.
- [83] 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.
- [84] 蒋佩东. 基于 Bayesian 模型的长江流域年径流对变化环境响应的空间效应. 硕士学位论文, 武汉大学, 2023.
- [85] Zengchao Hao and Vijay P. Singh. Integrating entropy and copula theories for hydrologic modeling and analysis. *Entropy*, 17(4):2253–2280, 2015.
- [86] Francesca Condino. *La divergenza di Jensen-Shannon nell'algoritmo di clustering dinamico per dati descritti da distribuzioni multivariate*. PhD thesis, Università degli Studi di Napoli Federico II, 2009.
- [87] Jujie Wang, Wenjie Xu, Yue Zhang, and Jian Dong. A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization. *Chaos, Solitons & Fractals*, 158:112098, 2022.
- [88] Xiaoxuan Wu, Chen Zhang, Jun Zhu, and Xin Zhang. Research on PM2.5 concentration prediction based on the CE-AGA-LSTM model. *Applied Sciences*, 12(14):7009, 2022.
- [89] Jieyin Chen. Short-term prediction of PM2.5 concentration based on self-attention mechanism improved temporal convolution network. In *2023 International Seminar on Computer Science and Engineering Technology (SCSET)*, pages 528–534, 2023.
- [90] Qiao Guo, Haoyu Zhang, Yuhao Zhang, and Xuchu Jiang. Prediction of PM2.5 concentration based on the CEEMDAN-RLMD-BiLSTM-LEC model. *PeerJ*, 11:e15931, 2023.
- [91] 陈燕璇, 刘合香, and 倪增华. 基于 Copula 熵因子选取的 PSO-ELM 台风灾情预测模型. 气象研究与应用, 40(2):7–11, 2019.

- [92] 吴京鹏. 基于图嵌入表示的节点无特征网络链路预测研究. 硕士学位论文, 西北师范大学, 2022.
- [93] 金秀章, 乔鹏, and 史德金. 基于 VMD-Bayes-Lasso 算法带误差补偿的火电厂 NO<sub>x</sub> 浓度软测量. 华北电力大学学报 (自然科学版), 52(235):117–124+142, 2023.
- [94] 杨媛. 电站锅炉 NO<sub>x</sub> 排放浓度两阶段智能优化研究. 硕士学位论文, 东北电力大学, 2024.
- [95] 乔鹏. 基于 CatBoost-Bayes 带误差补偿的脱硫出口 SO<sub>2</sub> 浓度软测量研究. 硕士学位论文, 华北电力大学, 2024.
- [96] Jingkai Xue, Chengzhi Xing, Qihua Li, Shanshan Wang, Qihou Hu, Yizhi Zhu, Ting Liu, Chengxin Zhang, and Cheng Liu. Long-term spatiotemporal variations of ammonia in the Yangtze River Delta region of China and its driving factors. *Journal of Environmental Sciences*, 150:202–217, 2025.
- [97] 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.
- [98] Wanyu Li, Gangsheng Wang, Zirui Mu, Shanshan Qi, Shuhao Zhou, and Daifeng Xiang. Microbially-mediated soil carbon-nitrogen dynamics in response to future soil moisture change. *Earth's Future*, 13(3):e2024EF005521, 2025.
- [99] Yueying Li, Li Peng, Sainan Li, Yuemin Yue, and Kelin Wang. Integrating transfer entropy and network analysis to explore social-ecological resilience evolution: a case study in South China Karst. *Journal of Cleaner Production*, page 145926, 2025.
- [100] Francisco Escolano, Edwin R. Hancock, Miguel A. Lozano, and Manuel Curado. The mutual information between graphs. *Pattern Recognition Letters*, 87:12–19, 2017.
- [101] Soumik Purkayastha and Peter X.K. Song. Asymmetric predictability in causal discovery: an information theoretic approach. *arXiv preprint arXiv:2210.14455*, 2022.
- [102] 王子祥. 不同树种液流分析及预测方法研究. 硕士学位论文, 浙江农林大学, 2023.
- [103] Yane Li, Yuhang Jiang, Lijun Guo, Weibo Wang, Jiahao Wu, Xiang Weng, and Hailin Feng. CE-CNN-BiGRU-SA: A novel generalization sap flow prediction model for trees at different ages and species with historical environmental factors. *SSRN:5460406*, 2025.
- [104] Ziya Zhang, Yi Li, Chen Xinguo, Yanzi Wang, Ben Niu, De Li Liu, Jianqiang He, Bakhtiyor Pulatov, Ishtiaq Hassan, and Qingtao Meng. Impact of climate change and planting date shifts on growth and yields of double cropping rice in southeastern China in future. *Agricultural Systems*, 205:103581, 2023.

- [105] 张子雅. 气候变化对中国水稻物候和产量的影响及播期优化. 硕士学位论文, 西北农林科技大学, 2023.
- [106] 张春磊, 李颜娥, 丁煜, and 罗煦钦. 基于深度学习技术的水稻环境因素产量预测. 电子技术应用, 50(4):81–86, 2024.
- [107] Paolo Victor Redondo, Raphaël Huser, and Hernando Ombao. Measuring information transfer between nodes in a brain network through spectral transfer entropy. *The Annals of Applied Statistics*, 19(3):2386–2411, 2025.
- [108] Zhaojun Li, Yanyu Xing, Xinyan Wang, Yunlu Cai, Xiaoxia Zhou, and Xi Zhang. Estimating global phase synchronization by quantifying multivariate mutual information and detecting network structure. *Neural Networks*, 183:106984, 2025.
- [109] Wojciech Ciezborka, Joan Falcó-Roget, Cemal Koba, and Alessandro Crimi. End-to-End stroke imaging analysis using effective connectivity and interpretable artificial intelligence. *IEEE Access*, 13:10227–10239, 2025.
- [110] 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.
- [111] 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.
- [112] 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.
- [113] Etienne Combrisson, Michele Allegra, Ruggero Basanisi, Robin A.A. Ince, Bruno Giordano, Julien Bastin, and Andrea Brovelli. Group-level inference of information-based measures for the analyses of cognitive brain networks from neurophysiological data. *NeuroImage*, 258:119347, 2022.
- [114] 汪方毅, 唐杰庆, 刘倩, 余成新, 李博, and 丁帆. 基于静息态 fMRI 区分健康老年人认知水平的 MVPA 方法研究. 磁共振成像, 14(6):18–25, 2023.
- [115] Pieter De Clercq, Jonas Vanthornhout, Maaike Vandermosten, and Tom Francart. Beyond linear neural envelope tracking: a mutual information approach. *Journal of Neural Engineering*, 20(2):026007, 2023.

- [116] NA Pospelov, VP Sotskov, VV Plusnin, OS Rogozhnikova, KA Toropova, OI Ivashkina, and KV Anokhin. Searching for cognitive specializations of neurons using mutual information framework. *Genes & Cells*, 18(4):878–881, 2023.
- [117] Laouen Belloli, Pedro Mediano, Rodrigo Cofré, Diego Fernandez Slezak, and Rubén Herzog. THOI: An efficient and accessible library for computing higher-order interactions enhanced by batch-processing. *arXiv preprint arXiv:2501.03381*, 2025.
- [118] Maximilian Walden. Application of Bagged Copula-GP: Confirming Neural Dependency on Pupil Dilation. *Transactions on Machine Learning Research*, 2024.
- [119] Maximilian Walden. Bagged Copula-GPFA: A framework for estimating dynamic neuronal dependency relationships. Technical report, University of Edinburgh, 2024.
- [120] Dinu Johannes Kaufmann. *Semi-parametric Gaussian Copula Models for Machine Learning*. PhD thesis, University of Basel, 2017.
- [121] 吴亚婷, 余青山, 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.
- [122] 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.
- [123] David O’ Reilly and Ioannis Delis. A network information theoretic framework to characterise muscle synergies in space and time. *Journal of Neural Engineering*, 19(1):016031, 2022.
- [124] David O’Reilly. *Dissecting muscle synergies in the task space*. PhD thesis, University of Leeds, 2024.
- [125] Shaojun Zhu, Jinhui Zhao, Yating Wu, and Qingshan She. Intermuscular coupling network analysis of upper limbs based on R-vine copula transfer entropy. *Mathematical Biosciences and Engineering*, 19(9):9437–9456, 2022.
- [126] 金国美, 余青山, 马玉良, 张建海, and 孙明旭. 基于小波包-copula 互信息的肌间耦合特性. 传感技术学报, 35(10):1348–1353, 2022.
- [127] Johannes Leugering and Gordon Pipa. A unifying framework of synaptic and intrinsic plasticity in neural populations. *Neural Computation*, 30(4):945–986, 2018.
- [128] Johannes Leugering. *Neural mechanisms of information processing and transmission*. PhD thesis, Universität Osnabrück, 2021.

- [129] 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.
- [130] Vlad-Bogdan Coroian, Pedro Mediano, and Tolga Birdal. Scaling up synergy estimators. Technical report, Imperial College London, 2024.
- [131] Liesa Ravidts. Revealing temporal interactions around the heartbeat-evoked potential modulated by emotional perception. Master's thesis, Ghent University, 2019.
- [132] Agata Charzyńska and Anna Gambin. Improvement of the k-NN entropy estimator with applications in systems biology. *Entropy*, 18(1):13, 2015.
- [133] Farzaneh Farhangmehr, Daniel M. Tartakovsky, Parastou Sadatmousavi, Mano R. Mauurya, 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.
- [134] Aleksander Wieczorek and Volker Roth. Causal compression. *arXiv preprint arXiv:1611.00261*, 2016.
- [135] Qiang Wu and Dongxi Li. CRIA: An interactive gene selection algorithm for cancer prediction based on copy number variations. *Frontiers in Plant Science*, 13:839044, 2022.
- [136] 吴强. 基于 Copula 熵的交互式基因选择算法及其在癌症预测中的应用. 硕士学位论文, 太原理工大学, 2022.
- [137] Shilong Shang, Dongxi Li, Xiaoran Yan, and Yun Dang. CEFS+: An efficient and interactive feature selection approach based on copula entropy for high-dimensional genetic data. *SSRN:5014010*, 2024.
- [138] 商世龙. 基于 Copula 熵特征选择和顶点式动态 GAT 的基因预测研究. 硕士学位论文, 太原理工大学, 2024.
- [139] Xiaoran Yan, Shilong Shang, Dongxi Li, and Yun Dang. An efficient and interactive feature selection approach based on copula entropy for high-dimensional genetic data. *Scientific Reports*, 15(1):30100, 2025.
- [140] Hongyu Pan, Shanxiong Chen, and Hailing Xiong. A high-dimensional feature selection method based on modified Gray Wolf Optimization. *Applied Soft Computing*, 135:110031, 2023.
- [141] 钟琦. 基于信息度量的基因表达数据特征选择方法研究. 硕士学位论文, 曲阜师范大学, 2024.

- [142] 竺政彤. 基于单细胞测序数据构建基因调控网络的方法研究. 硕士学位论文, 内蒙古农业大学, 2023.
- [143] Jing Li. *Advanced Computational Framework for Dissecting Gene Regulatory Dynamics in Complex Diseases: Hepatocellular Carcinoma*. PhD thesis, University of Liverpool, 2024.
- [144] Antonio Lacalmita. *Integrazione di approcci di intelligenza artificiale e reti complesse per l'analisi dei dati genomici e la scoperta di biomarcatori in malattie complesse*. PhD thesis, Università degli studi di Bari, 2025.
- [145] Radko Mesiar and Ayyub Sheikhi. Nonlinear random forest classification, a copula-based approach. *Applied Sciences*, 11(15):15, 2021.
- [146] Jian Ma. Copula entropy based variable selection for survival analysis. *arXiv preprint arXiv:2209.01561*, 2022.
- [147] 付金露. 基于特征选择的乳腺癌患者预后模型研究. 硕士学位论文, 江西财经大学, 2023.
- [148] Yu Luo, Guangcan Xu, Hongyu Li, Tianju Ma, Zi Ye, and Zhaojun Li. Research on establishing corneal edema after phacoemulsification prediction model based on variable selection with copula entropy. *Journal of Clinical Medicine*, 12(4):1290, 2023.
- [149] 罗昱. 主动控制液流系统在白内障超声乳化手术中的应用研究. 硕士学位论文, 中国人民解放军医学院, 2023.
- [150] 潘红宇. 基于影像组学与深度学习的脑肿瘤图像分类研究. 硕士学位论文, 西南大学, 2023.
- [151] 汤宇飞. 基于脉搏波的糖尿病和高血压诊断算法研究. 硕士学位论文, 中国矿业大学, 2023.
- [152] 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).
- [153] 李润泽, 姚尧, 冯珂珂, 杨硕, 李佳丽, 程铁峰, 尹绍雅, and 徐桂芝. 重复经颅磁刺激改善帕金森病运动症状的脑功能网络分析. *生物化学与生物物理进展*, 50(1):126–134, 2023.
- [154] Jian Ma. Predicting TUG score from gait characteristics based on video analysis and machine learning. In *Proceedings of 2023 Chinese Intelligent Automation Conference*, pages 1–12, 2023. See also bioRxiv 963686 (2020).
- [155] Jian Ma. Associations between finger tapping, gait and fall risk with application to fall risk assessment. *arXiv preprint arXiv:2006.16648*, 2020.
- [156] 张婷婷, 王楠, 周天彤, 王苏弘, and 邹凌. 基于 Couple 熵的抑郁症相干性反馈指标提取. *电子测量技术*, 45(9):160–167, 2022.

- [157] 张婷婷. 基于脑电的抑郁症识别及虚拟现实康复训练研究. 硕士学位论文, 常州大学, 2022.
- [158] Di Han, Yuhu Shi, Lei Wang, Yueyang Li, and Weiming Zeng. The multi-frequency decomposition entropy learning for nonlinear fMRI data analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 33:68–80, 2025.
- [159] 王中华. 人类身高的表观基因组关联分析和基于 DNA 甲基化构建成人身高推断的深度神经网络模型. 博士学位论文, 河北医科大学, 2024.
- [160] 高浩田, 李东喜, 陈泽华, and 赵芊. 基于交互式多特征融合算法的药物靶标预测. 太原理工大学学报, 55(4):751–758, 2024.
- [161] Qingsong Shan and Qianning Liu. Binary trees for dependence structure. *IEEE Access*, 8:150989–150998, 2020.
- [162] 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.
- [163] Haonan Zhang, Jiapeng Dai, and Yousaf Ali Khan. Poverty improvement policies and household income: Evidence from China. *Heliyon*, 9(11):E21442, 2023.
- [164] Leonie Bossemeyer. Machine learning for causal discovery with applications in economics. Master's thesis, Ludwig-Maximilians-Universität München, 2021.
- [165] 韦颖璐. 基于 pair-copula 熵的相关性度量. 硕士学位论文, 苏州大学, 2021.
- [166] Muye Han and Jinsheng Zhou. Multi-scale characteristics of investor sentiment transmission based on wavelet, transfer entropy and network analysis. *Entropy*, 24(12):1786, 2022.
- [167] Omid M. Ardakani. Information content of inflation expectations: A copula-based model. *Studies in Nonlinear Dynamics & Econometrics*, 29(1):71–93, 2024.
- [168] Wuyue An, Lin Wang, and Dongfeng Zhang. Comprehensive commodity price forecasting framework using text mining methods. *Journal of Forecasting*, 42(7):1865–1888, 2023.
- [169] Yuliana Apaza Flores. Estudo de séries temporais dos preços da gasolina, etanol e açúcar no estado de São Paulo através da transfer entropy. Dissertação de Mestrado, Universidade Estadual Paulista, 2025.
- [170] Yu-Xin Tian and Chuan Zhang. An end-to-end deep learning model for solving data-driven news vendor problem with accessibility to textual review data. *International Journal of Production Economics*, 265:109016, 2023.

- [171] 王琳君. 中国企业海外并购的影响因素和绩效评价研究. 博士学位论文, 中国科学院大学, 2022.
- [172] 王修臻子, 魏云捷, and 王琳君. 宏观经济冲击对中国跨境并购的影响分析. 管理评论, 37(7):67–76, 2025.
- [173] 柳琼. 基于 Copula 和 MI 理论的相关性度量及其应用研究. 硕士学位论文, 三峡大学, 2018.
- [174] 陈作海, 钱恒, and 高永超. 一种基于知识图谱的城市热线派单方法及系统, 2023. CN115860436A.
- [175] Bowen Zhang, Jinping Lin, Man Luo, Changxian Zeng, Jiajia Feng, Meiqi Zhou, and Fuying Deng. Changes in public sentiment under the background of major emergencies – taking the Shanghai epidemic as an example. *International Journal of Environmental Research and Public Health*, 19(19):12594, 2022.
- [176] Mario Wieser. *Learning Invariant Representations for Deep Latent Variable Models*. PhD thesis, University of Basel, 2020.
- [177] Stuart William Card. Towards an information theoretic framework for evolutionary learning. Master's thesis, Syracuse University, 2011.
- [178] 张可, 刘施彤, 郑植, 贾宇明, and 黄乐天. 一种基于动态贝叶斯网络的目标意图识别方法, 2022. CN114997306A.
- [179] 许海云, 王超, 陈亮, 徐硕, 杨冠灿, and 朱礼军. 颠覆性技术的科学-技术-产业互动模式识别与分析. 情报学报, 42(7):816–831, 2023.
- [180] 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.
- [181] 朱正林 and 张冕. 基于 AO 优化 VMD-CE-BiGRU 的光伏发电功率预测. 国外电子测量技术, 41(10):56–61, 2022.
- [182] 杨秀, 闫钟宇, 孙改平, 熊雪君, and 冯煜尧. 基于多类型天气识别的光伏功率日前预测. 现代电力, 2025.
- [183] 王士涛 and 邹晨鑫. 一种光伏组件状态的预测方法及系统、计算机程序产品, 2024. CN118134033A.
- [184] 崔双双 and 孙单勋. 分工况下风电机组各变量相关性研究. 综合智慧能源, 44(12):49–55, 2022.

- [185] Qin Yan, Zhiying Lu, Hong Liu, Xingtang He, Xihai Zhang, and Jianlin Guo. Short-term prediction of integrated energy load aggregation using a bi-directional simple recurrent unit network with feature-temporal attention mechanism ensemble learning model. *Applied Energy*, 355:122159, 2024.
- [186] 阚超. 基于深度学习的综合能源系统多元负荷短期预测研究. 硕士学位论文, 贵州大学, 2023.
- [187] 胡程林. 基于深度多任务学习的多节点电力负荷预测研究. 硕士学位论文, 湘潭大学, 2022.
- [188] 吴迪, 颜俣, 鲁刚, 夏鹏, 闫晓卿, 孙广增, 杨帆, 贺永龙, 马国福, 秦满鑫, and 许辉. 基于混合机器学习的配电网负荷预测装置及方法, 2024. CN118585786A.
- [189] 王伊佳. 多源数据融合模型在时间序列预测中的应用研究. 硕士学位论文, 西南石油大学, 2024.
- [190] 董海艳, 赵炳文, 王运韬, 田宇, 傅彦博, 孟德群, and 张铁. 一种含源荷时序相似度约束的源储协同规划配置方法, 2022. CN110766314A.
- [191] Peili Liu, Song Han, Na Rong, and Junqiu Fan. Frequency Stability Prediction of Power Systems Using Vision Transformer and Copula Entropy. *Entropy*, 24(8):1165, 2022.
- [192] 刘沛力. 基于机器学习的低惯性电力系统频率稳定性预测方法研究. 硕士学位论文, 贵州大学, 2023.
- [193] 冯双, 杨浩, 雷家兴, 汤奕, 周吉, 钱俊良, and 郝珊珊. 一种电力系统宽频振荡影响因素和传播路径分析方法, 2022. CN114977222A.
- [194] 冯双, 杨浩, 崔昊, 汤奕, and 雷家兴. 基于 copula 传递熵的设备级和网络级宽频振荡传播路径分析及振荡源定位方法. *电工技术学报*, 39(16):4996–5010, 2023.
- [195] Ying Lu, Jutian Li, Zhen Zhang, Hao Yang, Rui Lin, Shuang Feng, Yongyan Chen, and Jiaxing Lei. An analysis method of influencing factors of wideband oscillations in fractional frequency transmission system for offshore wind power based on copula entropy. In *2024 6th International Conference on Power and Energy Technology (ICPET)*, pages 653–658, 2024.
- [196] Wentao Sun, Quanquan Wang, Sixuan Xu, Yi Ge, Feifei Zhao, Zhuyi Peng, and Wanchun Qi. A stability analysis and quantitative evaluation method for hybrid transmission systems. In *2023 4th International Conference on Advanced Electrical and Energy Systems (AEES)*, pages 412–417, 2023.
- [197] 王旭, 王之伟, 陈泉, 张文嘉, 孙文涛, 邹盛, 王荟荟, 宗炫君, 韩杏宁, 沈高峰, 张敏, 孙海森, and 王静怡. 一种交直流混联系统宽频振荡风险的识别方法及装置, 2024. CN117674116A.

- [198] Ting Yang, Yachuang Liu, Hao Li, Yanhou Chen, and Haibo Pen. Cooperative voltage control in distribution networks considering multiple uncertainties in communication. *Sustainable Energy, Grids and Networks*, 39:101459, 2024.
- [199] Wei Hu, Qiuting Guo, Wei Wang, Weiheng Wang, and Shuhong Song. Research on user loss contribution calculation of high-loss distribution area based on transfer entropy. In *2022 China International Conference on Electricity Distribution (CICED)*, pages 499–502, 2022.
- [200] 秦超 and 潘毓笙. 一种基于时空特征的配电网拓扑辨识方法, 2023. CN117154679A.
- [201] Xiaoping Xiong and Guohua Qing. A hybrid day-ahead electricity price forecasting framework based on time series. *Energy*, 264:126099, 2022.
- [202] Jiabei He and Lifeng Wu. Cross-conditions capacity estimation of lithium-ion battery with constrained adversarial domain adaptation. *Energy*, 277:127559, 2023.
- [203] 赵长胜. 动力电池健康状态的多维特征融合与迁移辨识方法研究. 硕士学位论文, 郑州轻工业大学, 2025.
- [204] Jian Ma. Root cause analysis on energy efficiency with transfer entropy flow. *arXiv preprint arXiv:2401.05664*, 2024.
- [205] Seyedeh Azadeh Fallah Mortezaejad, Ruochen Wang, Gholamreza Mohtashami Borzadaran, and Kim Phuc Tran. Non-parametric multivariate control chart using copula entropy. *Sankhya B*, 2025.
- [206] Marvin Lasserre, Régis Lebrun, and Pierre-Henri Wuillemin. Learning Continuous High-Dimensional Models using Mutual Information and Copula Bayesian Networks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 12139–12146. AAAI Press, 2021.
- [207] Marvin Lasserre. *Apprentissages dans les réseaux bayésiens à base de copules non-paramétriques*. PhD thesis, Sorbonne Université, 2022.
- [208] Zhiwei Li, Peng Wang, Jili Zhang, and Hua Guan. A model-free method for identifying time-delay characteristics of HVAC system based on multivariate transfer entropy. *Building and Environment*, 217:109072, 2022.
- [209] Chao Ma, Jingwei Chi, Dongxu Li, Fanchao Kong, Dechun Lu, and Weizhang Liao. Prediction on seismic performance levels of reinforced concrete beams by considering crack development. *Soil Dynamics and Earthquake Engineering*, 187:109006, 2024.
- [210] 迟经纬. 基于承载与防水功能的城市地下空间结构抗震韧性评价方法. 硕士学位论文, 北京建筑大学, 2024.

- [211] 常旭, 马财龙, 肖旭峰, and 鲁成凤. RC 柱受剪承载力神经网络预测模型及其可解释性. 新疆大学学报 (自然科学版中英文), 42(1):114–128, 2025.
- [212] Jin Liu, Changhai Zhai, Shunshun Pei, Zhuoru Song, and Bochang Zhou. Vine-Copula seismic functionality evaluation method of medical room systems based on shaking table tests. *Earthquake Engineering & Structural Dynamics*, 54(6):1499–1519, 2025.
- [213] 林超. 盾构轴线偏差预测与纠偏参数区间推荐. 硕士学位论文, 郑州大学, 2023.
- [214] 黄达. 基于模块链构建的大件货物多式联运方案研究. 博士学位论文, 北京交通大学, 2021.
- [215] 许罗豪, 刘金鑫, 张慧波, and 纪超. 基于熵与回归树的票价影响因素研究. 综合运输, 45(6):125–130, 2023.
- [216] 纪超. 基于旅客出行需求的高铁市场化定价研究. 博士学位论文, 北京交通大学, 2022.
- [217] 王升. 基于多源数据的城市轨道交通系统客流分析与预测. 硕士学位论文, 东南大学, 2022.
- [218] Jiahao Chang and Xiaoyu Song. A railway passenger flow prediction model based on improved Prophet. In *Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application*, pages 798–804, 2024.
- [219] 常家豪. 基于集成模型的铁路客流量预测研究. 硕士学位论文, 兰州交通大学, 2024.
- [220] 周泳江. 基于拓扑脑电特征选择与融合的驾驶员疲劳识别研究. 硕士学位论文, 西华大学, 2024.
- [221] 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.
- [222] Hongxia Cai and Zhiqiang Rong. Key quality feature identification and quality prediction in complex manufacturing processes. In *2023 15th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pages 229–232, 2023.
- [223] 王小巧. 复杂机械产品装配过程质量自适应控制方法及支持系统研究. 博士学位论文, 合肥工业大学, 2015.
- [224] Jie Dong, Keren Cao, and Kaixiang Peng. Hierarchical causal graph-based fault root cause diagnosis and propagation path identification for complex industrial process monitoring. *IEEE Transactions on Instrumentation and Measurement*, 72:1–11, 2023.
- [225] 刘鹏阳. 数据驱动的全流程分布式过程监控. 硕士学位论文, 北方工业大学, 2023.
- [226] 王晶, 刘鹏阳, 卢山, 周萌, and 陈晓露. 基于 CE-Louvain 分解和动态递归 SVDD 的分布式过程监测. 控制理论与应用, 2024.

- [227] Jie Hu, Min Wu, Weihua Cao, and Witold Pedrycz. Dynamic modeling framework based on automatic identification of operating conditions for sintering carbon consumption prediction. *IEEE Transactions on Industrial Electronics*, 71(3):3133–3141, 2023.
- [228] 李家宝, 鄢萍, 李蓬川, 翟鸿锦, 刘洋洲, 冯伟, and 易润忠. 基于特征融合和集成学习的涡轮盘模锻质量预测方法, 2024. CN118735067A.
- [229] 李家宝, 鄢萍, 李蓬川, 翟鸿锦, 刘洋洲, 冯伟, and 易润忠. 基于多目标优化问题建模的涡轮盘模锻工艺参数优化方法, 2024. CN118734701A.
- [230] 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.
- [231] 程毅. 基于深度学习的砂轮剩余使用寿命预测. 硕士学位论文, 江南大学, 2023.
- [232] Zong Meng, Shufan Ma, Wei Cao, Jimeng Li, Lixiao Cao, Fengjie Fan, and Xingzhao Wang. A remaining useful life prediction method of rolling bearings by RSA-BAFT combined with copula entropy feature selection. *Expert Systems with Applications*, 275:127100, 2025.
- [233] 耿妍竹. 基于多模型融合的风电机组故障预警研究. 硕士学位论文, 华北电力大学, 2024.
- [234] Zhifeng Luo and Haojiang Xi. Hybrid model based on copula mutual information and SSA-BP: Analysis of key factors and prediction of stable gas production. *Arabian Journal for Science and Engineering*, 50:4673–4685, 2025.
- [235] Zishang Yuan, Yong Wan, Lu Fan, Dong Sun, Yu Liu, Ligang Li, and Yongshou Dai. An analysis method of influencing factors of crude oil carbon footprint based on XGB-CE: a case study of an onshore oil production area in Shengli oilfield, China. *Clean Technologies and Environmental Policy*, 2025.
- [236] 高峰. 基于多源信息感知的煤-岩识别技术研究. 博士学位论文, 武汉大学, 2023.
- [237] 建中华 and 代伟. 基于 Copula 熵优化的煤炭重介过程多变量时延参数估计方法. In 2024 中国自动化大会, 2024.
- [238] 田庆华, 崔璇, 许志鹏, and 郭学益. 一种真空蒸馏制备高纯金属的优化方法及优化系统, 2024. CN117577229A.
- [239] Xu hui Lin, Xiang dong Xing, Yi ze Ren, Bao rong Wang, Zhi heng Yu, Ming Lv, and Zhong ze Du. Prediction model of permeability index for blast furnace based on WD-NL-transformer. *Ironmaking & Steelmaking*, 2024.
- [240] Min Yin, Jince Li, and Hongguang Li. A CNN approach based on correlation metrics to chemical process fault classifications with limited labeled data. *The Canadian Journal of Chemical Engineering*, 101(7):3982–3997, 2022.

- [241] Yingpeng Wei and Li Wang. Copula entropy-based PCA method and application in process monitoring. In *2022 4th International Conference on Intelligent Information Processing (IIP)*, pages 61–64, 2022.
- [242] 魏英鹏. 基于 copula 函数的过程监测方法研究. 硕士学位论文, 上海应用技术大学, 2023.
- [243] Shuangshuang Pan, Li Zhu, and Xirong Xu. Root cause and fault propagation analysis based on causal graph in chemical processes. In *2023 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)*, pages 1–6, 2023.
- [244] 潘爽爽. 面向化工过程的异常检测与故障溯源方法研究. 硕士学位论文, 大连理工大学, 2024.
- [245] Xiaotian Bi, Deyang Wu, Daoxiong Xie, Huawei Ye, and Jinsong Zhao. Large-scale chemical process causal discovery from big data with Transformer-based deep learning. *Process Safety and Environmental Protection*, 173:163–177, 2023.
- [246] 武昊. 基于深度学习的化工过程软测量建模方法研究. 博士学位论文, 北京化工大学, 2023.
- [247] 时和畅. 基于时序差分神经网络的发酵过程批次终点预测及优化研究. 硕士学位论文, 北京化工大学, 2025.
- [248] Chao Tang, Dongyao Jiang, and Badong Chen. MEG Channel Selection Using Copula Entropy-Based Transfer Entropy for Motor Imagery BCI. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4, 2024.
- [249] Chao Tang, Dongyao Jiang, Yi Guo, Liangjun Chen, and Badong Chen. Copula transfer entropy-based channel selection for MEG motor imagery brain computer interfaces. *Tsinghua Science and Technology*, 2025.
- [250] 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.
- [251] Hao Liu, Dechang Pi, Shuyuan Qiu, Xixuan Wang, and Chang Guo. Data-driven identification model for associated fault propagation path. *Measurement*, 188:110628, 2022.
- [252] 刘昊. 面向时序遥测数据的故障传播路径挖掘方法研究. 硕士学位论文, 南京航空航天大学, 2023.
- [253] 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.

- [254] 贾如侠. 涡扇发动机故障预测及剩余寿命分析方法研究. 硕士学位论文, 哈尔滨师范大学, 2023.
- [255] 孙秀慧. 基于特征相似性的航空发动机剩余寿命预测. 硕士学位论文, 鲁东大学, 2024.
- [256] 吴格, 陈旭, 傅之凤, 李忠虎, and 杨程屹. 一种因果关系分析方法及装置, 2020. CN110766314A.
- [257] Xiaowei Tang, Mengfan Ye, Jiaqi Wu, and Shengrun Zhang. Two stages of arrival aircraft: Influencing factors and prediction of integrated arrival time. *Aerospace*, 12(3):250, 2025.
- [258] 陈爱真, 罗汝斌, 董帅, 方娟, 刘朝阳, 白云飞, 董柏顺, 曹达, 王时雨, 李艳, 李梦阳, 侯婷婷, 王琭珉, and 周桃. 一种效能评估指标体系的约简方法和系统, 2024. CN117634946A.
- [259] Sheng Gao, Linchuan Zhang, Lei He, Xiaoyang Deng, Huilin Yin, and Hao Zhang. Attack detection for intelligent vehicles via CAN- bus: A lightweight image network approach. *IEEE Transactions on Vehicular Technology*, 72(12):16624–16636, 2023.
- [260] Jie Xu, Hong Tao, Xiaowen Zhang, Dengyu Xu, Xin Lu, Yiran Guo, and Yanhui Wang. An adaptive multi-signal framework for real-time fault diagnosis of rolling bearings. *IEEE Transactions on Instrumentation and Measurement*, 74:1–16, 2025.
- [261] Zhiwei Li, Zhuo Chen, Jili Zhang, and Song Mu. Passenger-centric model predictive control for indoor temperature to facilitate energy-efficient airport terminal. *Building and Environment*, 283:113344, 2025.
- [262] 刘勃. 基于机器学习的封装材料加速预测. 硕士学位论文, 哈尔滨理工大学, 2022.
- [263] 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.
- [264] 王旭, 金梁, 刘璐, 李明亮, and 黄开枝. 均匀散射环境中物理层安全密钥容量分析. *通信学报*, 37(9):75–81, 2016.
- [265] 傅宇舟, 程文驰, 陈小军, and 李赞. 面向 6G 网络的基于语义通信的端到端服务框架. *移动通信*, 47(6):35–40, 2023.
- [266] Andreas Gocht-Zech. *Ein Framework zur Optimierung der Energieeffizienz von HPC-Anwendungen auf der Basis von Machine-Learning-Methoden*. PhD thesis, Technische Universität Dresden, 2022.
- [267] Lin Liu, Cong Hu, and Xiao-Jun Wu.  $CE^2$ : A copula entropic mutual information estimator for enhancing adversarial robustness. In *Pattern Recognition and Computer Vision*, pages 163–174. Springer Singapore, 2024.

- [268] 王倩, 王学航, 刘韩, and 孟伟伦. 一种物联网入侵检测方法, 2024. CN118473739A.
- [269] Qian Wang, Xiang Liu, Yifan Cheng, Yongqiang Cheng, and Bing Zhang. A domain adaptation network intrusion detection algorithm based on class-balanced knowledge transfer and multi-structure domain alignment. *Research Square*, 2025.
- [270] Salwa Shakir Baawi, Zahraa Ch. Olewi, Abbas M. Ali Al-Muqarm, Dhiah Al-Shammary, and Fahim Sufi. Efficient malware detection based on machine learning for enhanced cloud privacy protection. *Evolving Systems*, 16(1):30, 2025.
- [271] Xuexing Zeng and T S Durrani. Band selection for hyperspectral images using copula-based mutual information. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 341–344, 2009.
- [272] 曹久慧, 卞桂荣, 陈剑, 李宝枝, 顾晓彬, 徐士月, 夏显文, and 张旭辉. 一种基于自注意力机制的变形监测方法, 2023. CN116378120A.
- [273] 张旭辉. 顾及邻近点相似性的机器学习变形分析方法研究. 硕士学位论文, 武汉大学, 2022.
- [274] 赵廷. 多波束声呐海底底质智能分类关键技术研究. 博士学位论文, 哈尔滨工程大学, 2022.
- [275] Hudson and Thames. Machine learning financial laboratory (MLFinLab). GitHub, 2021. URL: <https://github.com/hudson-and-thames/mlfinlab>.
- [276] Hudson and Thames. The ArbitrageLab package in Python. GitHub, 2024. URL: <https://github.com/hudson-and-thames/arbitragelab>.
- [277] Qiutong Wang. Social networks, asset allocation and portfolio diversification. Master's thesis, University of Waterloo, 2015.
- [278] 廖铁楠. 基于 Copula 熵选股及集成神经网络预测的投资组合管理研究. 硕士学位论文, 南京信息工程大学, 2023.
- [279] 朱仲儿. 多种机器学习方法的股票分类预测. 硕士学位论文, 上海师范大学, 2022.
- [280] Zhonger Zhu and Wansheng Wang. Stock type prediction based on multiple machine learning methods. *Journal of Intelligent Learning Systems and Applications*, 16(3):242–261, 2024.
- [281] 徐泽晖. 基于 GAS-CE-LGBM 的“一带一路”指数收益率预测研究. *统计学与应用*, 13(4):1431–1441, 2024.
- [282] Rafael Calsaverini and Renato Vicente. An information-theoretic approach to statistical dependence: Copula information. *EPL (Europhysics Letters)*, 88(6):68003, 2009.

- [283] Rafael S. Calsaverini. *Tópicos em Mecânica Estatística de Sistemas Complexos*. PhD thesis, Universidade de São Paulo, 2013.
- [284] Fadhah Amer Alanazi. Truncating Regular Vine Copula Based on Mutual Information: An Efficient Parsimonious Model for High-Dimensional Data. *Mathematical Problems in Engineering*, 2021:4347957, 2021.
- [285] 王念鸽. 基于互信息的 Vine Copula 模型的高频数据投资组合风险测度研究. 硕士学位论文, 浙江财经大学, 2023.
- [286] 熊靖宇. 基于 Copula 熵的行业风险溢出效应分析. 硕士学位论文, 东北财经大学, 2020.
- [287] 丁永辉. 中国金融系统的风险联动研究. 硕士学位论文, 东北财经大学, 2024.
- [288] Mengyuan Chen, Jilan Liu, Ning Zhang, and Yichao Zheng. Vulnerability analysis method based on network and copula entropy. *Entropy*, 26(3):192, 2024.
- [289] Omid M. Ardakani and Rawan Ajina. Tail risks in household finance. *Finance Research Letters*, 69:106065, 2024.
- [290] 孔祥永, 王浩, 袁伟, and 蔡明. 一种自动化特征工程信用评价系统及方法, 2021. CN114049198A.
- [291] 彭翊庭. 个人信用风险评估模型比较——基于 Copula 熵的特征选择. 硕士学位论文, 清华大学, 2022.
- [292] 王钊颖. 基于集成算法的上市公司绿色信贷风险评估研究. 硕士学位论文, 重庆大学, 2023.
- [293] Dabin Zhang, Ruibin Lin, Tingting Wei, Liwen Ling, and Junjie Huang. A novel deep transfer learning framework with adversarial domain adaptation: application to financial time-series forecasting. *Neural Computing and Applications*, 35:24037–24054, 2023.
- [294] Henryk Gurgul and Robert Syrek. Mutual information between Polish subindexes –the use of copula entropy around the time of the COVID-19 pandemic. *Statistics in Transition new series*, 25(1):23–41, 2024.
- [295] Henryk Gurgul and Robert Syrek. Mutual information between the main foreign subindices: The application of copula entropy around WHO’s declaration date at the time of the COVID-19 pandemic. *International Entrepreneurship Review*, 10(2):7–24, 2024.
- [296] 栗嵩林. 保险科技发展对保险公司经营绩效的影响研究. 硕士学位论文, 中央财经大学, 2023.
- [297] Jamal Uddin. Copula entropy-based variable screening with an application to insurance data analysis. Master’s thesis, Bowling Green State University, 2025.

- [298] Amanda Mahmudovic. Forecasting cryptocurrency returns in adaptive markets with an extended copula based feature selection and extended copula divergence hybrid loss function. Master's thesis, Linnaeus University, 2025.
- [299] M. Mohammadi, M. Hashempour, and M. Emadi. Semiparametric estimation of mutual information for elliptical copulas. In *The 7th Seminar on Copula Theory and its Applications*, page 44, 2023.
- [300] Guillaume Marrelec and Alain Giron. An inferential measure of dependence between two systems using Bayesian model comparison. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 55(3):1671–1683, 2025.
- [301] Edwin T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620, 1957.
- [302] Milan Bubák and Mirko Navara. Fitting copulas with maximal entropy. *Entropy*, 27(1):87, 2025.
- [303] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint arXiv:physics/0004057*, 2000.
- [304] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.
- [305] Laouen Belloli and Rubén Herzog. THOI: An efficient library for higher order interactions analysis based on gaussian copulas enhanced by batch-processing. GitHub, 2024. URL: <https://github.com/Laouen/THOI>.
- [306] Joe Suzuki and Tian-Le Yang. Generalization of LiNGAM that allows confounding. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 3540–3545. IEEE, 2024.
- [307] 陈琦琦. 基于 copula 熵和偏秩相关系数的时序因果网络构建算法及其应用. 硕士学位论文, 合肥工业大学, 2023.
- [308] Jing Yang and Xinzhi Rao. Copula entropy based causal network discovery from non-stationary time series. In *Pattern Recognition*, pages 115–131. Springer Cham, 2025.
- [309] Zihao Jiang, Tao Deng, Lei You, Siwei Feng, and Juncheng Jia. CoPruning: Exploring the parameter-gradient nonlinear correlation for neural network pruning using copula function. *OpenReview.net*, 2024. URL: <https://openreview.net/forum?id=Yqqa9aNwB0>.
- [310] 王筱萍. 基于分导 Copula 函数的分布估计算法研究. 博士学位论文, 兰州理工大学, 2013.

- [311] 张轶棠. 基于 copula entropy 的变数选取方法与节点选取方法. 硕士学位论文, 国立政治大学, 2024.
- [312] A.I. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, 1957.
- [313] Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.
- [314] Jian Ma. *copent*: Estimating copula entropy and transfer entropy in R. *arXiv preprint arXiv:2005.14025*, 2021.
- [315] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):66138, 2004.
- [316] Michael Irwin Jordan. *Learning in Graphical Models*. MIT press, 1999.
- [317] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.
- [318] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.
- [319] Neville Kenneth Kitson, Anthony C. Constantinou, Zhigao Guo, Yang Liu, and Kiattikum Chobtham. A survey of Bayesian network structure learning. *Artificial Intelligence Review*, 56(8):8721–8814, 2023.
- [320] Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.
- [321] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.
- [322] Claudia Czado. *Analyzing Dependent Data with Vine Copulas*. Springer Cham, 2019.
- [323] Dorota Kurowicka and Harry Joe, editors. *Dependence Modeling: Vine Copula Handbook*. World Scientific, 2010.
- [324] Tim Bedford and Roger M Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence*, 32(1):245–268, 2001.
- [325] Tim Bedford and Roger M Cooke. Vines—a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068, 2002.

- [326] Ingrid Hobæk Haff, Kjersti Aas, and Arnoldo Frigessi. On the simplified pair-copula construction —simply useful or too simplistic? *Journal of Multivariate Analysis*, 101(5):1296–1310, 2010.
- [327] Thomas Nagler. Simplified vine copula models: State of science and affairs. *Risk Sciences*, 1:100022, 2025.
- [328] Arthur Asuncion and David Newman. UCI machine learning repository, 2007.
- [329] Albert M Liebetrau. *Measures of Association*. SAGE, 1983.
- [330] Philippe Barbe, Christian Genest, Kilani Ghoudi, and Bruno Rémillard. On Kendall’s process. *Journal of Multivariate Analysis*, 58(2):197–229, 1996.
- [331] Harry Joe. Multivariate concordance. *Journal of Multivariate Analysis*, 35(1):12–30, 1990.
- [332] Edward F. Wolff. N-dimensional measures of dependence. *Stochastica*, 4(3):175–188, 1980.
- [333] Harry Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall, 1997.
- [334] Corrado Gini. *L’ammontare e la composizione della ricchezza delle nazioni*, volume 62. Fratelli Bocca, 1914.
- [335] Javad Behboodian, Ali Dolati, and Manuel Úbeda-Flores. A multivariate version of Gini’s rank association coefficient. *Statistical Papers*, 48:295–304, 2007.
- [336] B. Schweizer and E. F. Wolff. On Nonparametric Measures of Dependence for Random Variables. *The Annals of Statistics*, 9(4):879–885, 1981.
- [337] Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10:441–451, 1959.
- [338] Friedrich Schmid, Rafael Schmidt, Thomas Blumentritt, Sandra Gaißer, and Martin Ruppert. Copula-based measures of multivariate association. In *Copula Theory and Its Application*, pages 209–236, 2010.
- [339] Julie Josse and Susan Holmes. Measuring multivariate association and beyond. *Statistics Surveys*, 10:132–167, 2016.
- [340] National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey, 2013-2014.
- [341] Edward I. George. The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308, 2000.

- [342] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [343] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [344] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [345] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B-Methodological*, 58(1):267–288, 1996.
- [346] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.
- [347] 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.
- [348] 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.
- [349] 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.
- [350] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.
- [351] Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.
- [352] Hui Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.
- [353] 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.
- [354] Wassily Hoeffding. A non-parametric test of independence. *Annals of Mathematical Statistics*, 19(4):546–557, 1948.
- [355] 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.

- [356] 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.
- [357] David G. Kleinbaum and Mitchel Klein. *Survival Analysis*. Springer New York, NY, 2011.
- [358] Jianqing Fan, Gang Li, and Runze Li. *Contemporary Multivariate Analysis and Design of Experiments*, chapter An Overview on Variable Selection for Survival Analysis, pages 315–336. World Scientific, 2005.
- [359] Stephen Salerno and Yi Li. High-dimensional survival analysis: Methods and applications. *Annual Review of Statistics and Its Application*, 10:25–49, 2023.
- [360] Helen Beebee, Christopher Hitchcock, and Peter Menzies, editors. *The Oxford Handbook of Causation*. Oxford University Press, 2009.
- [361] Stephen Mumford and Rani Lill Anjum. *Causation: A Very Short Introduction*. Oxford University Press, 2013.
- [362] Phyllis Illari and Federica Russo. *Causality: Philosophical Theory Meets Scientific Practice*. Oxford University Press, 2014.
- [363] Judea Pearl. *Causality*. Cambridge University Press, 2009.
- [364] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.
- [365] Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery*, 12(2):e1449, 2022.
- [366] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D’ya like DAGs? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.
- [367] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.
- [368] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.
- [369] Norbert Wiener. The theory of prediction. Modern mathematics for engineers. *New York*, 165(6), 1956.
- [370] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

- [371] Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.
- [372] Ali Shojaie and Emily B. Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289–319, 2022.
- [373] Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical Review Letters*, 103(23):238701, 2009.
- [374] 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.
- [375] 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.
- [376] 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.
- [377] A. Chiuso and G. Pillonetto. System identification: A machine learning perspective. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):281–304, 2019.
- [378] Gustau Camps-Valls, Andreas Gerhardus, Urmilla Ninad, Gherardo Varando, Georg Martius, Emili Balaguer-Ballester, Ricardo Vinuesa, Emiliano Diaz, Laure Zanna, and Jakob Runge. Discovering causal relations and equations from data. *Physics Reports*, 1044:1–68, 2023.
- [379] Joshua S. North, Christopher K. Wikle, and Erin M. Schliep. A review of data-driven discovery for dynamic systems. *International Statistical Review*, 91(3):464–492, 2023.
- [380] Alan A. Kaptanoglu, Lanyue Zhang, Zachary G. Nicolaou, Urban Fasel, and Steven L. Brunton. Benchmarking sparse system identification with low-dimensional chaos. *Nonlinear Dynamics*, 111(14):13143–13164, 2023.
- [381] Max Champneys, Gerben I. Beintema, R. Tóth, Maarten Schoukens, and Timothy J. Rogers. Baseline results for selected nonlinear system identification benchmarks. *IFAC-PapersOnLine*, 58(15):474–479, 2024.
- [382] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

- [383] Gianluigi Pillonetto, Francesco Dinuzzo, Tianshi Chen, Giuseppe De Nicolao, and Lennart Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682, 2014.
- [384] Lawrence C Evans. Entropy and partial differential equations. *Lecture Notes at UC Berkeley*, 2004.
- [385] Y. Sinai. Kolmogorov-Sinai entropy. *Scholarpedia*, 4(3):2034, 2009. revision #91407.
- [386] Pasquale Nardone and Giorgio Sonnino. Entropy of difference: A new tool for measuring complexity. *Axioms*, 13(2):130, 2024.
- [387] K.R. Chernyshov and E.Ph. Jharko. An information-theoretic approach to system identification with applying Tsallis entropy. *IFAC-PapersOnLine*, 51(6):24–29, 2018.
- [388] A.A. Stoorvogel and Jan van Schuppen. System identification with information theoretic criteria. Technical Report R 9513, Centrum voor Wiskunde en Informatica, 1995.
- [389] Edward N Lorenz. Maximum simplification of the dynamic equations. *Tellus*, 12(3):243–254, 1960.
- [390] Otto E Rössler. An equation for continuous chaos. *Physics Letters A*, 57(5):397–398, 1976.
- [391] Otto E Rössler. An equation for hyperchaos. *Physics Letters A*, 71(2-3):155–157, 1979.
- [392] Henry D. I. Abarbanel, Reggie Brown, John J. Sidorowich, and Lev Sh. Tsimring. The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics*, 65:1331–1392, 1993.
- [393] G.C. Carter. Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–255, 1987.
- [394] Svante Björklund and Lennart Ljung. A review of time-delay estimation techniques. In *42nd IEEE International Conference on Decision and Control*, volume 3, pages 2502–2507, 2003.
- [395] N.J.I. Mars and G.W. van Arragon. Time delay estimation in non-linear systems using average amount of mutual information analysis. *Signal Processing*, 4(2):139–153, 1982.
- [396] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
- [397] Sheldon M Ross. *A First Course in Probability*. Pearson, 2020.
- [398] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2019.

- [399] Keya Rani Das and AHMR Imon. A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, 5(1):5–12, 2016.
- [400] Henry C Thode. *Testing for Normality*. CRC press, 2002.
- [401] Bruno Ebner and Norbert Henze. Tests for multivariate normality—a critical review with emphasis on weighted  $L^2$ -statistics. *TEST*, 29(4):845–892, 2020.
- [402] Wanfang Chen and Marc G Genton. Are you all normal? it depends! *International Statistical Review*, 91(1):114–139, 2023.
- [403] B. W. Yap and C. H. Sim. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155, 2011.
- [404] Berna Yazici and Senay Yolacan. A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, 77(2):175–183, 2007.
- [405] S. S. Shapiro, M. B. Wilk, and H. J. Chen. A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63(324):1343–1372, 1968.
- [406] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.
- [407] Bernard L Welch. The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.
- [408] Jay L. Devore, Kenneth N. Berk, and Matthew A. Carlton. *Modern Mathematical Statistics with Applications*. Springer Cham, 2021.
- [409] John A Rice. *Mathematical Statistics and Data Analysis*. Thomson Brooks/Cole, 2007.
- [410] N. A. C. Cressie and H. J. Whitford. How to use the two sample t-test. *Biometrical Journal*, 28(2):131–148, 1986.
- [411] Vance W. Berger and YanYan Zhou. Kolmogorov–Smirnov test: Overview. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, 2014.
- [412] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.
- [413] Michèle Basseville and Igor V Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.
- [414] B. E. Brodsky and B. S. Darkhovsky. *Nonparametric Methods in Change Point Problems*. Springer Dordrecht, 1993.

- [415] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.
- [416] Ewan Stafford Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.
- [417] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.
- [418] Samaneh Aminikhahhahi and Diane J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.
- [419] David J. Gross. The role of symmetry in fundamental physics. *Proceedings of the National Academy of Sciences*, 93(25):14256–14259, 1996.
- [420] Kurt Sundermeyer. *Symmetries in Fundamental Physics*. Springer Cham, 2014.
- [421] Mark A Armstrong. *Groups and Symmetry*. Springer New York, NY, 1988.
- [422] Pierre Simon de Laplace. *Théorie Analytique des Probabilités*. Courcier, 1820.
- [423] Paul Bartha and Richard Johns. Probability and symmetry. *Philosophy of Science*, 68(S3):S109–S122, 2001.
- [424] Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer New York, NY, 2005.
- [425] A Philip Dawid. Probability, symmetry and frequency. *The British Journal for the Philosophy of Science*, 36(2):107–128, 1985.
- [426] Philip Prescott. Student’s t-tests. In *Encyclopedia of Statistical Sciences*, pages 8371–8377. John Wiley & Sons, 2006.
- [427] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.
- [428] Ronald H. Randles. Wilcoxon signed rank test. In *Encyclopedia of Statistical Sciences*, pages 9150–9153. John Wiley & Sons, 2006.
- [429] E.L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Cham, 2022.
- [430] James S Allison and Charl Pretorius. A Monte Carlo evaluation of the performance of two new tests for symmetry. *Computational Statistics*, 32(4):1323–1338, 2017.
- [431] L Baringhaus and N Henze. A characterization of and new consistent tests for symmetry. *Communications in statistics-theory and methods*, 21(6):1555–1566, 1992.

- [432] Vladimir Božin, Bojana Milošević, Ya Yu Nikitin, and Marko Obradović. New characterization-based symmetry tests. *Bulletin of the Malaysian Mathematical Sciences Society*, 43:297–320, 2020.
- [433] Weiwen Miao, Yulia R Gel, and Joseph L Gastwirth. A new test of symmetry about an unknown median. In *Random walk, sequential analysis and related topics: A festschrift in honor of Yuan-Shih Chow*, pages 199–214. World Scientific, 2006.
- [434] Bojana Milošević and Marko Obradović. Characterization based symmetry tests and their asymptotic efficiencies. *Statistics & Probability Letters*, 119:155–162, 2016.
- [435] Bojana Milošević and Marko Obradović. Comparison of efficiencies of some symmetry tests around an unknown centre. *Statistics*, 53(1):43–57, 2019.
- [436] Antonietta Mira and. Distribution-free test for symmetry based on bonferroni’s measure. *Journal of Applied Statistics*, 26(8):959–972, 1999.
- [437] Paul Cabilio and Joe Masaro. A simple test of symmetry about an unknown median. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 24(3):349–361, 1996.
- [438] Andrey Feuerverger and Roman A Mureika. The empirical characteristic function and its applications. *The Annals of Statistics*, 5(1):88–97, 1977.
- [439] Daniel Gaigall. Rothman–Woodrooffe symmetry test statistic revisited. *Computational Statistics & Data Analysis*, 142:106837, 2020.
- [440] VV Litvinova. New nonparametric test for symmetry and its asymptotic efficiency. *Vestnik St. Petersburg University Mathematics*, 34(4):12–14, 2001.
- [441] Ya Yu Nikitin and Mohammad Ahsanullah. New U-empirical tests of symmetry based on extremal order statistics, and their efficiencies. In *Mathematical Statistics and Limit Theorems: Festschrift in Honour of Paul Deheuvels*, pages 231–248. Springer Cham, 2015.
- [442] Robert J. Serfling. Multivariate symmetry and asymmetry. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, 2014.
- [443] George E. Martin. *Transformation Geometry: An Introduction to Symmetry*. Springer New York, NY, 1982.
- [444] Gábor J Székely and Maria L Rizzo. *The Energy of Data and Distance Correlation*. Chapman and Hall/CRC, 2023.
- [445] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

- [446] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.
- [447] Tianhong Sheng and Bharath K. Sriperumbudur. On distance and kernel measures of conditional dependence. *Journal of Machine Learning Research*, 24(7):1–16, 2023.
- [448] Junli Lin. *Copula Versions of RKHS-Based and Distance-Based Criteria*. PhD thesis, Pennsylvania State University, 2017.
- [449] Björn Böttcher. Copula versions of distance multivariance and dHSIC via the distributional transform –a general approach to construct invariant dependence measures. *Statistics*, 54(3):577–594, 2020.
- [450] Barnabás Póczos, Zoubin Ghahramani, and Jeff Schneider. Copula-based kernel dependency measures. *arXiv preprint arXiv:1206.4682*, 2012.
- [451] Björn Böttcher, Martin Keller-Ressel, and René L. Schilling. Distance multivariance: New dependence measures for random vectors. *The Annals of Statistics*, 47(5):2757–2789, 2019.
- [452] Gábor J. Székely and Maria L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.
- [453] Jérémie Kellner and Alain Celisse. A one-sample test for normality with kernel methods. *Bernoulli*, 25(3):1816–1837, 2019.
- [454] Gábor J. Székely and Maria L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.
- [455] Zaïd Harchaoui, Eric Moulines, and Francis Bach. Kernel change-point analysis. In *Advances in Neural Information Processing Systems*, volume 21, 2008.
- [456] Bo Chen, Feifei Chen, Junxin Wang, and Tao Qiu. An efficient and distribution-free symmetry test for high-dimensional data based on energy statistics and random projections. *Computational Statistics & Data Analysis*, 206:108123, 2025.
- [457] Dag Tjøstheim, Håkon Otneim, and Bård Støve. Statistical Dependence: Beyond Pearson’s  $\rho$ . *Statistical Science*, 37(1):90–109, 2022.
- [458] Wenliang Pan, Yuan Tian, Xueqin Wang, and Heping Zhang. Ball Divergence: Nonparametric two sample test. *The Annals of Statistics*, 46(3):1109–1137, 2018.
- [459] Kai Zhang. BET on independence. *Journal of the American Statistical Association*, 114(528):1620–1637, 2019.

- [460] Wolfgang Trutschnig. On a strong metric on the space of copulas and its induced dependence measure. *Journal of Mathematical Analysis and Applications*, 384(2):690–705, 2011.
- [461] Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.
- [462] C Genest, J G Nešlehová, B Rémillard, and O A Murphy. Testing for independence in arbitrary distributions. *Biometrika*, 106(1):47–68, 2019.
- [463] Arturo Erdely. A subcopula based dependence measure. *Kybernetika*, 53(2):231–243, 2017.
- [464] Xiaofeng Shao and Jingsi Zhang. Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318, 2014.
- [465] Fred Viole and David N. Nawrocki. Deriving nonlinear correlation coefficients from partial moments. *SSRN:2148522*, 2012.
- [466] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B-Methodological*, 41(1):1–15, 1979.
- [467] Chun Li and Xiaodan Fan. On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics*, 12(3):e1489, 2020.
- [468] Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019.
- [469] Lasse Petersen and Niels Richard Hansen. Testing conditional independence via quantile regression based partial copulas. *Journal of Machine Learning Research*, 22(70):1–47, 2021.
- [470] Trevor Park, Xiaofeng Shao, and Shun Yao. Partial martingale difference correlation. *Electronic Journal of Statistics*, 9(1):1492–1517, 2015.
- [471] Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6):3070–3102, 2021.
- [472] Zhen Huang, Nabarun Deb, and Bodhisattva Sen. Kernel partial correlation coefficient — a measure of conditional dependence. *Journal of Machine Learning Research*, 23(216):1–58, 2022.
- [473] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*, 2018.

- [474] Samuel Burkart and Franz J Király. Predictive independence testing, predictive conditional independence testing, and predictive graphical modelling. *arXiv preprint arXiv:1711.05869*, 2017.
- [475] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [476] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947, 2017.
- [477] Octavio César Mesner and Cosma Rohilla Shalizi. Conditional mutual information estimation for mixed, discrete and continuous data. *IEEE Transaction on Information Theory*, 67(1):464–484, 2021.
- [478] George Udny Yule and Olaus Magnus Friedrich Erdmann Henrici. On the theory of correlation for any number of variables, treated by a new system of notation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 79(529):182–193, 1907.
- [479] Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.
- [480] Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.
- [481] Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022.
- [482] Anton Rask Lundborg, Ilmun Kim, Rajen D. Shah, and Richard J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *arXiv preprint arXiv:2211.02039*, 2024.
- [483] Kanti V Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.
- [484] Kanti V Mardia. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 115–128, 1974.
- [485] Norbert Henze and Bernd Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in statistics-Theory and Methods*, 19(10):3595–3617, 1990.

- [486] J Patrick Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(2):115–124, 1982.
- [487] J Patrick Royston. Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 32(2):121–133, 1983.
- [488] Patrick Royston. Approximating the shapiro-wilk w-test for non-normality. *Statistics and computing*, 2:117–119, 1992.
- [489] Jurgen A Doornik and Henrik Hansen. An omnibus test for univariate and multivariate normality. *Oxford bulletin of economics and statistics*, 70:927–939, 2008.
- [490] Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769, 1954.
- [491] Theodore W. Anderson. Anderson–Darling Tests of Goodness-of-Fit. In *International Encyclopedia of Statistical Science*. Springer Berlin, Heidelberg, 2025.
- [492] James A. Koziol. A class of invariant procedures for assessing multivariate normality. *Biometrika*, 69(2):423–427, 1982.
- [493] Vassilly Voinov, Natalie , Pya, Rashid , Makarov, and Yevgeniy Voinov. New invariant and consistent chi-squared type goodness-of-fit tests for multivariate normality and a related comparative simulation study. *Communications in Statistics - Theory and Methods*, 45(11):3249–3263, 2016.
- [494] Charles E. McCulloch. Relationships among some chi-square goodness of fit statistics. *Communications in Statistics - Theory and Methods*, 14(3):593–603, 1985.
- [495] K. O. Dzaparidze and M. S. Nikulin. On a modification of the standard statistics of pearson. *Theory of Probability & Its Applications*, 19(4):851–853, 1975.
- [496] Norbert Henze and Thorsten Wagner. A new approach to the BHEP tests for multivariate normality. *Journal of Multivariate Analysis*, 62(1):1–23, 1997.
- [497] D. R. Cox and N. J. H. Small. Testing multivariate normality. *Biometrika*, 65(2):263–272, 1978.
- [498] Philip Dörr, Bruno Ebner, and Norbert Henze. A new test of multivariate normality by a double estimation in a characterizing PDE. *Metrika*, 84(3):401–427, 2021.

- [499] Philip Dörr, Bruno Ebner, and Norbert Henze. Testing multivariate normality by zeros of the harmonic oscillator in characteristic function spaces. *Scandinavian Journal of Statistics*, 48(2):456–501, 2021.
- [500] Bruno Ebner, Norbert Henze, and David Stroeder. Testing normality in any dimension by Fourier methods in a multivariate Stein equation. *Canadian Journal of Statistics*, 50(3):992–1033, 2022.
- [501] Norbert Henze and María Dolores Jiménez-Gamero. A new class of tests for multinormality with i.i.d. and GARCH data based on the empirical moment generating function. *TEST*, 28:499–521, 2019.
- [502] Norbert Henze and Jaco Visagie. Testing for normality in any dimension based on a partial differential equation involving the moment generating function. *Annals of the Institute of Statistical Mathematics*, 72:1109–1136, 2020.
- [503] James A. Koziol. A note on measures of multivariate kurtosis. *Biometrical Journal*, 31(5):619–624, 1989.
- [504] J. F. Malkovich and A. A. Afifi. On tests for multivariate normality. *Journal of the American Statistical Association*, 68(341):176–179, 1973.
- [505] Alessandro Manzotti and Adolfo J. Quiroz. Spherical harmonics in quadratic forms for testing multivariate normality. *Test*, 10(1):87–104, 2001.
- [506] T. F. Móri, V. K. Rohatgi, and G. J. Székely. On multivariate skewness and kurtosis. *Theory of Probability & Its Applications*, 38(3):547–551, 1994.
- [507] Jan Pudełko. On a new affine invariant and consistent test for multivariate normality. *Probability and Mathematical Statistics*, 25:43–54, 2005.
- [508] Selcuk Korkmaz, Dincer Goksuluk, and Gokmen Zararsiz. MVN: An R Package for Assessing Multivariate Normality. *The R Journal*, 6(2):151–162, 2014.
- [509] Natalya Pya, Vassilly Voinov, Rashid Makarov, and Yevgeniy Voinov. mvnTest: Goodness of Fit Tests for Multivariate Normality. CRAN, 2016. R package version 1.1-0, URL: <https://cran.r-project.org/package=mvnTest>.
- [510] Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons, 2015.
- [511] Simon Hediger, Loris Michel, and Jeffrey Näf. On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis*, 170:107435, 2022.

- [512] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.
- [513] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.
- [514] Andrey Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.
- [515] Nikolai Smirnov. Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.
- [516] Harald Cramér. *Sannolikhetskalkylen Och Några Av Dess Användningar*. Gjallarhornet, 1927.
- [517] Richard Von Mises. *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*. Deuticke, 1931.
- [518] Nicolaas H. Kuiper. Tests concerning random points on a circle. *Indagationes Mathematicae (Proceedings)*, 63:38–47, 1960.
- [519] Aaditya Ramdas, Nicolás García Trillo, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.
- [520] Connor Dowd. A new ECDF two-sample test statistic. *arXiv preprint arXiv:2007.01360*, 2020.
- [521] A. N. Pettitt. A two-sample Anderson-Darling rank statistic. *Biometrika*, 63(1):161–168, 1976.
- [522] Sinda Ammous, Olivier Bouaziz, Anatole Dedecker, Jérôme Dedecker, Jonathan El Methni, Mohamed Mellouk, and Florence Muri. The **robustTest** package: two-sample tests revisited. *arXiv preprint arXiv:2211.08784*, 2024.
- [523] E. L. Lehmann. Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, 22(2):165–179, 1951.
- [524] M. Rosenblatt. Limit theorems associated with variants of the von mises statistic. *The Annals of Mathematical Statistics*, 23(4):617–623, 1952.
- [525] Jin Zhang. Powerful two-sample tests based on the likelihood ratio. *Technometrics*, 48(1):95–103, 2006.

- [526] Fazil Aliev, Levent Özbek, Mehmet Fedai Kaya, Coşkun Kuş, Hon Keung Tony Ng, and Haikady N Nagaraja. A nonparametric test for the two-sample problem based on order statistics. *Communications in Statistics-Theory and Methods*, 53(10):3688–3712, 2024.
- [527] Apratim Guha and Tom Chothia. A two sample test based on mutual information. *Calcutta Statistical Association Bulletin*, 66(1-2):39–54, 2014.
- [528] Robert Drake and Apratim Guha. A mutual information-based k-sample test for discrete distributions. *Journal of Applied Statistics*, 41(9):2011–2027, 2014.
- [529] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.
- [530] Marta Cousido-Rocha, Jacobo de Uña-Álvarez, and Jeffrey D. Hart. A two-sample test for the equality of univariate marginal distributions for high-dimensional data. *Journal of Multivariate Analysis*, 174:104537, 2019.
- [531] G. Fasano and A. Franceschini. A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170, 1987.
- [532] J. A. Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627, 1983.
- [533] Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, volume 24, 2011.
- [534] Regina Y. Liu and Kesar Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260, 1993.
- [535] William H. Kruskal. A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, 23(4):525–540, 1952.
- [536] Karin Neubert and Edgar Brunner. A studentized permutation test for the non-parametric Behrens–Fisher problem. *Computational Statistics & Data Analysis*, 51(10):5192–5204, 2007.
- [537] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer Berlin, Heidelberg, 2012.
- [538] Yichuan Bai and Lynna Chu. A robust framework for graph-based two-sample tests using weights. *arXiv preprint arXiv:2307.12325*, 2023.

- [539] Zhen Huang and Bodhisattva Sen. A kernel measure of dissimilarity between M distributions. *Journal of the American Statistical Association*, 119(548):3020–3032, 2024.
- [540] Alexandros Karatzoglou, Alexandros Smola, Kurt Hornik, and Achim Zeileis. `kernlab` – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.
- [541] Marta Cousido-Rocha and Jacobo de Uña-Álvarez. `TwoSampleTest.HD`: An R package for the two-sample problem with high-dimensional data. *The R Journal*, 15:79–92, 2023.
- [542] Connor Puritz, Elan Ness-Cohn, and Rosemary Braun. `fasano.franceschini.test`: An implementation of a multivariate KS test in R. *The R Journal*, 15:159–171, 2023.
- [543] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- [544] Jaxk Reeves, Jien Chen, Xiaolan L. Wang, Robert Lund, and Qi Qi Lu. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915, 2007.
- [545] A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512, 1974.
- [546] Nicholas A. James and David S. Matteson. `ecp`: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62(7):1–25, 2015.
- [547] Xinyuan Fan and Weichi Wu. Random interval distillation for detecting multiple changes in general dependent data. *arXiv preprint arXiv:2403.00600*, 2024.
- [548] Euan T. McGonigle and Haeran Cho. Nonparametric data segmentation in multivariate time series via joint characteristic functions. *arXiv preprint arXiv:arXiv:2305.07581*, 2023.
- [549] Gordon J. Ross. Nonparametric detection of multiple location-scale change points via wild binary segmentation. *arXiv preprint arXiv:2107.01742*, 2021.
- [550] Michael Messer, Stefan Albert, and Gaby Schneider. The multiple filter test for change point detection in time series. *Metrika*, 81(6):589–607, 2018.
- [551] Michael Messer. Bivariate change point detection: Joint detection of changes in expectation and variance. *Scandinavian Journal of Statistics*, 49(2):886–916, 2022.
- [552] Tengyao Wang and Richard J. Samworth. High Dimensional Change Point Estimation via Sparse Projection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):57–83, 2017.

- [553] Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 77(2):475–507, 2015.
- [554] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.
- [555] Thomas Grundy, Rebecca Killick, and Gueorgui Mihaylov. High-dimensional changepoint detection via a geometrically inspired mapping. *Statistics and Computing*, 30(4):1155–1166, 2020.
- [556] Birte Eichinger and Claudia Kirch. A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564, 2018.
- [557] Zifeng Zhao, Feiyu Jiang, and Xiaofeng Shao. Segmenting time series via self-normalisation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1699–1725, 2022.
- [558] Jie Ding, Yu Xiang, Lu Shen, and Vahid Tarokh. Multiple change point analysis: Fast implementation and strong consistency. *IEEE Transactions on Signal Processing*, 65(17):4495–4510, 2017.
- [559] Andreas Anastasiou and Piotr Fryzlewicz. Detecting multiple generalized change-points by isolating single ones. *Metrika*, 85(2):141–174, 2022.
- [560] Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.
- [561] Piotr Fryzlewicz. Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society*, 49(4):1027–1070, 2020.
- [562] Tijana Levajković and Michael Messer. Multiscale change point detection via gradual bandwidth adjustment in moving sum processes. *Electronic Journal of Statistics*, 17(1):70–101, 2023.
- [563] Jiaqi Li, Likai Chen, Weining Wang, and Wei Biao Wu.  $l^2$  Inference for change points in high-dimensional time series via a Two-Way MOSUM. *The Annals of Statistics*, 52(2):602–627, 2024.
- [564] Toby Dylan Hocking, Guillem Rigaill, Paul Fearnhead, and Guillaume Bourque. Constrained dynamic programming and supervised penalty learning algorithms for peak detection in genomic data. *Journal of Machine Learning Research*, 21(87):1–40, 2020.

- [565] Per August Jarval Moen, Ingrid Kristine Glad, and Martin Tveten. Efficient sparsity adaptive changepoint estimation. *arXiv preprint arXiv:2306.04702*, 2023.
- [566] Reza Drikvandi and Reza Modarres. A distribution-free method for change point detection in non-sparse high dimensional data. *Journal of Computational and Graphical Statistics*, 34(1):290–305, 2025.
- [567] Jun Li, Minya Xu, Ping-Shou Zhong, and Lingjun Li. Change point detection in the mean of high-dimensional time series data under dependence. *arXiv preprint arXiv:1903.07006*, 2019.
- [568] Sean Ryan and Rebecca Killick. Detecting changes in covariance via random matrix theory. *Technometrics*, 65(4):480–491, 2023.
- [569] Changliang Zou, Guanghui Wang, and Runze Li. Consistent selection of the number of change-points via sample-splitting. *The Annals of Statistics*, 48(1):413–439, 2020.
- [570] Rebecca Killick and Idris A. Eckley. **changepoint**: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.
- [571] Alexander Meier, Claudia Kirch, and Haeran Cho. **mosum**: A package for moving sums in change-point analysis. *Journal of Statistical Software*, 97(8):1–42, 2021.
- [572] Shubo Sun, Zifeng Zhao, Feiyu Jiang, and Xiaofeng Shao. **SNSeg**: An R package for time series segmentation via self-normalization. *The R Journal*, 16:46–66, 2025.
- [573] Andreas Anastasiou, Yining Chen, Haeran Cho, and Piotr Fryzlewicz. **breakfast**: Methods for fast multiple change-point/break-point detection and estimation. CRAN, 2024. URL: <https://CRAN.R-project.org/package=breakfast>.
- [574] Vincent Runge, Toby Dylan Hocking, Gaetano Romano, Fatemeh Afghah, Paul Fearnhead, and Guillem Rigaill. **gfpop**: An R package for univariate graph-constrained change-point detection. *Journal of Statistical Software*, 106(6):1–39, 2023.
- [575] Emmanuel Pilliat, Alexandra Carpentier, and Nicolas Verzelen. Optimal multiple change-point detection for high-dimensional data. *Electronic Journal of Statistics*, 17(1):1240–1315, 2023.
- [576] Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1):479–487, 1988.
- [577] Constantino Tsallis. The nonadditive entropy  $S_q$  and its applications in physics and elsewhere: Some remarks. *Entropy*, 13(10):1765–1804, 2011.

- [578] Velimir M. Ilić and Miomir S. Stanković. Generalized Shannon-Khinchin axioms and uniqueness theorem for pseudo-additive entropies. *Physica A: Statistical Mechanics and its Applications*, 411:138–145, 2014.
- [579] Frank Lad, Giuseppe Sanfilippo, and Gianna Agrò. Extropy: Complementary Dual of Entropy. *Statistical Science*, 30(1):40–58, 2015.
- [580] Narayanaswamy Balakrishnan, Francesco Buono, and Maria Longobardi. On weighted extropies. *Communications in Statistics - Theory and Methods*, 51(18):6250–6267, 2022.
- [581] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.
- [582] Maria Ribeiro, Teresa Henriques, Luísa Castro, André Souto, Luís Antunes, Cristina Costa-Santos, and Andreia Teixeira. The entropy universe. *Entropy*, 23(2):222, 2021.
- [583] Misaki Ozawa and Nina Javerzat. Perspective on physical interpretations of Rényi entropy in statistical mechanics. *Europhysics Letters*, 147(1):11001, 2024.
- [584] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [585] Lisa Borland, Angel R. Plastino, and Constantino Tsallis. Information gain within nonextensive thermostatistics. *J. Math. Phys.*, 39(12):6490–6501, 1998.
- [586] S. Furuichi, K. Yanagi, and K. Kuriyama. Fundamental properties of Tsallis relative entropy. *J. Math. Phys.*, 45(12):4868–4877, 2004.
- [587] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.
- [588] Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, (136):210–271, 1909.
- [589] Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8(1-2):85–108, 1963.
- [590] David F Kerridge. Inaccuracy and inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(1):184–194, 1961.

- [591] Reuven Y. Rubinstein and Dirk P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer New York, NY, 2004.
- [592] 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.
- [593] 刘辰昊, 张蕾, and 庞思平. 含能材料机器学习研究的数据优化策略. 含能材料, 2025.
- [594] 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.
- [595] Florian Hodel. *cylcop*: Circular-linear copulas with angular symmetry for movement data. CRAN, 2022. R package version 0.2.0, URL: <https://cran.r-project.org/package=cylcop>.
- [596] Robin A.A. Ince, Jan-Mathijs Schoffelen, Lukas Snoek, and Danylo Ulianych. *gcmi* : Gaussian-Copula Mutual Information. GitHub, 2020. URL: <https://github.com/robince/gcmi>.
- [597] Danylo Ulianych. The *pytorch-mighty* package in Python. GitHub, 2023. URL: <https://github.com/dizcza/pytorch-mighty>.
- [598] BraiNets. The *HOI* package in Python. Github, 2024. URL: <https://brainnets.github.io/hoi/>.
- [599] Etienne Combrisson, Ruggero Basanisi, Vinicius Lima Cordeiro, Robin A.a Ince, and Andrea Brovelli. *Frites*: A Python package for functional connectivity analysis and group-level statistics of neurophysiological data. *Journal of Open Source Software*, 7(79):3842, 2022.
- [600] Etienne Combrisson, Timothy Nest, Andrea Brovelli, Robin A. A. Ince, Juan L. P. Soto, Aymeric Guillot, and Karim Jerbi. *Tensorpac*: An open-source Python toolbox for tensor-based phase-amplitude coupling measurement in electrophysiological brain signals. *PLoS computational biology*, 16(10):e1008302, 2020.
- [601] Institute for Advanced Brain Studies. *driada*: Dimensionality reduction for integrated activity data. Github, 2024. URL: <https://github.com/iabs-neuro/driada>.

- [602] Nina Kudryashova, Theoklitos Amvrosiadis, Nathalie Dupuy, Nathalie Rochefort, and Arno Onken. Parametric Copula-GP model for analyzing multidimensional neuronal and behavioral relationships. *PLoS computational biology*, 18(1):e1009799, 2022.
- [603] Nina Kudryashova. Parametric Copula-GP framework. GitHub, 2022. URL: <https://github.com/NinelK/CopulaGP>.
- [604] Tianren Qin. *Polars-ds*: Polars extension for general data science use. GitHub, 2024. URL: [https://github.com/abstractqqq/polars\\_ds\\_extension](https://github.com/abstractqqq/polars_ds_extension).
- [605] Alessandro Crimi. The *effconnp* package in Python. GitHub, 2025. URL: <https://github.com/alecrimi/effconnp>.
- [606] Kai Xu. The *CopEnt.jl* package in Julia. Github, 2021. URL: <https://github.com/xukai92/CopEnt.jl>.
- [607] Kristian Agasøster Haaga. The *CausalityTools.jl* package v2.10.1 in Julia. Github, 2023. URL: <https://github.com/JuliaDynamics/CausalityTools.jl>.
- [608] Oskar Laverny and Santiago Jimenez. *Copulas.jl*: A fully distributions.jl-compliant copula package. *Journal of Open Source Software*, 9(94):6189, 2024. URL: <https://github.com/lrnv/Copulas.jl>.
- [609] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. *FieldTrip*: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011(1):156869, 2011.
- [610] Igor Volobouev. *NPStat* – non-parametric statistical modeling and analysis in C++ and Python. HEPForge, 2023. URL: <https://npstat.hepforge.org>.