中国科学: 数学 2024 年 第 54 卷 第 8 期: 1169 ~ 1194
SCIENTIA SINICA Mathematica
《中国科学》杂志社
SCIENCE CHINA PRESS
综 述

非线性相依数据关联分析

周叶青[1,2,3], 许凯[4], 朱利平[5*]

1. 同济大学数学科学学院, 上海 200092;
2. 同济大学经济与管理学院, 上海 200092;
3. 同济大学智能计算与应用教育部重点实验室, 上海 200092;
4. 安徽师范大学数学与统计学院, 芜湖 241002;
5. 中国人民大学统计与大数据研究院, 北京 100872
E-mail: zhouyeqing@tongji.edu.cn, tjxxukai@163.com, zhu.liping@ruc.edu.cn

收稿日期: 2023-06-09; 接受日期: 2024-03-22; 网络出版日期: 2024-07-25; * 通信作者
国家重点研发计划 (批准号: 2023YFA1008702)、国家自然科学基金 (批准号: 12001405, 12271005, 11901006, 12225113 和 12171477)、上海市自然科学基金 (批准号: 23ZR1469000)、安徽省自然科学基金 (批准号: 2308085Y06 和 1908085QA06)、安徽省青年拔尖人才青年学者 (2023) 和中央高校基本科研业务费专项资金 (批准号: 22120240274) 资助项目

摘要 度量和检验复杂数据非线性相依关系是统计学领域中的基本问题. 在过去的一个多世纪里, 非线性相依数据关联分析方法和理论取得了一些重要进展. 本文讨论分布相依、均值相依和分位数相依 3 种不同关联关系; 按照一维、多维到高维的顺序, 介绍 3 种不同关联关系下所取得的部分重要研究进展; 最后讨论这些关联分析方法和理论在高维数据变量筛选和模型检验中的应用.

关键词 关联分析 相关系数 非线性 独立 条件独立

MSC (2020) 主题分类 62H20, 62M10

1 引言

统计学 “始于度量, 兴于相关”. 探求复杂数据关联关系是统计学领域的基本问题, 在生命健康、经济管理等许多领域都有非常重要的应用. 例如, 在基因组学研究中, 生物学家需要识别成千上万个基因位点上的基因表达是否与某种疾病存在关联[15,56,74]. 早在 19 世纪末, Pearson[60,62] 开创性地提出了矩相关系数, 来度量一维随机变量之间的线性相依关系, 并建立了矩相关系数与线性回归系数之间的联系. 针对分类数据, Pearson[61] 提出了列联表分析方法, 该方法度量一维随机变量之间的非线性相依关系. 这些原创思想和重要发现非常具有启发意义. 本文所指的非线性相依关系包括线性相依作为特殊情形. 事实上, 准确地描述或刻画非线性相依关系是非常困难的. 本文分别讨论分布相依、均值相依和分位数相依 3 种不同情形下的非线性相依关系. 即便如此, 任何一种情形下的非线性关系还是存在无穷多种可能性, 因而不得不去从反面的角度来分别定义分布独立、均值独立和分位数独立. 当

英文引用格式: Zhou Y Q, Xu K, Zhu L P. Association analysis for nonlinearly dependent data (in Chinese). Sci Sin Math, 2024, 54: 1169–1194, doi: 10.1360/SSM-2023-0175

我们说两个变量分布 (均值、分位数) 独立时, 意味着这两个变量在分布 (均值、分位数) 意义下不存在任何的非线性关联关系.

为了度量多维随机变量 $\boldsymbol{x} \in \mathbb{R}^p$ 和 $\boldsymbol{y} \in \mathbb{R}^q$ 之间的非线性关系, 文献 [30, 79] 拓展了 Pearson 相关系数, 分别提出了似然比检验和典型相关系数. 但是这两种方法在检验独立性时需要随机向量满足正态分布, 并且样本量要大于向量维数. 这样的强假设在实际应用中很有可能不成立. 事实上, 将一维非线性度量拓展到多维情形下非常困难, 直到距离相关系数 [75] 被提出, 才有了突破性的进展. 距离相关系数是将多维数据投影到一维空间, 利用特征函数度量独立性, 遍历所有可能的投影方向将结果汇总. 距离相关系数可以分析连续型、离散型或者混合型等多样化的数据类型, 其计算形式简洁, 数值为 0 表明随机向量相互独立. 文献 [23] 提出了基于 Hilbert-Schmidt 算子的独立性准则, 距离相关系数也可以等价地看成其选择一种特定核函数的结果. 这些特性使距离相关系数获得了不少学者的青睐, 被应用至变量筛选 [44]、条件独立性检验 [77]、独立主成分分析 [51]、交互效应识别 [36] 和相互独立检验 [89] 等领域. 但需要指出的是, 距离相关系数需要假设随机向量的矩存在. 基于距离相关系数构建的独立性检验, 其统计量的渐近分布依赖于随机向量本身的分布, 含有待估计的未知参数. 这也意味着检验临界值的确定需要采用自助抽样法或者随机置换法, 极大增加了独立性检验的计算量. 鉴于这些理论与计算上的困难, 不少新方法应运而出. 文献 [28] 基于距离的秩提出了度量准则, 文献 [105] 通过投影的方式从分布函数的角度提出了投影相关系数. 文献 [13, 68] 利用多元秩变换, 对距离相关系数进行推广. 文献 [54] 考虑使用基于 Euclid 距离排序的 τ^* 度量非线性关系. 文献 [69] 将多元秩变换应用至广义对称协方差, 使独立性检验统计量的极限分布与随机向量的概率分布无关. 值得注意的是, 这些方法的改进也同样需要付出一定的代价. 例如, 文献 [105] 虽然免除了矩条件的假设, 但却将统计量的计算复杂度从样本量的平方阶提升至样本量的三次方阶, 增加了计算上的困难. 此外, 刻画分布独立的方法还可以被拓展至均值 (分位数) 独立的框架下.

在数据维数发散的高维情形下, 可以借鉴一维或多维思路, 将高维数据投影到一维或多维情形下. 但是, 高维向量的低维投影往往是渐近正态的, 使得为多维数据而设计的度量准则在维数发散时性质发生巨大的改变. 基于投影的方法会简化成度量线性的相依关系. 例如, 在数据维数趋于无穷、样本量固定或以稍慢速度发散时, 文献 [104] 证明了样本距离协方差可以近似表示为各分量间的样本协方差之和. 这表明在固定维数下能够探测复杂非线性关系的距离相关系数, 在高维情形下只能度量线性相关性, 使得基于其构造的独立性检验功效急剧下降. 文献 [21] 发现在某一类特定的备择下, 当数据维数满足一定的条件时, 距离相关系数才能刻画非线性关系. 在均值 (分位数) 独立情形下, 相似的现象也同样存在. 鉴于此, 本文进一步介绍如何将固定维数下的度量方法推广到发散维数的大数据场景下.

除了关联关系, 因果关系也是统计学中关注的重要问题之一. 关联关系的存在并不意味着一定有因果关系. 文献 [52] 介绍了进行因果推断的统计方法. 条件独立性是进行因果推断的基础. 为了帮助挖掘因果关系, 本文介绍条件关联性的度量与检验方法, 即在控制混杂因素的前提下, 研究事物之间是否具有非线性关联关系. 由于条件关联性度量考虑了混杂因素的存在, 估计涉及非参数估计, 因此其理论和计算的难度更加升级. 不少学者尝试将条件独立与独立建立等价关系, 希望借助非线性关联的工具解决条件独立的问题, 来缓解维数祸根的影响.

本文还探讨关联性度量准则的两个应用: 高维数据变量筛选与模型检验. 在高维数据分析中, 常常需要剔除数据中的冗余变量来增强模型的可解释性. 变量筛选的主要思路是将各种类型的关联性度量作为边际的筛选准则, 选出对响应变量最有影响的协变量保留在模型中. 高维数据的模型检验则是利用关联度量的方法, 研究模型残差与协变量的关联性, 以此判断高维情形下的模型假设是否合理.

本文余下内容的结构如下. 第 2 节讨论多种关联关系的联系与区别. 第 3 节介绍一维随机变量的关联度量. 第 4 节介绍多维随机向量的关联度量. 第 5 节介绍高维随机向量的关联度量. 第 6 节讨论条件关联度量. 第 7 节进一步论述关联度量的应用. 第 8 节给出本文结论.

2 多种关联关系的联系与区别

2.1 分布独立

度量随机变量之间的非线性关系一般可以从刻画随机变量分布特征的密度函数、分布函数和特征函数出发, 分析其联合分布与边际分布乘积之间的差异. 令 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分别表示维数为 p 和 q 的随机向量, 它们之间的独立性可以用符号表示为 $\boldsymbol{x}\perp\!\!\!\perp\boldsymbol{y}$. 相互独立表明 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 之间不存在任何的非线性关系, 即 $\boldsymbol{x}$ 不会对 $\boldsymbol{y}$ 的分布特征产生任何的影响.

从分布函数出发, $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 相互独立当且仅当 $(\boldsymbol{x},\boldsymbol{y})$ 联合分布函数等于 $\boldsymbol{x}$ 的边际分布函数与 $\boldsymbol{y}$ 的边际分布函数乘积. 基于此, 文献 [8, 29] 分别提出了 Hoeffding 相关系数和 Blum-Kiefer-Rosenblatt (BKR) 相关系数用于检验一维随机变量的独立性. 文献 [103] 为了对同时存在离散型与连续型变量的高维数据进行边际效用的准确排序, 提出了修正的 BKR 相关系数. 文献 [33,105] 通过投影的方式, 分别将 Hoeffding 相关系数和 BKR 相关系数推广到多元的情形. 文献 [94] 从计算复杂度、原假设下的渐近分布、高维情形下的功效这 3 个方面对投影相关系数进行了改进. 文献 [78] 将分布函数的概念拓展至非 Euclid 空间中, 并提出了相互独立的检验方法.

从特征函数出发, $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 相互独立当且仅当 $(\boldsymbol{x},\boldsymbol{y})$ 联合特征函数等于 $\boldsymbol{x}$ 的边际特征函数与 $\boldsymbol{y}$ 的边际特征函数乘积. 基于此, 文献 [75] 选择了特定的权函数, 使得特征函数的积分具有显式表达式, 提出了距离相关系数. 不同权函数的选择会使得到的相关系数形式不同. 例如, 文献 [24] 选择了密度函数为权函数, 得到了稳定相关系数 (stable correlation). 文献 [96] 对于属性变量数据, 提出了半参数距离相关系数.

从密度函数出发, $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 相互独立当且仅当 $(\boldsymbol{x},\boldsymbol{y})$ 联合密度函数等于 $\boldsymbol{x}$ 的边际密度函数与 $\boldsymbol{y}$ 的边际密度函数乘积. 基于此, 文献 [5, 34, 63, 72] 对互信息进行了讨论. 为了估计互信息, 文献 [73] 将数据离散化, 文献 [55] 使用核密度估计, 文献 [5] 则是考虑基于 k 近邻的熵来构造有效的估计量. 但这些方法的估计效果在很大程度上依赖于一些调节参数, 而如何选择最佳调节参数以达到较好的假设检验效果在文献中鲜有提及.

2.2 均值独立与分位数独立

$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 之间均值独立用符号可以表示为 $\mathrm{E}(\boldsymbol{y}\mid\boldsymbol{x})=\mathrm{E}(\boldsymbol{y})$, 即 $\boldsymbol{y}$ 在给定 $\boldsymbol{x}$ 下的均值与 $\boldsymbol{x}$ 相互独立. 虽然随机变量之间的分布独立可以推导出均值独立成立, 但分布独立与均值独立并不完全等价, 两个互不独立的随机变量依然有可能均值独立.

为了度量均值独立, 文献 [66] 通过拓展度量分布独立的距离相关系数, 基于特征函数提出了鞅差相关系数. 在高维回归模型中, 文献 [91] 利用边际的鞅差相关系数聚合, 在不假设模型结构的前提下检验高维协变量对响应变量均值的作用. 文献 [39] 为函数型数据提供了条件均值独立的检验方法. 文献 [27,37] 基于核函数检验了条件均值独立性. 文献 [42] 基于对称的 Lévy 测度拓展了鞅差相关系数.

文献 [97] 发现对于一维随机变量 X 和 Y, 有

$$\mathrm{E}(Y\mid X)=\mathrm{E}(Y)\ \text{几乎处处成立}$$

$$\begin{aligned}&\Leftrightarrow \mathrm{E}(Y \mid X < x_0) = \mathrm{E}(Y) \text{ 对所有的 } x_0 \in \mathrm{supp}(X) \text{ 成立}\\&\Leftrightarrow \mathrm{cov}\{Y,\ I(X < x_0)\} = 0 \text{ 对所有的 } x_0 \in \mathrm{supp}(X) \text{ 成立}\\&\Leftrightarrow \mathrm{E}[\mathrm{cov}^2\{Y,\ I(X < \widetilde{X}) \mid \widetilde{X}\}] = 0,\end{aligned}$$

其中, $\mathrm{supp}(X)$ 表示 X 的支撑, $\widetilde{X}$ 为 X 的独立复制, $I(\cdot)$ 表示示性函数. 因此, 文献 [97] 将条件独立转化为度量 Y 与 X 示性函数之间的协方差是否为 0, 并提出了累积协方差. 在估计累积协方差时, 使用的是 X 的秩, 即利用了 X 分布函数的信息. 文献 [85] 通过投影平均在多元空间中推广了累积协方差. 文献 [43] 针对高维异质性数据, 对每一维数的累积协方差进行求和聚合, 检验高维协变量的均值效应.

为了研究分位数独立, 令 $Q_{\boldsymbol{y}}(\tau_{\boldsymbol{y}})$ 表示 $\boldsymbol{y}$ 的无条件 $\tau_{\boldsymbol{y}}$ 分位数, $Q_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\tau_{\boldsymbol{y}})$ 表示给定 $\boldsymbol{x}$ 时 $\boldsymbol{y}$ 的条件 $\tau_{\boldsymbol{y}}$ 分位数. 与均值独立类似, 条件分位数独立关心的是, 对于 $\tau_{\boldsymbol{y}} \in (0,1)$, $Q_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\tau_{\boldsymbol{y}}) = Q_{\boldsymbol{y}}(\tau_{\boldsymbol{y}})$ 是否成立. 若成立, 则表明在给定的 $\tau_{\boldsymbol{y}}$ 分位数水平下, $\boldsymbol{x}$ 对 $\boldsymbol{y}$ 的条件分位数没有影响. 文献 [40] 提出了分位数相关系数. 文献 [106] 提出了区间分位数的概念, 并建立了分位数独立与分布独立之间的联系. 同时, 条件分位数独立等价于 $\mathrm{E}(\boldsymbol{w} \mid \boldsymbol{x}) = \mathrm{E}(\boldsymbol{w})$, 其中

$$\boldsymbol{w} = \tau_{\boldsymbol{y}} - I\{\boldsymbol{y} - Q_{\boldsymbol{y}}(\tau_{\boldsymbol{y}}) \leqslant 0\}.$$

因此, 文献 [66, 91] 利用鞅差相关系数来度量与检验条件分位数独立.

2.3 两样本检验与对称性检验

假设 $p = q$, 检验 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 是否具有相同分布依然可以通过比较两总体的密度函数、分布函数和特征函数完成检验方法的构建. $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 具有相同分布用符号可以表示为 $\boldsymbol{x} \overset{\mathrm{D}}{=} \boldsymbol{y}$. 若将 $\boldsymbol{y}$ 记成与 $\boldsymbol{x}$ 对称的随机向量 $-\boldsymbol{x}$, 则两样本的检验方法也可以应用至对称性检验.

在一维情形下, Kolmogorov-Smirnov 检验 [70] 和 Cramér-von Mises 检验 [2, 64] 是两个经典的基于经验分布函数的方法. 由于相应的检验统计量仅与样本的秩有关, 所以两个检验统计量的极限零分布不依赖于任何冗余参数, 不需要任何矩条件. 但随着维数 p 的增加, 它们会遭遇维数祸根问题 [16]. 虽然利用投影方法去检验一样本的多元分布拟合优度问题在很早的时候就受到国内学者的关注 [12, 107], 但鲜有文献在两样本/高维两样本框架下考虑. 为了缓解维数祸根问题, 文献 [87] 利用投影推广了经典的 Cramér-von Mises 度量, 提出了能够适应于高维数据的两样本分布相等的非参数检验.

从密度函数的角度, 文献 [1] 通过比较两总体密度的差值构造了检验统计量. 文献 [98] 通过比较两总体密度的商值构造了检验统计量. 这些方法属于非参数光滑检验, 需要考虑密度函数估计中至关重要的窗宽或者节点数的选择等问题, 且无法在高维数据下直接使用.

从特征函数的角度, 文献 [3] 推荐势能检验统计量. 势能检验是一个很受欢迎的方法, 文献 [6] 基于文献 [25] 证明了两样本势能检验能够应用到高维数据中. 但遗憾的是, 势能检验需要总体分布满足某些矩条件因而对厚尾数据或者含异常值点的数据不够稳健.

从两总体分布最大均值偏差的角度, 文献 [22] 提出了基于正定核的检验方法, 但需要选择合理的正定核及窗宽. 在高维数据下, 最大均值偏差的有效性问题尚未得到解决.

从图结构的角度, 文献 [20] 基于最小生成树构造了检验统计量, 文献 [53] 基于最邻近法构造了检验统计量, 文献 [26] 基于组合样本的秩构造了检验统计量, 文献 [7] 则考虑了最短 Hamilton 路径. 尽管这些图结构方法很有用, 某些方法在高维数据情形下也可使用, 但它们需要选择一些调节参数, 如

最临近数、最优权重和最优路径等. 文献 [57] 在可分离的 Banach 空间上, 提出了利用球偏离系数构造了检验统计量. 基于球偏离系数的检验方法, 对于非平衡数据和尺度差异的备择结构非常有效.

3 一维随机变量的关联度量

3.1 Pearson 相关系数

对于一维随机变量 X 和 Y, 度量相关性最经典的方法是采用 Pearson 相关系数. 如果 $0<\mathrm{var}(X)<\infty$ 和 $0<\mathrm{var}(Y)<\infty$, 则 X 和 Y 不相关等价于 $\mathrm{cov}(X,Y)=0$. 但是, $\mathrm{cov}(X,Y)$ 会受到 X 和 Y 的方差的影响. 利用 Cauchy 不等式, 可知

$$\mathrm{cov}(X,Y)\leqslant\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}.$$

因此, 对 X 和 Y 标准化, Pearson 相关系数的定义为

$$\mathrm{corr}(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}=\frac{\mathrm{E}[\{X-\mathrm{E}(X)\}\{Y-\mathrm{E}(Y)\}]}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}.$$

当 $\mathrm{corr}(X,Y)=1$ 时, 易证

$$\mathrm{P}\left\{Y=\sqrt{\frac{\mathrm{var}(Y)}{\mathrm{var}(X)}}X+\mathrm{E}(Y)-\sqrt{\frac{\mathrm{var}(Y)}{\mathrm{var}(X)}}\mathrm{E}(X)\right\}=1.$$

同理, 若 $\mathrm{corr}(X,Y)=-1$, 可得

$$\mathrm{P}\left\{Y=-\sqrt{\frac{\mathrm{var}(Y)}{\mathrm{var}(X)}}X+\mathrm{E}(Y)+\sqrt{\frac{\mathrm{var}(Y)}{\mathrm{var}(X)}}\mathrm{E}(X)\right\}=1.$$

因此, Pearson 相关系数的绝对值为 1 时, 表示随机变量 X 和 Y 完全线性相关. Pearson 相关系数在 R 语言中可以用 stats::cor ( ) 计算. 在单变量线性回归中, 响应变量与解释变量 Pearson 相关系数的平方反映了解释变量对响应变量方差的解释比例. 但是, Pearson 相关系数无法度量随机变量之间的非线性关系, 其等于 0 仅说明随机变量之间不相关, 并且在应用时容易受到样本观测异常值的影响.

3.2 Kendall τ 和 Spearman ρ 秩相关系数

作为 Pearson 相关系数的推广, Kendall τ 秩相关系数 [32] 和 Spearman ρ 秩相关系数 [71] 能够对单调变换后的随机变量保持度量的数值不变, 对样本观测的异常值保持稳健. Kendall τ 秩相关系数的定义为

$$\tau=\mathrm{E}\{\mathrm{sgn}(X_1-X_2)\mathrm{sgn}(Y_1-Y_2)\},$$

其中 $\mathrm{sgn}(\cdot)$ 表示符号函数. Spearman ρ 秩相关系数的定义为

$$\rho_S=\mathrm{corr}\{R(X),R(Y)\}=\frac{\mathrm{cov}\{R(X),R(Y)\}}{\sqrt{\mathrm{var}\{R(X)\}\mathrm{var}\{R(Y)\}}},$$

其中 $R(\cdot)$ 表示对随机变量进行秩变换. Kendall τ 和 Spearman ρ 秩相关系数在 R 语言中也可以用 stats::cor ( ) 计算. 当 X 和 Y 独立时, 易知 $\tau=\rho_S=0$. 但是, 反之不然. 并且, Kendall τ 和 Spearman ρ 秩相关系数只能检测随机变量之间单调的相依关系. 以 Spearman ρ 为例, 易证如果 $\rho_S=1$, 则有

$$\mathrm{P}(Y=F_Y^{-1}\{F_X(X)\})=1,$$

其中 $F_X(\cdot)$ 和 $F_Y(\cdot)$ 分别表示 X 和 Y 的累积分布函数. 类似地, 如果 $\rho_S=-1$, 则有

$$\mathrm{P}(Y=F_Y^{-1}\{1-F_X(X)\})=1.$$

针对二元联合正态随机向量 (X,Y), 若其相关系数为 ρ, 文献 [31] 论述了

$$\tau=\frac{2}{\pi}\arcsin(\rho),\quad \rho_S=\frac{6}{\pi}\arcsin\left(\frac{\rho}{2}\right).$$

3.3 分布独立

3.3.1 Hoeffding 相关系数和 BKR 相关系数

为了度量各种类型的非线性关系, 许多独立性度量方法应运而生. Hoeffding 相关系数 [29] 和 BKR 相关系数 [8] 均基于联合分布函数和边际分布函数乘积之间的差异来判断两个一维随机变量之间是否存在非线性关系. Hoeffding 相关系数与 BKR 相关系数的定义分别为

$$\mathrm{H}(X,Y)=\int_{\mathbb{R}^2}\{F_{X,Y}(x,y)-F_X(x)F_Y(y)\}^2dF_{X,Y}(x,y)$$

以及

$$\mathrm{BKR}(X,Y)=\int_{\mathbb{R}^1}\int_{\mathbb{R}^1}\{F_{X,Y}(x,y)-F_X(x)F_Y(y)\}^2dF_X(x)dF_Y(y),$$

其中 $F_{X,Y}(\cdot,\cdot)$ 表示 X 和 Y 的联合分布函数. Hoeffding 相关系数与 BKR 相关系数具有非常相似的形式. 它们均是非负的相关系数, 并且在两个随机变量独立时等于 0. 但是 Hoeffding 相关系数等于 0 时, 并不能说明两个随机变量相互独立. 例如, 在 $\mathrm{P}(X=0,Y=1)=\mathrm{P}(X=1,Y=0)=1/2$ 时, $\mathrm{H}(X,Y)=0$ 但 X 和 Y 并不相互独立 [29]. 相比之下, $\mathrm{BKR}(X,Y)=0$ 当且仅当随机变量 X 和 Y 相互独立.

对于二元联合正态随机向量 (X,Y), 若其相关系数为 ρ, 文献 [101, 定理 5] 证明了

$$\begin{aligned}\mathrm{H}(X,Y)=(4\pi^2)^{-1}&\left[(\arcsin\rho)^2-\left\{\arcsin\left(\frac{\rho}{2}\right)\right\}^2\right]\\&+\pi^{-2}\left[\int_0^{\arcsin\frac{\rho}{2}}\arcsin\left(\frac{\sin x}{2\cos 2x+1}\right)dx\right.\\&\left.+\int_0^{\arcsin\frac{\rho}{2}}\arcsin\left\{\left(\frac{2\cos 2x-1}{6\cos 2x+3}\right)^{1/2}\sin x\right\}dx\right]\\&-(2\pi^2)^{-1}\left\{\int_0^{\arcsin\rho}\arcsin\left(\frac{\sin x}{3}\right)dx\right.\\&\left.+\int_0^{\arcsin\frac{\rho}{2}}\arcsin\left(\frac{2\cos 2x+3}{2\cos 2x+1}\sin x\right)dx\right\},\end{aligned}$$

$$\mathrm{BKR}(X,Y)=(2\pi^2)^{-1}\Big\{\int_0^{\arcsin\frac{\rho}{2}}\arcsin\left(\frac{2\cos 2x+3}{2\cos 2x+1}\sin x\right)dx - \int_0^{\arcsin\frac{\rho}{2}}\arcsin\left(\frac{\sin x}{2\cos 2x+1}\right)dx\Big\}$$

以及

$$\inf_{\rho\neq 0}\frac{\mathrm{H}(X,Y)}{\rho^2}=(12\pi^2)^{-1}\leqslant\frac{\mathrm{H}(X,Y)}{\rho^2}\leqslant\sup_{\rho\neq 0}\frac{\mathrm{H}(X,Y)}{\rho^2}=\frac{1}{30},$$
$$\inf_{\rho\neq 0}\frac{\mathrm{BKR}(X,Y)}{\rho^2}=(12\pi^2)^{-1}\leqslant\frac{\mathrm{BKR}(X,Y)}{\rho^2}\leqslant\sup_{\rho\neq 0}\frac{\mathrm{BKR}(X,Y)}{\rho^2}=\frac{1}{90}.$$

3.3.2 τ^* 相关系数

由于 Kendall τ 秩相关系数仅能检测随机变量之间单调的相依关系, 文献 [4] 对原有的 Kendall τ 秩相关系数进行拓展, 提出了可以度量独立性的 τ^* 相关系数, 其定义为

$$\tau^*(X,Y)=\mathrm{E}\{a(X_1,X_2,X_3,X_4)a(Y_1,Y_2,Y_3,Y_4)\},$$

其中

$$a(z_1,z_2,z_3,z_4)=\mathrm{sgn}(|z_1-z_2|+|z_3-z_4|-|z_1-z_3|-|z_2-z_4|).$$

对于任意随机变量 X 和 Y, 始终有 $\tau^*(X,Y)\geqslant 0$, 其等于 0 当且仅当 X 和 Y 相互独立. τ^* 相关系数的计算可以用 R 语言中 TauStar 包完成.

针对二元联合正态随机向量 (X,Y), 若其相关系数为 ρ, 文献 [101, 定理 5] 论证了

$$\begin{aligned}\tau^*(X,Y)=3\pi^{-2}&\left[(\arcsin\rho)^2-\left\{\arcsin\left(\frac{\rho}{2}\right)\right\}^2\right]\\&+12\pi^{-2}\int_0^{\arcsin\frac{\rho}{2}}\arcsin\left\{\left(\frac{2\cos 2x-1}{6\cos 2x+3}\right)^{1/2}\sin x\right\}dx\\&-6\pi^{-2}\Big\{\int_0^{\arcsin\rho}\arcsin\left(\frac{\sin x}{3}\right)dx\\&-\int_0^{\arcsin\frac{\rho}{2}}\arcsin\left(\frac{2\cos 2x+3}{2\cos 2x+1}\sin x\right)dx\Big\}\end{aligned}$$

以及

$$\inf_{\rho\neq 0}\frac{\tau^*(X,Y)}{\rho^2}=3\pi^{-2}\leqslant\frac{\tau^*(X,Y)}{\rho^2}\leqslant\sup_{\rho\neq 0}\frac{\tau^*(X,Y)}{\rho^2}=\frac{2}{3}.$$

3.3.3 修正的 BKR 相关系数

为了给同时存在离散型与连续型随机变量的高维数据提供有效的特征筛选准则, 文献 [103] 提出了修正的 BKR (modified Blum-Kiefer-Rosenblatt, MBKR) 相关系数, 其定义为

$$\mathrm{MBKR}(X,Y)=\int_{\mathbb{R}^1}\int_{\mathbb{R}^1}\frac{\{F_{X,Y}(x,y)-F_X(x)F_Y(y)\}^2}{F_X(x)\{1-F_X(x)\}F_Y(y)\{1-F_Y(y)\}}dF_X(x)dF_Y(y).$$

MBKR 相关系数继承了 BKR 相关系数在度量独立性上的一些良好的性质. 例如, $\mathrm{MBKR}(X,Y)=0$ 当且仅当随机变量 X 和 Y 相互独立. 它与原本的 BKR 相关系数差别体现在: MBKR 是对 $\mathrm{corr}^2\{I(X$

$\leqslant x), I(Y \leqslant y)\}$ 进行积分, 而原本的 BKR 相关系数采用的是 $\mathrm{cov}^2\{I(X \leqslant x), I(Y \leqslant y)\}$. 如果 X 和 Y 均是连续型随机变量, 则采用 Pearson 相关系数或是协方差并不会产生本质的差别. 因为 $F_X(X)$ 和 $F_Y(Y)$ 都服从 $[0,1]$ 上的均匀分布. 而当随机变量中出现一部分离散型或分类型随机变量时, 采用 Pearson 相关系数与协方差产生的差异就不可忽视了. 文献 [103] 通过数值模拟说明了两者在高维特征筛选中的不同表现. 计算 MBKR 相关系数的代码在 github.com/Yeqing-TJ 上.

3.3.4 切片独立性度量

定义 $s(t;X) = \mathrm{pr}(Y \geqslant t \mid X)$. 一维随机变量 T 的概率密度函数与累积分布函数分别为 $f(t)$ 和 $F_T(t)$. T 的支撑集记作 $\mathrm{supp}(T) = \{t : f(t) > 0\}$. 假设 $\mathrm{supp}(Y) \subseteq \mathrm{supp}(T)$. 因此, X 和 Y 是相互独立的当且仅当 $\mathrm{var}\{s(t;X)\} = 0$ 对于所有 $t \in \mathbb{R}^1$ 成立.

文献 [10,14,35,92] 分别提出用下列度量来衡量变量间独立性的程度:

$$\mathcal{S}(X,Y) = \frac{\int \mathrm{var}\{s(t;X)\}dF_T(t)}{\int \mathrm{var}\{I(Y \geqslant t)\}dF_T(t)}.$$

上式中的分母部分保证了 $\mathcal{S}(X,Y)$ 的取值能够在 0 到 1 之间. 通过方差分解可以得到

$$\mathcal{S}(X,Y) = 1 - \frac{\int \mathrm{E}[\mathrm{var}\{I(Y \geqslant t) \mid X\}]dF_T(t)}{\int \mathrm{var}\{I(Y \geqslant t)\}dF_T(t)}.$$

文献 [10,14,35] 假设 T 是 Y 的一个独立复制, 而文献 [92] 允许 T 是一个满足 $\mathrm{supp}(Y) \subseteq \mathrm{supp}(T)$ 条件的任意随机变量.

文献 [35, 引理 1] 和 [10, 定理 1] 给出了关于 $\mathcal{S}(X,Y)$ 的性质. $\mathcal{S}(X,Y) = 0$ 当且仅当 X 和 Y 相互独立, $\mathcal{S}(X,Y) = 1$ 当且仅当 Y 是 X 的某个确定函数. 若 (X,Y) 是一个二元正态随机变量, 相关系数为 ρ, 则 $\mathcal{S}(X,Y)$ 随着 $|\rho|$ 的增大严格单调递增. 如果对 X 和 Y 作严格单调的变换, 则 $\mathcal{S}(X,Y)$ 的数值仍保持不变.

给定一组样本 $\{(X_i,Y_i), i = 1,\ldots,n\}$ 来估计 $\mathcal{S}(X,Y)$. 第一类方法是采用核光滑的方法. 文献 [14,35] 提出了用核光滑来估计给定的 t 时 $\mathrm{var}\{I(Y \geqslant t) \mid X\}$ 的值. 通过核光滑来估计 $\mathcal{S}(X,Y)$ 的复杂度约为 $O(n^2)$, 并且估计量在原假设下的渐近分布依赖于核函数. 文献 [10] 提出了 $\mathcal{S}(X,Y)$ 的秩估计, 主要是根据 X_i 的值对 $\{(X_i,Y_i), i = 1,\ldots,n\}$ 进行排序, 排序后的样本记作 $\{(X_{(i)},Y_{(i)}), i = 1,\ldots,n\}$, 其中, $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 是 X_i 的次序统计量, $Y_{(i)}$ 是与 $X_{(i)}$ 相对应的样本. 若 r_i 是 $Y_{(i)}$ 的秩, 则文献 [10] 提出的估计为

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

由于 $\xi_n(X,Y)$ 的定义是基于秩, 其计算复杂度为 $O(n\log n)$. 在 X 和 Y 相互独立的情形下, $\xi_n(X,Y)$ 渐近服从正态分布. 计算 $\xi_n(X,Y)$ 可以用 R 包 XICOR 实现. 文献 [92] 介绍了一种估计 $\mathcal{S}(X,Y)$ 的切片方法. 将排序后的样本根据 $X_{(i)}$ 的取值分成 H 个切片, 使得每个切片内包含 c 个样本. 为了方便, 假设 $n = Hc$. 记 $X_{(h,j)} = X_{(c(h-1)+j)}$, $Y_{(h,j)} = Y_{(c(h-1)+j)}$, 其中 $j = 1,\ldots,c$, $h = 1,\ldots,H$. 在第 h 个切片内的观测样本为 $\{(X_{(h,j)},Y_{(h,j)}), j = 1,\ldots,c\}$. 给定 t, 在每个切片内估计 $\mathrm{var}\{I(Y \geqslant t) \mid X\}$, 之后再对所有的切片进行平均.

此外, 文献 [11] 介绍了关于该度量的最新研究进展. 文献 [67] 从构造独立性检验的相合性、计算效率和统计推断效率 3 个角度对 Hoeffding 相关系数、BKR 相关系数、τ^* 相关系数和 Chatterjee ξ 进

行了详细的比较. 文献 [46] 探讨了如何提升 Chatterjee ξ 的检验功效. 文献 [45] 建立了 Chatterjee ξ 的中心极限定理, 文献 [47] 探讨了标准的自助抽样法不能用于 Chatterjee ξ 的推断.

3.4 均值独立与分位数独立

3.4.1 累积散度

除了不相关与独立性, 我们在均值回归中还关心一种关系—均值独立, 即研究 $\mathrm{E}(Y \mid X) = \mathrm{E}(Y)$ 是否成立. 文献 [97] 发现 $\mathrm{E}(Y \mid X) = \mathrm{E}(Y)$ 几乎处处成立与 $\mathrm{E}[\mathrm{cov}^2\{Y, I(X \leqslant \tilde{X}) \mid \tilde{X}\}] = 0$ 等价. 基于此, 文献 [97] 提出了累积协方差和累积散度, 分别为

$$\mathrm{CCov}(Y \mid X) = \mathrm{E}[\mathrm{cov}^2\{Y,\ I(X < \widetilde{X}) \mid \widetilde{X}\}]$$

和

$$\mathrm{CD}(Y \mid X) = \frac{\mathrm{CCov}(Y \mid X)}{\mathrm{var}(Y)}.$$

从上述定义可以看出, $\mathrm{CD}(Y \mid X)$ 允许 X 的方差为无穷, 因此对厚尾分布的 X 依然能够保持稳健. $\mathrm{CD}(Y \mid X)$ 关于 X 和 Y 并不对称, 其等于 0 当且仅当 $\mathrm{E}(Y \mid X) = \mathrm{E}(Y)$ 成立. 对于 $a, b \in \mathbb{R}^1$ 且 $a \neq 0$, 对于任意严格单调变换 $M(X)$, 总有 $\mathrm{CD}(Y \mid X) = \mathrm{CD}\{aY + b \mid M(X)\}$.

针对二元联合正态随机向量 (X, Y), 若其相关系数为 ρ, 文献 [97, 定理 1] 给出了

$$\mathrm{CD}(Y \mid X) = \mathrm{CD}(X \mid Y) = \frac{\rho^2}{2\sqrt{3}\pi}.$$

3.4.2 分位数相关系数

在均值回归模型之外, 分位数回归也受到了广泛的关注. 文献 [40] 指出了分位数回归模型受欢迎的两个原因: 一是对于非 Gauss 或者厚尾数据表现十分稳健, 二是分位数回归模型通过不同的分位数可以获得更容易解释的回归估计. 与均值独立类似, 分位数独立关心的是 $Q_{Y \mid X}(\tau_Y) = Q_Y(\tau_Y)$ 是否成立. 文献 [40] 发现 $Q_{Y|X}(\tau_Y) = Q_Y(\tau_Y)$ 当且仅当 $I\{Y - Q_Y(\tau_Y) > 0\}$ 与 X 相互独立. 受到该发现的启发, 文献 [40] 提出了分位数协方差与分位数相关系数. 对于 $\tau_Y \in (0, 1)$, 有

$$\mathrm{qcov}_{\tau_Y}(Y, X) = \mathrm{cov}\{I(Y - Q_Y(\tau_Y) > 0), X\}$$

和

$$\mathrm{qcorr}_{\tau_Y}(Y, X) = \frac{\mathrm{qcov}_{\tau_Y}(Y, X)}{\{\tau_Y(1 - \tau_Y)\mathrm{var}(X)\}^{1/2}}.$$

由于分位数相关系数基于协方差来构造, 而协方差仅能判断线性相关关系, 因此在应用时分位数相关系数可能无法准确地识别非线性的分位数相依关系.

文献 [40, 引理 1] 建立了分位数相关系数与分位数回归系数之间的关系. 考虑分位数线性回模型

$$(a_0, b_0) = \arg\min_{a,b} \mathrm{E}[\rho_{\tau_Y}(Y - a - bX)],$$

其中 $\rho_{\tau_Y}(\omega) = w[\tau_Y - I(\omega < 0)]$. 对于 $\mathrm{E}(X^2) < \infty$, 有 $\mathrm{qcov}_{\tau_Y}(Y, X) = \varrho(b_0)$, 其中 $\varrho(\cdot)$ 是一个连续且递增的函数, $\varrho(b) = 0$ 当且仅当 $b = 0$.

3.4.3 区间分位数系数

分位数相关系数研究的是对于给定的单个分位数 τ_Y, 是否有 $Q_{Y|X}(\tau_Y) = Q_Y(\tau_Y)$ 成立, 但其只能识别线性的分位数相依关系. 为了将其进一步拓展, 文献 [106] 考虑了检验 $Q_{Y\,|\,X=Q_X(\tau_X)}(\tau_Y) = Q_Y(\tau_Y)$ 对于 $(\tau_Y,\tau_X) \in \mathcal{I}_Y \otimes \mathcal{I}_X \subseteq (0,1)\otimes(0,1)$ 是否成立. 文献 [106] 定义了如下的区间分位数系数

$$q(Y,X,\mathcal{I}_Y,\mathcal{I}_X) = \int_{\mathcal{I}_Y}\int_{\mathcal{I}_X} \frac{\mathrm{cov}^2\{I(Y\leqslant Q_Y(\tau_Y)), I(X\leqslant Q_X(\tau_X))\}}{\tau_Y(1-\tau_Y)\tau_X(1-\tau_X)} d\mu_1(\tau_Y)d\mu_2(\tau_X).$$

从定义可以看出, 如果要检验传统的分位数相依关系 $Q_{Y\,|\,X}(\tau_Y) = Q_Y(\tau_Y)$, 只需要令 $\mathcal{I}_Y = \{\tau_Y\}$ 和 $\mathcal{I}_X = (0,1)$. 引入区间分位数系数这一概念, 拓展了传统的独立性基本概念.

文献 [106, 命题 1] 证明了, 如果 $Q_{Y\,|\,X=Q_X(\tau_X)}(\tau_Y)$ 是唯一的, 则 $q(Y,X,\mathcal{I}_Y,\mathcal{I}_X) = 0$ 当且仅当 $Q_{Y\,|\,X=Q_X(\tau_X)}(\tau_Y) = Q_Y(\tau_Y)$ 对于 $(\tau_Y,\tau_X)\in\mathcal{I}_Y\otimes\mathcal{I}_X$ 成立. 此外, $q\{Y,X,(0,1),(0,1)\} = 0$ 当且仅当 Y 和 X 相互独立. 若对 X 和 Y 进行单调递增的变换, 分位数相关系数的值保持不变.

在应用时, 可以选择在任意感兴趣的分位数区间上, 度量随机变量之间的分位数相依关系. 模型 $Y = \exp(X^2)+\varepsilon$, 其中 X 和 ε 均服从标准正态分布. 从模型的结构可以看出, 并不是所有 Y 的分位数都依赖于 X. $Q_{Y\,|\,X=Q_X(\tau_X)}(\tau_Y)$ 依赖于 X 仅在区间 $\tau_Y\in\mathcal{I}_Y = (0,0.5)\cup(0.5,1)$ 上. 由此可见, 在数据分析时选择合适的分位数区间, 有助于刻画 X 与 Y 之间的关系.

4 多维随机向量的关联度量

4.1 分布独立

4.1.1 多维 Kendall τ 相关系数

Kendall τ 和 τ^* 相关系数均基于秩来构造, 尽管它们在一维框架下具有很多优良的性质, 如稳健性和渐近分布不依赖随机变量的分布等. 但在多维数据框架下, 基于向量的排序会使其遭遇 “维数祸根” 问题. 文献 [54] 推荐基于 Euclid 距离排序的 τ^*, 用符号 IPR-τ^* 表示, 其定义为

$$\begin{aligned}
\text{IPR-}\tau^*(\boldsymbol{x},\boldsymbol{y}) &= \mathrm{E}\{a(\|\boldsymbol{x}_0-\boldsymbol{x}_1\|,\|\boldsymbol{x}_0-\boldsymbol{x}_2\|,\|\boldsymbol{x}_0-\boldsymbol{x}_3\|,\|\boldsymbol{x}_0-\boldsymbol{x}_4\|)\\
&\quad\times a(\|\boldsymbol{y}_0-\boldsymbol{y}_1\|,\|\boldsymbol{y}_0-\boldsymbol{y}_2\|,\|\boldsymbol{y}_0-\boldsymbol{y}_3\|,\|\boldsymbol{y}_0-\boldsymbol{y}_4\|)\}\\
&\leqslant \mathrm{E}^{1/2}\{a(\|\boldsymbol{x}_0-\boldsymbol{x}_1\|,\|\boldsymbol{x}_0-\boldsymbol{x}_2\|,\|\boldsymbol{x}_0-\boldsymbol{x}_3\|,\|\boldsymbol{x}_0-\boldsymbol{x}_4\|)\}^2\\
&\quad\times \mathrm{E}^{1/2}\{a(\|\boldsymbol{y}_0-\boldsymbol{y}_1\|,\|\boldsymbol{y}_0-\boldsymbol{y}_2\|,\|\boldsymbol{y}_0-\boldsymbol{y}_3\|,\|\boldsymbol{y}_0-\boldsymbol{y}_4\|)\}^2\\
&= \{\text{IPR-}\tau^*(\boldsymbol{x},\boldsymbol{x})\}^{1/2}\{\text{IPR-}\tau^*(\boldsymbol{y},\boldsymbol{y})\}^{1/2},
\end{aligned}$$

其中函数 $a(z_1,z_2,z_3,z_4)$ 定义在第 3.3.2 小节中. 与原本的 τ^* 相比, IPR-$\tau^*(\boldsymbol{x},\boldsymbol{y})$ 采用了 $(\boldsymbol{x},\boldsymbol{y})$ 的 5 个简单随机样本, 比 τ^* 多需要一重样本.

文献 [54] 证明了 IPR-$\tau^*(\boldsymbol{x},\boldsymbol{y})$ 总是大于等于 0, 其等于 0 当且仅当 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 独立. 多维 Kendall τ 相关系数可相应地定义为

$$\text{IPR-}\tau_b^*(\boldsymbol{x},\boldsymbol{y}) = \frac{\text{IPR-}\tau^*(\boldsymbol{x},\boldsymbol{y})}{\sqrt{\text{IPR-}\tau^*(\boldsymbol{x},\boldsymbol{x})\text{IPR-}\tau^*(\boldsymbol{y},\boldsymbol{y})}}.$$

4.1.2 距离相关系数

文献 [75] 采用加权 L_2 范数度量了联合特征函数与边际特征函数乘积之间的距离, 提出了距离协方差, 其定义为

$$\mathrm{dCov}^2(\boldsymbol{x},\boldsymbol{y})=\int_{\mathbb{R}^{p+q}}\frac{\|\mathrm{E}\exp(\mathrm{i}\langle\boldsymbol{t},\boldsymbol{x}\rangle+\mathrm{i}\langle\boldsymbol{s},\boldsymbol{y}\rangle)-\mathrm{E}\exp(\mathrm{i}\langle\boldsymbol{t},\boldsymbol{x}\rangle)\mathrm{E}\exp(\mathrm{i}\langle\boldsymbol{s},\boldsymbol{y}\rangle)\|^2}{c_pc_q\|\boldsymbol{t}\|^{1+p}\|\boldsymbol{s}\|^{1+q}}d\boldsymbol{t}d\boldsymbol{s},$$

其中, $c_p=\pi^{(1+p)/2}/\Gamma\{(1+p)/2\}$, $c_q=\pi^{(1+q)/2}/\Gamma\{(1+q)/2\}$, $\Gamma\{\cdot\}$ 表示 Gamma 函数, $\langle\cdot,\cdot\rangle$ 表示内积, $\|\cdot\|$ 表示 L_2 范数. 从定义可以看出, $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 独立当且仅当 $\mathrm{dCov}^2(\boldsymbol{x},\boldsymbol{y})=0$. 因此, 距离协方差可以完全度量与检验独立性, 而且并不需要 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的维数相等. 由于距离协方差是从特征函数的角度出发, 所以其需要假设随机向量的矩是有限的. 随着数据维数的增高, 在实际应用时, 矩条件可能很难被满足.

上述定义中的特定权函数使得距离协方差具有简洁的显式表达式:

$$\begin{aligned}\mathrm{dCov}^2(\boldsymbol{x},\boldsymbol{y})&=\mathrm{E}(\|\boldsymbol{x}_1-\boldsymbol{x}_2\|\|\boldsymbol{y}_1-\boldsymbol{y}_2\|)+\mathrm{E}(\|\boldsymbol{x}_1-\boldsymbol{x}_2\|)\mathrm{E}(\|\boldsymbol{y}_1-\boldsymbol{y}_2\|)\\&\quad-2\mathrm{E}(\|\boldsymbol{x}_1-\boldsymbol{x}_2\|\|\boldsymbol{y}_1-\boldsymbol{y}_3\|)\\&=4^{-1}\mathrm{E}\{(\|\boldsymbol{x}_1-\boldsymbol{x}_2\|+\|\boldsymbol{x}_3-\boldsymbol{x}_4\|-\|\boldsymbol{x}_1-\boldsymbol{x}_3\|-\|\boldsymbol{x}_2-\boldsymbol{x}_4\|)\\&\quad\times(\|\boldsymbol{y}_1-\boldsymbol{y}_2\|+\|\boldsymbol{y}_3-\boldsymbol{y}_4\|-\|\boldsymbol{y}_1-\boldsymbol{y}_3\|-\|\boldsymbol{y}_2-\boldsymbol{y}_4\|)\}\\&\leqslant 4^{-1}\mathrm{E}^{1/2}\{(\|\boldsymbol{x}_1-\boldsymbol{x}_2\|+\|\boldsymbol{x}_3-\boldsymbol{x}_4\|-\|\boldsymbol{x}_1-\boldsymbol{x}_3\|-\|\boldsymbol{x}_2-\boldsymbol{x}_4\|)^2\}\\&\quad\times\mathrm{E}^{1/2}\{(\|\boldsymbol{y}_1-\boldsymbol{y}_2\|+\|\boldsymbol{y}_3-\boldsymbol{y}_4\|-\|\boldsymbol{y}_1-\boldsymbol{y}_3\|-\|\boldsymbol{y}_2-\boldsymbol{y}_4\|)^2\}\\&=\mathrm{dCov}(\boldsymbol{x},\boldsymbol{x})\mathrm{dCov}(\boldsymbol{y},\boldsymbol{y}),\end{aligned}$$

其中不等式根据 Cauchy 不等式推导出. 故距离相关系数的定义为

$$\mathrm{dCor}^2(\boldsymbol{x},\boldsymbol{y})=\frac{\mathrm{dCov}^2(\boldsymbol{x},\boldsymbol{y})}{\mathrm{dCov}(\boldsymbol{x},\boldsymbol{x})\mathrm{dCov}(\boldsymbol{y},\boldsymbol{y})}.$$

对于二元联合正态随机向量 (X,Y), 若其相关系数为 ρ, 文献 [75] 证明了

$$\mathrm{dCor}^2(\boldsymbol{x},\boldsymbol{y})=\frac{\rho\arcsin(\rho)+\sqrt{1-\rho^2}-\rho\arcsin(\rho/2)-\sqrt{4-\rho^2}+1}{1+\pi/3-\sqrt{3}}.$$

距离相关系数在 R 语言中可以用 energy::dcor ( ) 计算. 文献 [23] 在再生核 Hilbert 空间中提出了 Hilbert-Schmidt 独立性准则. 文献 [65] 证明了存在核函数使得距离相关系数与 Hilbert-Schmidt 独立性准则等价. Hilbert-Schmidt 独立性准则在 R 语言中可以用 dHSIC::dhsic ( ) 计算.

4.1.3 HHG 相关系数和球相关系数

令 $\overline{B}_\rho(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 表示以 $\boldsymbol{x}_1$ 为中心、$\|\boldsymbol{x}_1-\boldsymbol{x}_2\|$ 为半径的球. 定义 $\delta^{\boldsymbol{x}}_{ij,k}=I\{\boldsymbol{x}_k\in\overline{B}_\rho(\boldsymbol{x}_i,\boldsymbol{x}_j)\}$, $\delta^{\boldsymbol{x}}_{ij,kl}=\delta^{\boldsymbol{x}}_{ij,k}\delta^{\boldsymbol{x}}_{ij,l}$ 以及 $\delta^{\boldsymbol{x}}_{ij,klst}=(\delta^{\boldsymbol{x}}_{ij,kl}+\delta^{\boldsymbol{x}}_{ij,st}-\delta^{\boldsymbol{x}}_{ij,ks}-\delta^{\boldsymbol{x}}_{ij,lt})/2$. 类似地, 针对 $\boldsymbol{y}$, 定义记号 $\delta^{\boldsymbol{y}}_{ij,k}$、$\delta^{\boldsymbol{y}}_{ij,kl}$ 和 $\delta^{\boldsymbol{y}}_{ij,klst}$. 引入权重函数 $\omega_1(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 和 $\omega_2(\boldsymbol{y}_1,\boldsymbol{y}_2)$. 令 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ $(i=1,\ldots,6)$ 是来自总体 $(\boldsymbol{x},\boldsymbol{y})$ 的简单随机样本, 文献 [59] 定义加权球协方差度量为

$$\mathrm{BCov}^2_\omega(\boldsymbol{x},\boldsymbol{y})=\mathrm{E}\{\delta^{\boldsymbol{x}}_{12,3456}\delta^{\boldsymbol{y}}_{12,3456}\omega_1(\boldsymbol{x}_1,\boldsymbol{x}_2)\omega_2(\boldsymbol{y}_1,\boldsymbol{y}_2)\}.$$

利用 Cauchy 不等式和标准化 $\mathrm{BCov}^2_\omega(\boldsymbol{x},\boldsymbol{y})$, 相应的加权球相关系数定义为

$$\mathrm{BCor}^2_\omega(\boldsymbol{x},\boldsymbol{y})=\frac{\mathrm{BCov}^2_\omega(\boldsymbol{x},\boldsymbol{y})}{\mathrm{BCov}_\omega(\boldsymbol{x},\boldsymbol{x})\mathrm{BCov}_\omega(\boldsymbol{y},\boldsymbol{y})}.$$

球相关系数在 R 语言中可以用 Ball::bcor ( ) 计算.

如果令权函数

$$\omega_1(\boldsymbol{x}_1,\boldsymbol{x}_2)=[\mathrm{E}(\delta^{\boldsymbol{x}}_{12,3}\mid\boldsymbol{x}_3)\{1-\mathrm{E}(\delta^{\boldsymbol{x}}_{12,3}\mid\boldsymbol{x}_3)\}]^{-1},$$
$$\omega_2(\boldsymbol{y}_1,\boldsymbol{y}_2)=[\mathrm{E}(\delta^{\boldsymbol{y}}_{12,3}\mid\boldsymbol{y}_3)\{1-\mathrm{E}(\delta^{\boldsymbol{y}}_{12,3}\mid\boldsymbol{y}_3)\}]^{-1},$$

则球相关系数等价于文献 [28] 中的 Heller-Heller-Gorfine (HHG) 相关系数. 基于 HHG 相关系数的独立性检验在 R 语言中可以用 HHG::hhg.test ( ) 计算. 特别地, 文献 [59] 证明了 $\mathrm{BCov}^2_\omega(\boldsymbol{x},\boldsymbol{y})$ 属于 Hoeffding 型相依性度量, 并且能够用随机积分等价表示为

$$\begin{aligned}\mathrm{BCov}^2_\omega(\boldsymbol{x},\boldsymbol{y})=\int(\theta-\mu\otimes\nu)^2\{\overline{B}_\rho(\boldsymbol{x}_1,\boldsymbol{x}_2)\times\overline{B}_\zeta(\boldsymbol{y}_1,\boldsymbol{y}_2)\}\\ \times\omega_1(\boldsymbol{x}_1,\boldsymbol{x}_2)\omega_2(\boldsymbol{y}_1,\boldsymbol{y}_2)\theta(d\boldsymbol{x}_1,d\boldsymbol{y}_1)\theta(d\boldsymbol{x}_2,d\boldsymbol{y}_2),\end{aligned}$$

其中, $(\boldsymbol{x},\boldsymbol{y})\sim\theta$, $\boldsymbol{x}\sim\mu$, $\boldsymbol{y}\sim\nu$ 以及 $(\theta-\mu\otimes\nu)^2\{\overline{B}_\rho(\boldsymbol{x}_1,\boldsymbol{x}_2)\times\overline{B}_\zeta(\boldsymbol{y}_1,\boldsymbol{y}_2)\}=[\theta\{\overline{B}_\rho(\boldsymbol{x}_1,\boldsymbol{x}_2)\times\overline{B}_\zeta(\boldsymbol{y}_1,\boldsymbol{y}_2)\}-\mu\{\overline{B}_\rho(\boldsymbol{x}_1,\boldsymbol{x}_2)\}\nu\{\overline{B}_\zeta(\boldsymbol{y}_1,\boldsymbol{y}_2)\}]^2$. 从而易知, 在 Banach 空间上, $\mathrm{BCov}^2_\omega(\boldsymbol{x},\boldsymbol{y})$ 是 0 当且仅当 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 独立.

在二元联合正态情形下, 文献 [58] 的补充材料证明了球相关系数是 Pearson 相关系数绝对值的增函数, 即

$$\mathrm{BCov}^2_\omega(X,Y)\asymp\mathrm{cor}^2(X,Y).$$

4.1.4 投影相关系数

文献 [88, 105] 从分布函数的角度出发, 通过投影的方式将 Hoeffding 相关系数推广到了多维的情形, 提出了投影相关系数. 文献 [105] 发现衡量随机向量 $\boldsymbol{x}\in\mathbb{R}^p$ 和 $\boldsymbol{y}\in\mathbb{R}^q$ 是否独立, 等价于对于所有的单位投影方向 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$, 检验 $U=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}$ 和 $V=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{y}$ 是否始终相互独立, 即检验

$$\iiint\mathrm{cov}^2\{I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}\leqslant u),I(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{y}\leqslant v)\}dF_{U,V}(u,v)d\boldsymbol{\alpha}d\boldsymbol{\beta}=0.$$

基于此, 文献 [105] 提出了投影协方差

$$\begin{aligned}\{\mathrm{Pcov}(\boldsymbol{x},\boldsymbol{y})\}^2=&\,\mathrm{E}\left[\arccos\left\{\frac{(\boldsymbol{x}_1-\boldsymbol{x}_3)^{\mathrm{T}}(\boldsymbol{x}_4-\boldsymbol{x}_3)}{\|\boldsymbol{x}_1-\boldsymbol{x}_3\|\|\boldsymbol{x}_4-\boldsymbol{x}_3\|}\right\}\arccos\left\{\frac{(\boldsymbol{y}_1-\boldsymbol{y}_3)^{\mathrm{T}}(\boldsymbol{y}_4-\boldsymbol{y}_3)}{\|\boldsymbol{y}_1-\boldsymbol{y}_3\|\|\boldsymbol{y}_4-\boldsymbol{y}_3\|}\right\}\right]\\&+\mathrm{E}\left[\arccos\left\{\frac{(\boldsymbol{x}_1-\boldsymbol{x}_3)^{\mathrm{T}}(\boldsymbol{x}_4-\boldsymbol{x}_3)}{\|\boldsymbol{x}_1-\boldsymbol{x}_3\|\|\boldsymbol{x}_4-\boldsymbol{x}_3\|}\right\}\arccos\left\{\frac{(\boldsymbol{y}_2-\boldsymbol{y}_3)^{\mathrm{T}}(\boldsymbol{y}_5-\boldsymbol{y}_3)}{\|\boldsymbol{y}_2-\boldsymbol{y}_3\|\|\boldsymbol{y}_5-\boldsymbol{y}_3\|}\right\}\right]\\&-2\mathrm{E}\left[\arccos\left\{\frac{(\boldsymbol{x}_1-\boldsymbol{x}_3)^{\mathrm{T}}(\boldsymbol{x}_4-\boldsymbol{x}_3)}{\|\boldsymbol{x}_1-\boldsymbol{x}_3\|\|\boldsymbol{x}_4-\boldsymbol{x}_3\|}\right\}\arccos\left\{\frac{(\boldsymbol{y}_2-\boldsymbol{y}_3)^{\mathrm{T}}(\boldsymbol{y}_4-\boldsymbol{y}_3)}{\|\boldsymbol{y}_2-\boldsymbol{y}_3\|\|\boldsymbol{y}_4-\boldsymbol{y}_3\|}\right\}\right].\end{aligned}$$

从定义可看出, $\mathrm{Pcov}(\boldsymbol{x},\boldsymbol{y})$ 的一个显著特征是它的计算中只涉及 $(\boldsymbol{x}_k-\boldsymbol{x}_l)/\|\boldsymbol{x}_k-\boldsymbol{x}_l\|$ 和 $(\boldsymbol{y}_k-\boldsymbol{y}_l)/\|\boldsymbol{y}_k-\boldsymbol{y}_l\|$, 表明投影协方差消除了距离相关系数所需要的 $(\boldsymbol{x},\boldsymbol{y})$ 矩条件. 文献 [88] 进一步论证了 $\{\mathrm{Pcov}(\boldsymbol{x},\boldsymbol{y})\}^2$ 可等价地表示为

$$\{\mathrm{Pcov}(\boldsymbol{x},\boldsymbol{y})\}^2=4^{-1}\mathrm{E}[\{\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_5,\boldsymbol{x}_2-\boldsymbol{x}_5)+\mathrm{ang}(\boldsymbol{x}_3-\boldsymbol{x}_5,\boldsymbol{x}_4-\boldsymbol{x}_5)$$

$$
\begin{aligned}
&- \operatorname{ang}(\boldsymbol{x}_1-\boldsymbol{x}_5,\boldsymbol{x}_3-\boldsymbol{x}_5)-\operatorname{ang}(\boldsymbol{x}_2-\boldsymbol{x}_5,\boldsymbol{x}_4-\boldsymbol{x}_5)\}\\
&\times\{\operatorname{ang}(\boldsymbol{y}_1-\boldsymbol{y}_5,\boldsymbol{y}_2-\boldsymbol{y}_5)+\operatorname{ang}(\boldsymbol{y}_3-\boldsymbol{y}_5,\boldsymbol{y}_4-\boldsymbol{y}_5)\\
&- \operatorname{ang}(\boldsymbol{y}_1-\boldsymbol{y}_5,\boldsymbol{y}_3-\boldsymbol{y}_5)-\operatorname{ang}(\boldsymbol{y}_2-\boldsymbol{y}_5,\boldsymbol{y}_4-\boldsymbol{y}_5)\}]\\
\leqslant\ & 4^{-1}\mathrm{E}^{1/2}\{\operatorname{ang}(\boldsymbol{x}_1-\boldsymbol{x}_5,\boldsymbol{x}_2-\boldsymbol{x}_5)+\operatorname{ang}(\boldsymbol{x}_3-\boldsymbol{x}_5,\boldsymbol{x}_4-\boldsymbol{x}_5)\\
&- \operatorname{ang}(\boldsymbol{x}_1-\boldsymbol{x}_5,\boldsymbol{x}_3-\boldsymbol{x}_5)-\operatorname{ang}(\boldsymbol{x}_2-\boldsymbol{x}_5,\boldsymbol{x}_4-\boldsymbol{x}_5)\}^2\\
&\times\mathrm{E}^{1/2}\{\operatorname{ang}(\boldsymbol{y}_1-\boldsymbol{y}_5,\boldsymbol{y}_2-\boldsymbol{y}_5)+\operatorname{ang}(\boldsymbol{y}_3-\boldsymbol{y}_5,\boldsymbol{y}_4-\boldsymbol{y}_5)\\
&- \operatorname{ang}(\boldsymbol{y}_1-\boldsymbol{y}_5,\boldsymbol{y}_3-\boldsymbol{y}_5)-\operatorname{ang}(\boldsymbol{y}_2-\boldsymbol{y}_5,\boldsymbol{y}_4-\boldsymbol{y}_5)\}^2\\
=\ & \operatorname{Pcov}(\boldsymbol{x},\boldsymbol{x})\operatorname{Pcov}(\boldsymbol{y},\boldsymbol{y}),
\end{aligned}
$$

其中 $\operatorname{ang}(\cdot,\cdot)$ 表示两同型向量之间的角度.

自然地, 投影相关系数的定义为

$$\{\operatorname{PC}(\boldsymbol{x},\boldsymbol{y})\}^2=\frac{\{\operatorname{Pcov}(\boldsymbol{x},\boldsymbol{y})\}^2}{\operatorname{Pcov}(\boldsymbol{x},\boldsymbol{x})\operatorname{Pcov}(\boldsymbol{y},\boldsymbol{y})},$$

并且如果 $\operatorname{Pcov}(\boldsymbol{x},\boldsymbol{x})=0$ 或 $\operatorname{Pcov}(\boldsymbol{y},\boldsymbol{y})=0$, 则令 $\operatorname{PC}(\boldsymbol{x},\boldsymbol{y})=0$. 通常情形下, $0\leqslant\operatorname{PC}(\boldsymbol{x},\boldsymbol{y})\leqslant 1$, 并且 $\operatorname{PC}(\boldsymbol{x},\boldsymbol{y})=0$ 当且仅当 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 相互独立. 投影相关系数还具有正交变换的不变性, 即对于任意两组正交矩阵 $\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$, 有

$$\operatorname{PC}(\boldsymbol{x},\boldsymbol{y})=\operatorname{PC}(a_1+\boldsymbol{b}_1\boldsymbol{C}_1\boldsymbol{x},a_2+\boldsymbol{b}_2\boldsymbol{C}_2\boldsymbol{y}).$$

在 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 相互独立的原假设下, $\operatorname{PC}(\boldsymbol{x},\boldsymbol{y})$ 依分布收敛到权数未知的无穷卡方和分布. 检验的临界值可以由随机置换的方法来获得. 文献 [94] 从计算复杂度、原假设下的渐近分布和高维情形下的功效这 3 个方面对投影相关系数进行了改进. 在文献 [16,105] 中投影想法的基础上, 文献 [33,87] 构造能够适应高维数据的两样本分布相等的稳健非参数检验方法.

4.1.5 互信息

文献 [5] 基于互信息 (mutual information) 来度量两个随机向量的非线性相关性. 令 $\mathcal{Z}=\{(\boldsymbol{x},\boldsymbol{y}):f(\boldsymbol{x},\boldsymbol{y})>0\}$. 互信息的定义为

$$\operatorname{MI}(\boldsymbol{x},\boldsymbol{y})=\int_{\mathcal{Z}}f(\boldsymbol{x},\boldsymbol{y})\log\frac{f(\boldsymbol{x},\boldsymbol{y})}{f(\boldsymbol{x})f(\boldsymbol{y})}d\lambda_d(\boldsymbol{x},\boldsymbol{y}),$$

其中对 $D\in\mathbb{N}$, λ_D 表示 $\mathbb{R}^D$ 的 Lebesgue 测度. 从定义可以看出, 互信息度量了 $(\boldsymbol{x},\boldsymbol{y})$ 联合密度函数与边际密度函数乘积之间的 Kullback-Leibler 差异. 因此, 互信息是一个非负的度量准则, 其等于 0 当且仅当 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 相互独立. 在对 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 进行可逆变换时, 互信息的数值保持不变. 在 R 语言中实现互信息独立性检验可以用 fastmit::mi.test ( ) 计算.

文献 [34] 指出, 互信息定义为对数似然比的形式, 因而可以为任何相依性检验的表现提供一个严格的上界. 文献 [63] 基于互信息, 定义了极大信息系数 (maximal information coefficient, MIC). 对于网格 G, 令 I_G 表示基于 G 概率分布的互信息. 特征矩阵中第 (x,y) 项为 $m_{x,y}=\max(I_G)/\log\min(x,y)$, 则对有序对 $(x,y)<B$, 当 B 为样本的函数时, MIC 为 $m_{x,y}$ 的极大值. 文献 [63] 证明了对于非常数的无噪声函数关系, MIC 以趋于 1 的概率取值为 1. 对于相互独立的随机变量, MIC 以趋于 1 的概率取值为 0. 文献 [63] 还发现对于函数关系, MIC 的取值大致等于回归函数的决定系数 R^2. 文献 [72]

讨论了 MIC 的相关性质, 并对将 MIC 拓展至条件非线性度量提出了展望. MIC 在 R 语言中可以用 metrica::MIC ( ) 计算.

4.1.6 基于多元分布函数转换的相关系数

当 $p=q=1$ 时, 对两个一维随机变量分别进行相应的分布函数变换, 利用变换后的非线性相依度量构造统计量, 其极限零分布是分布自由的, 即不依赖于未知总体的分布. 但针对多维随机向量, 使用距离相关系数和投影相关系数的独立性检验统计量在零假设下的分布具有加权的无穷卡方和形式, 并且依赖于未知非参数总体的某些数字特征.

为克服这一问题, 文献 [68] 引入了 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 多维分布函数的转换分布 $\boldsymbol{F}_{1,\pm}(\cdot)$ 和 $\boldsymbol{F}_{2,\pm}(\cdot)$ 使得 $\boldsymbol{F}_{1,\pm}(\boldsymbol{x})\sim\boldsymbol{U}_p$ 和 $\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\sim\boldsymbol{U}_q$, 其中, $\boldsymbol{U}_d$ 是随机向量 $R\boldsymbol{z}/\|\boldsymbol{z}\|$ 的分布函数, R 是服从 [0, 1] 上均匀分布的随机变量, $\boldsymbol{z}$ 是服从标准正态分布的随机向量, 并且 R 和 $\boldsymbol{z}$ 独立. 根据文献 [69, 命题 4.2], $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 独立等价于 $\boldsymbol{F}_{1,\pm}(\boldsymbol{x})$ 和 $\boldsymbol{F}_{2,\pm}(\boldsymbol{y})$ 独立. 因此, 相应的多维 Kendall τ 相关系数、距离相关系数、投影相关系数、HHG 相关系数、球相关系数和互信息分别定义为

$$
\begin{aligned}
&\text{IPR-}\tau_b^*\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}=\frac{\text{IPR-}\tau^*\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}}{[\text{IPR-}\tau^*\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}]^{1/2}[\text{IPR-}\tau^*\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}]^{1/2}},\\
&\text{dCor}^2\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}=\frac{\text{dCov}^2\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}}{\text{dCov}\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}\text{dCov}\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}},\\
&\text{PC}^2\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}=\frac{[\text{Pcov}\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}]^2}{\text{Pcov}\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}\text{Pcov}\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}},\\
&\text{BCor}_\omega^2\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}=\frac{\text{BCov}_\omega^2\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}}{\text{BCov}_\omega\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}\text{BCov}_\omega\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}},\\
&\text{MI}\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}=\int_{\mathcal{Z}} f\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}\log\frac{f\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}}{f\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x})\}f\{\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}}d\lambda_d\{\boldsymbol{F}_{1,\pm}(\boldsymbol{x}),\boldsymbol{F}_{2,\pm}(\boldsymbol{y})\}.
\end{aligned}
$$

为了在统一的框架下考虑一类渐近分布自由的度量, 文献 [69] 定义了广义对称协方差 (generalized symmetric covariance), 存在两个核函数 $f_1:(\mathbb{R}^{d_1})^m\to\mathbb{R}_{\geqslant 0}$ 和 $f_2:(\mathbb{R}^{d_2})^m\to\mathbb{R}_{\geqslant 0}$, 子群 H 包含数量相等的奇数、偶数置换,

$$
\mu(\boldsymbol{x},\boldsymbol{y})=\mu_{f_1,f_2,H}(\boldsymbol{x},\boldsymbol{y})=\text{E}[k_{f_1,f_2,H}\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_m,\boldsymbol{y}_m)\}],
$$

其中, $(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_m,\boldsymbol{y}_m)$ 为 m 个简单随机样本, 核函数 $k_{f_1,f_2,H}(\cdot)$ 定义为

$$
\begin{aligned}
&k_{f_1,f_2,H}\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_m,\boldsymbol{y}_m)\}\\
&\quad=\Big\{\sum_{\sigma\in H}\text{sgn}(\sigma)f_1(\boldsymbol{x}_{\sigma(1)},\ldots,\boldsymbol{x}_{\sigma(m)})\Big\}\Big\{\sum_{\sigma\in H}\text{sgn}(\sigma)f_2(\boldsymbol{y}_{\sigma(1)},\ldots,\boldsymbol{y}_{\sigma(m)})\Big\}.
\end{aligned}
$$

广义对称协方差的概念涵盖了许多上述提及的度量, 如距离相关系数 [75]、投影相关系数 [105] 和基于投影的多维 BKR 相关系数 [33] 等. 为了使渐近分布自由, 文献 [69] 紧接着考虑了对广义对称协方差的变换. 定义得分函数 $J_1,J_2:[0,1)\to\mathbb{R}_{\geqslant 0}$. 则对于 $k=1,2$, 令 $\boldsymbol{J}_k(\boldsymbol{u})=J_k(\|\boldsymbol{u}\|)\boldsymbol{u}/\|\boldsymbol{u}\|$. 若 $\boldsymbol{u}$ 为零向量, 则 $\boldsymbol{J}_k(\boldsymbol{u})=\boldsymbol{0}$. 基于得分的分布函数转换则为 $\boldsymbol{G}_{k,\pm}(\cdot)=\boldsymbol{J}_k(\boldsymbol{F}_{k,\pm}(\cdot))$. 文献 [69] 定义了转换后的度量为

$$
\mu_\pm(\boldsymbol{x},\boldsymbol{y})=\mu_{\pm;f_1,f_2,H}(\boldsymbol{x},\boldsymbol{y})=\mu_{f_1,f_2,H}(\boldsymbol{G}_{1,\pm}(\boldsymbol{x}),\boldsymbol{G}_{2,\pm}(\boldsymbol{y})).
$$

4.2 均值独立

4.2.1 鞅差相关系数

为了度量随机变量 Y 和随机向量 $\boldsymbol{x}\in\mathbb{R}^p$ 之间的均值相依性, 通过推广距离协方差的定义, 文献 [66] 引入了鞅差散度的定义

$$\mathrm{MDD}(Y\mid\boldsymbol{x})^2=\int_{\mathbb{R}^p}\frac{\|\mathrm{E}\{Y\exp(\mathrm{i}\langle\boldsymbol{t},\boldsymbol{x}\rangle)\}-\mathrm{E}\exp(\mathrm{i}\langle\boldsymbol{t},\boldsymbol{x}\rangle)\mathrm{E}(Y)\|^2}{c_p\|\boldsymbol{t}\|^{1+p}}d\boldsymbol{t}.$$

由该定义可知, $\mathrm{MDD}(Y\mid\boldsymbol{x})=0$ 当且仅当 $P\{\mathrm{E}(Y\mid\boldsymbol{x})=\mathrm{E}(Y)\}=1$. 如果 $\mathrm{E}(Y^2+\|\boldsymbol{x}\|^2)<\infty$, 则 $\mathrm{MDD}(Y\mid\boldsymbol{x})^2$ 具有如下的显式代数表示:

$$\begin{aligned}\mathrm{MDD}(Y\mid\boldsymbol{x})^2&=-\mathrm{E}[\{Y_1-\mathrm{E}(Y)\}\{Y_2-\mathrm{E}(Y)\}\|\boldsymbol{x}_1-\boldsymbol{x}_2\|]\\&=8^{-1}\mathrm{E}[\{(Y_1-Y_2)^2+(Y_3-Y_4)^2-(Y_1-Y_3)^2-(Y_2-Y_4)^2\}\\&\quad\times(\|\boldsymbol{x}_1-\boldsymbol{x}_2\|+\|\boldsymbol{x}_3-\boldsymbol{x}_4\|-\|\boldsymbol{x}_1-\boldsymbol{x}_3\|-\|\boldsymbol{x}_2-\boldsymbol{x}_4\|)]\\&\leqslant 8^{-1}\mathrm{E}^{1/2}[\{(Y_1-Y_2)^2+(Y_3-Y_4)^2-(Y_1-Y_3)^2-(Y_2-Y_4)^2\}^2\\&\quad\times\mathrm{E}^{1/2}(\|\boldsymbol{x}_1-\boldsymbol{x}_2\|+\|\boldsymbol{x}_3-\boldsymbol{x}_4\|-\|\boldsymbol{x}_1-\boldsymbol{x}_3\|-\|\boldsymbol{x}_2-\boldsymbol{x}_4\|)^2\\&=\mathrm{var}(Y)\mathrm{dCov}(\boldsymbol{x},\boldsymbol{x}).\end{aligned}$$

因而, 给定 $\boldsymbol{x}$ 时 Y 的鞅差相关系数很自然地可定义为

$$\mathrm{MDC}(Y\mid\boldsymbol{x})^2=\frac{\mathrm{MDD}(Y\mid\boldsymbol{x})^2}{\mathrm{var}(Y)\mathrm{dCov}(\boldsymbol{x},\boldsymbol{x})}.$$

鞅差相关系数在 R 语言中可以用 EDMeasure::mdc ( ) 计算. 文献 [90] 在因变量为一般度量空间情形下, 提出了充分降维方法, 并讨论了与 MDD 的联系.

针对二元联合正态随机向量 (X,Y), 若其相关系数为 ρ, 文献 [66] 得到了恒等式

$$\mathrm{MDC}(X,Y)^2=\frac{\rho^2}{\sqrt{4(1-3^{1/2}+\pi/3)}}.$$

4.2.2 累积投影散度

对于一维随机变量, 累积散度可以充分度量它们之间的均值独立性. 由于累积散度是基于秩的度量准则, 因此其不需要假设条件变量的矩条件, 并且对条件变量的任意单调变换保持数值不变. 虽然累积散度具有上述优点, 但仅适用于单变量使其应用范围受限. 文献 [85] 在多元空间中通过对投影平均推广了累积协方差, 提出了累积投影协方差

$$\{\mathrm{PCCov}(Y\mid\boldsymbol{x})\}^2=c(p)^{-1}\int\int\mathrm{cov}^2\{Y,\ I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}\leqslant u)\}dF_U(u)d\boldsymbol{\alpha},$$

其中 $c(p)=\pi^{p/2-1}/\Gamma(p/2)$, $\Gamma(\cdot)$ 表示 Gamma 函数. $F_U(u)$ 表示随机变量 $U=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}\in\mathbb{R}^1$ 的边际分布函数. 由定义可知, $\mathrm{PCCov}(Y\mid\boldsymbol{x})=0$ 当且仅当 $\mathrm{P}\{\mathrm{E}(Y\mid\boldsymbol{x})=\mathrm{E}(Y)\}=1$.

假设 $\mathrm{E}(|Y|)<\infty$, 则

$$\{\mathrm{PCCov}(Y\mid\boldsymbol{x})\}^2=-\mathrm{E}\{Y_1Y_2\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_3,\boldsymbol{x}_2-\boldsymbol{x}_3)\}-\mathrm{E}\{Y_1Y_2\mathrm{ang}(\boldsymbol{x}_3-\boldsymbol{x}_5,\boldsymbol{x}_4-\boldsymbol{x}_5)\}$$

$$
\begin{aligned}
&+2\mathrm{E}\{Y_1Y_2\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_4,\boldsymbol{x}_3-\boldsymbol{x}_4)\}\\
&\geqslant 0,
\end{aligned}
$$

其中

$$
\mathrm{ang}(\boldsymbol{x}_k-\boldsymbol{x}_r,\boldsymbol{x}_l-\boldsymbol{x}_r)=\arccos\frac{(\boldsymbol{x}_k-\boldsymbol{x}_r)^{\mathrm{T}}(\boldsymbol{x}_l-\boldsymbol{x}_r)}{\|\boldsymbol{x}_k-\boldsymbol{x}_r\|\,\|\boldsymbol{x}_l-\boldsymbol{x}_r\|},
$$

其中, $\|\cdot\|$ 表示向量的 L_2 范数, 并且 $(\boldsymbol{x}_i,Y_i)$ $(i=1,\ldots,5)$ 为 $(\boldsymbol{x},Y)$ 的独立副本. 从上式可以看出, 累积投影协方差的显式表达形式简洁且不包含任何的调节参数, 计算只用到了 $\mathrm{ang}(\boldsymbol{x}_k-\boldsymbol{x}_r,\boldsymbol{x}_l-\boldsymbol{x}_r)$, 并且 $\{\mathrm{PCCov}(Y\mid\boldsymbol{x})\}^2$ 可用角度的双中心距离表示, 即

$$
\begin{aligned}
\{\mathrm{PCCov}(Y\mid\boldsymbol{x})\}^2&=8^{-1}\mathrm{E}[\{(Y_1-Y_2)^2+(Y_3-Y_4)^2-(Y_1-Y_3)^2-(Y_2-Y_4)^2\}\\
&\quad\times\{\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_5,\boldsymbol{x}_2-\boldsymbol{x}_5)+\mathrm{ang}(\boldsymbol{x}_3-\boldsymbol{x}_5,\boldsymbol{x}_4-\boldsymbol{x}_5)\\
&\quad-\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_5,\boldsymbol{x}_3-\boldsymbol{x}_5)-\mathrm{ang}(\boldsymbol{x}_2-\boldsymbol{x}_5,\boldsymbol{x}_4-\boldsymbol{x}_5)\}]\\
&\leqslant 8^{-1}\mathrm{E}^{1/2}\{(Y_1-Y_2)^2+(Y_3-Y_4)^2-(Y_1-Y_3)^2-(Y_2-Y_4)^2\}^2\\
&\quad\times\mathrm{E}^{1/2}\{\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_5,\boldsymbol{x}_2-\boldsymbol{x}_5)+\mathrm{ang}(\boldsymbol{x}_3-\boldsymbol{x}_5,\boldsymbol{x}_4-\boldsymbol{x}_5)\\
&\quad-\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_5,\boldsymbol{x}_3-\boldsymbol{x}_5)-\mathrm{ang}(\boldsymbol{x}_2-\boldsymbol{x}_5,\boldsymbol{x}_4-\boldsymbol{x}_5)\}^2\\
&=\mathrm{var}(Y)\mathrm{Pcov}(\boldsymbol{x},\boldsymbol{x}).
\end{aligned}
$$

因此, 累积投影相关系数的定义为

$$
\mathrm{PCD}(Y\mid\boldsymbol{x})=\frac{\mathrm{PCCov}(Y\mid\boldsymbol{x})}{\sqrt{\mathrm{var}(Y)\mathrm{Pcov}(\boldsymbol{x},\boldsymbol{x})}},
$$

其中 $\mathrm{var}(Y)$ 表示 Y 的方差, $\mathrm{Pcov}(\boldsymbol{x},\boldsymbol{x})$ 为定义在文献 [105] 中的投影协方差. 当 $\mathrm{var}(Y)\mathrm{Pcov}(\boldsymbol{x},\boldsymbol{x})=0$ 时, 定义 $\mathrm{PCD}(Y\mid\boldsymbol{x})=0$.

针对二元联合正态随机向量 (X,Y), 若其相关系数为 ρ, $\mathrm{PCD}(Y\mid X)$ 退化为 $\mathrm{CD}(Y\mid X)$, 第 3.4.1 小节已经讨论了 $\mathrm{CD}(Y\mid X)$ 与 ρ 之间的具体关系.

5 高维随机向量的关联度量

对于许多经典的独立性度量, 维数的固定与发散会对它们的性质产生决定性的影响. 例如, 若独立性检验统计量是一个退化的 U 统计量, 则在固定维数下, 其零假设下的渐近分布为无穷多卡方分布的加权和, 权重依赖于随机向量的分布. 而在维数发散的高维情形下, 可能会收敛到正态分布. 原本在固定维数下可以度量任意非线性关系的准则, 而在高维情形下, 可能仅能捕捉线性相关.

5.1 分布独立

5.1.1 距离相关系数

当随机向量的维数发散速度快于样本量时, 文献 [104] 发现距离相关系数不再具有度量任意非线性相关的能力, 而是退化为只能捕捉随机向量之间的线性关系. 具体而言,

$$
\mathrm{dCov}^2(\boldsymbol{x},\boldsymbol{y})\approx c^{-1}\sum_{k=1}^{p}\sum_{l=1}^{q}\mathrm{cov}^2(X_k,Y_l),
$$

其中, X_k 和 Y_l 为随机向量的单个分量, c 为一个依赖于 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分布以及数据维数的常数. 这一式子表明, 在高维情形下, 距离相关系数与 Pearson 相关系数具有类似的表现.

为了在高维情形下能够度量非线性关系, 文献 [104] 提出将非线性度量边际地进行计算后再聚合, 以损失高维数据的相关性信息为代价来避免维数的影响, 即采用组装的度量

$$\mathrm{mdCov}^2(\boldsymbol{x},\boldsymbol{y}) = \{C(n,2)\}^{1/2}\sum_{k=1}^{p}\sum_{l=1}^{q}\mathrm{dCov}^2(X_k,Y_l),$$

其中 $C(n,d)$ 表示从 $\{1,\ldots,n\}$ 中抽取 d 个不同元素的所有组合数. 从定义可以看出, 该度量可以捕捉每一数据维数的非线性相依性.

5.1.2 基于秩的相关系数

除了距离相关系数, 在高维情形下, 我们还关心基于秩的相关系数在度量非线性关系上的表现. 文献 [101] 考虑了如何利用 Hoeffding、BKR 和 τ^* 相关系数在高维情形下进行独立性检验. 与文献 [104] 的思路一致, 文献 [101] 采用基于秩的相关系数边际计算随机向量单个分量之间的相依性, 再进行求和聚合.

在高维数据的框架下, 文献 [101] 定义了核函数

$$U_h^{(kl)} = \begin{cases} \mathrm{H}(X_k,Y_l), & h = h^{(\mathrm{H})}, \\ \mathrm{BKR}(X_k,Y_l), & h = h^{(\mathrm{BKR})}, \\ \tau^*(X_k,Y_l), & h = h^{(\tau^*)}, \end{cases}$$

并定义相应的组装的秩相关系数

$$T_h = \Delta_h\{C(d,2)\}^{-1}\sum_{k=1}^{p}\sum_{l=1}^{q}U_h^{(kl)},$$

其中, d 为核函数的阶数, Δ_h 为一个常数调节因子,

$$d = \begin{cases} 5, & h = h^{(\mathrm{H})}, \\ 6, & h = h^{(\mathrm{BKR})}, \\ 4, & h = h^{(\tau^*)}. \end{cases} \quad \Delta_h = \begin{cases} 40, & h = h^{(\mathrm{H})}, \\ 60, & h = h^{(\mathrm{BKR})}, \\ \dfrac{2}{3}, & h = h^{(\tau^*)}. \end{cases}$$

在功效分析中, 文献 [101] 考虑了一类 Gauss 等相关的备择假设, 证明了在随机向量同方差的情形下, 基于秩的独立性检验相较于基于距离相关系数的检验会有效率损失. 但在异方差的情形下, 基于秩的独立性检验会使得检验的效率有着极大的提升. 实现检验的 R 语言代码在 github.com/Yeqing-TJ 上.

5.2 均值独立

5.2.1 线性回归模型的系数检验

在线性模型的框架下, 检验 $\mathrm{E}(Y \mid \boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$ 等价于检验 $\boldsymbol{\beta} = 0$ 是否成立. 传统的 F 检验可以用来检验线性回归系数的整体显著性. 但是当回归模型中存在高维协变量时, F 检验的功效将会随着数据维数与样本量之比的上升而逐渐下降 [95]. 当维数大于样本量时, F 检验会由于样本协方差不可逆而完全失效.

为了能够处理高维回归问题, 文献 [95] 修正了原来的 F 检验, 并建议使用

$$\sum_{s=1}^{p}\mathrm{cov}^2(Y,X_s)$$

来检验线性回归系数的整体显著性. 文献 [95] 在不限制 n 和 p 发散速率的前提下, 研究了其估计量的渐近正态性.

除了求和聚合, 还可以通过寻找极大值来聚合边际的检验统计量. 通常来说, 基于求和的检验统计量在密集型的备择假设下表现较好, 而基于极大值的检验统计量在稀疏型的备择假设下具有更强的功效. 文献 [86] 基于 Wilcoxon 得分为检验线性回归的系数提出了极大值型的检验统计量, 增强对非正态分布的厚尾变量的稳健性. 同时也将协变量之间的相依关系融入到检验之中, 有助于提升检验的功效.

5.2.2 鞅差相关系数

文献 [95] 的检验基于协方差函数, 因此仅能够度量线性关系. 在高维情形下, 不假设任何的模型形式, 研究均值独立是非常困难的. 因此, 文献 [91] 将研究的目标定在一个稍弱的假设检验问题上:

$$H_0':\mathrm{E}(Y\mid X_s)=\mathrm{E}(Y),\quad \text{对所有的 } 1\leqslant s\leqslant p \text{ 几乎处处成立.}$$

为了解决这一问题, 文献 [91] 考虑使用每一维数的鞅差相关系数进行求和聚合, 即

$$\sum_{s=1}^{p}\mathrm{MDD}(Y\mid X_s)^2.$$

与文献 [95] 相比, 这一度量可以捕捉非线性的均值关系, 因而应用范围更为广泛.

5.2.3 累积散度

与固定维数下相似, 鞅差相关系数依赖于回归模型协变量的矩条件. 在高维空间中, 随机向量不同分量之间方差差异可能很大. 文献 [91, 95] 中的检验均不是尺度不变的. 因此, 高维异方差结构会对均值检验的功效产生影响. 为了解决这个问题, 一种常用的策略是在进行上述检验之前, 通过除以其相应的标准差对每个协变量进行标准化. 但是这一策略在发散的数据维数下带来了理论论证的困难, 并且隐含要求所有协变量的方差是有限的.

文献 [43] 从累积差异的角度分析和解决该问题, 考虑使用每一维数的累积协方差进行求和聚合, 即

$$\sum_{s=1}^{p}\mathrm{CCov}(Y\mid X_s).$$

文献 [43] 在线性模型的框架下比较了文献 [91, 95] 中检验的功效. 当每个协变量具有相同方差时, 文献 [95] 渐近功效最优, 主要是由于其专门为线性模型设计, 而另外两个检验均不假设模型的具体形式. 当每个协变量的方差差异很大时, 文献 [43] 中检验方法的功效会优于其他两种方法. 实现检验的 R 语言代码在 github.com/Yeqing-TJ 上.

6 条件关联度量

作为独立性的拓展, 条件独立性度量了在给定向量 $\boldsymbol{z}=(Z_1,\ldots,Z_d)^{\mathrm{T}}\in\mathbb{R}^d$ 时, $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 之间的关联性. 条件独立性是处理很多科学问题的重要假设. 例如, 在进行因果推断时, 条件独立性假设可以控制实验的随机化. 当处理模型的内生性时, 引入工具变量需要满足条件独立性条件. 条件独立性还可以刻画图模型中节点与节点之间的因果关系. 鉴于此, 接下来的章节将讨论如何度量与检验数据之间的条件关联性.

6.1 偏相关系数

度量 X 和 Y 在给定 $\boldsymbol{z}$ 时的条件相关性, 最经典的方法之一是偏相关系数 [38]. 偏相关系数计算了 X 和 Y 分别关于 $\boldsymbol{z}$ 作线性回归之后两个残差的 Pearson 相关系数, 即

$$\mathrm{pCor}=\frac{\mathrm{cov}(X-\alpha_1-\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{z},Y-\alpha_2-\boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{z})}{\sqrt{\mathrm{var}(X-\alpha_1-\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{z})\mathrm{var}(Y-\alpha_2-\boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{z})}}.$$

偏相关系数在 R 语言中可以用 ppcor::pcor ( ) 计算.

当 X、Y 和 $\boldsymbol{z}$ 的联合分布是正态分布时, 零偏相关系数等价于条件独立. 但偏相关系数只考虑了两个残差的线性相关性, 因此它不具备度量 X 和 Y 之间非线性条件相关的能力.

为了解决这一问题并能够处理高维的条件变量, 文献 [18] 先利用惩罚最小二乘获得回归系数的估计, 再采用距离相关系数 [75] 度量两个残差的独立性. 与偏相关系数一样, 它也假定了 $\boldsymbol{x}$ 和 $\boldsymbol{z}$ 以及 $\boldsymbol{y}$ 和 $\boldsymbol{z}$ 之间满足线性模型. 在分位数的框架下, 文献 [40] 还考虑了分位数偏相关系数.

与偏相关系数类似, 条件相关系数也从线性关系出发, 定义为

$$\mathrm{cCor}(X,Y\mid Z)=\frac{\mathrm{cov}(X,Y\mid Z)}{\sqrt{\mathrm{var}(X\mid Z)\mathrm{var}(Y\mid Z)}}.$$

6.2 条件距离相关系数

文献 [77] 提出了条件距离相关系数 (conditional distance correlation) 去度量与检验条件相依性. 距离相关系数是从特征函数的角度分析随机向量的非线性关系. 沿着这个思路, 条件距离相关系数采用条件特征函数来完成条件非线性关系的度量. 条件距离协方差相应的定义为

$$\mathrm{CDCov}^2(\boldsymbol{x},\boldsymbol{y}\mid\boldsymbol{z})=\int_{\mathbb{R}^{p+q}}\frac{\|\mathrm{E}\{\exp(\mathrm{i}\langle\boldsymbol{t},\boldsymbol{x}\rangle+\mathrm{i}\langle\boldsymbol{s},\boldsymbol{y}\rangle)\mid\boldsymbol{z}\}-\mathrm{E}\{\exp(\mathrm{i}\langle\boldsymbol{t},\boldsymbol{x}\rangle)\mid\boldsymbol{z}\}\mathrm{E}\{\exp(\mathrm{i}\langle\boldsymbol{s},\boldsymbol{y}\rangle)\mid\boldsymbol{z}\}\|^2}{c_pc_q\|\boldsymbol{t}\|^{1+p}\|\boldsymbol{s}\|^{1+q}}d\boldsymbol{t}d\boldsymbol{s},$$

其中, $c_p=\pi^{(1+p)/2}/\Gamma\{(1+p)/2\}$, $c_q=\pi^{(1+q)/2}/\Gamma\{(1+q)/2\}$, $\Gamma\{\cdot\}$ 表示 Gamma 函数, $\langle\cdot,\cdot\rangle$ 表示内积, $\|\cdot\|$ 表示 L_2 范数. 条件距离相关系数在 R 语言中可以用 cdcsis::cdcor ( ) 计算.

文献 [77, 定理 1] 介绍了条件协方差的性质. CDCov^2 总是非负的, 其等于 0 当且仅当在给定 $\boldsymbol{z}$ 时, $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 相互条件独立. 对于任意向量 $\boldsymbol{a}_1\in\mathbb{R}^p$, $\boldsymbol{a}_2\in\mathbb{R}^q$, 常数 b_1 和 b_2, $p\times p$ 正交矩阵 $\boldsymbol{C}_1$, $q\times q$ 正交矩阵 $\boldsymbol{C}_2$, 有

$$\mathrm{CDCov}(\boldsymbol{a}_1+b_1\boldsymbol{C}_1\boldsymbol{x},\boldsymbol{a}_2+b_2\boldsymbol{C}_2\boldsymbol{y}\mid\boldsymbol{z})=\sqrt{|b_1b_2|}\mathrm{CDCov}(\boldsymbol{x},\boldsymbol{y}\mid\boldsymbol{z}).$$

如果 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 在给定 $\boldsymbol{z}$ 时相互独立, 则

$$\mathrm{CDCov}(\boldsymbol{x}+\boldsymbol{y}\mid\boldsymbol{z})\leqslant\mathrm{CDCov}(\boldsymbol{x}\mid\boldsymbol{z})+\mathrm{CDCov}(\boldsymbol{y}\mid\boldsymbol{z}).$$

文献 [77] 还定义了条件距离相关系数

$$\mathrm{CDCor}^2(\boldsymbol{x},\boldsymbol{y}\mid\boldsymbol{z})=\frac{\mathrm{CDCov}^2(\boldsymbol{x},\boldsymbol{y}\mid\boldsymbol{z})}{\mathrm{CDCov}(\boldsymbol{x},\boldsymbol{x}\mid\boldsymbol{z})\mathrm{CDCov}(\boldsymbol{y},\boldsymbol{y}\mid\boldsymbol{z})}.$$

6.3 基于 BKR 相关系数的条件相依性度量

文献 [100] 通过对随机变量 X 和 Y 进行 Rosenblatt 变换[64], 证明了在温和的条件下, X 和 Y 之间给定 $\boldsymbol{z}$ 的条件独立性等价于条件分布 $F_1(X\mid\boldsymbol{z})$ 和 $F_2(Y\mid\boldsymbol{z})$ 之间的无条件独立性. 假设对于任意给定的 $\boldsymbol{z}$, X 和 Y 均是连续的, 则

(1) 条件独立性 $X\perp\!\!\!\perp Y\mid\boldsymbol{z}$ 可以推导出无条件独立性 $F_1(X\mid\boldsymbol{z})\perp\!\!\!\perp F_2(Y\mid\boldsymbol{z})$;

(2) 在假设 $\{F_1(X\mid\boldsymbol{z}),F_2(Y\mid\boldsymbol{z})\}\perp\!\!\!\perp\boldsymbol{z}$ 的情形下, 无条件独立性 $F_1(X\mid\boldsymbol{z})\perp\!\!\!\perp F_2(Y\mid\boldsymbol{z})$ 可以推导出条件独立性 $X\perp\!\!\!\perp Y\mid\boldsymbol{z}$.

这种等价性可以将检验条件独立性的问题转化成检验无条件独立性的问题. 第一部分确保了只要 X 和 Y 是连续的, 检验的尺度会得到很好的控制; 第二部分保证了对于所有满足 $\{F_1(X\mid\boldsymbol{z}),F_2(Y\mid\boldsymbol{z})\}\perp\!\!\!\perp\boldsymbol{z}$ 的备择假设, 检验都具有非平凡的功效.

基于该等价性, 非线性相依性度量均可以运用于检验条件相依性. 文献 [100] 选择了 BKR 相关系数来检验 $F_1(X\mid\boldsymbol{z})$ 和 $F_2(Y\mid\boldsymbol{z})$ 之间的独立性. 为了符号的简洁性, 定义 $V=F_1(X\mid\boldsymbol{z})$ 和 $W=F_2(Y\mid\boldsymbol{z})$, 则有 $V\perp\!\!\!\perp\boldsymbol{z}$ 和 $W\perp\!\!\!\perp\boldsymbol{z}$. 令 $F_V(v)$ 和 $F_W(w)$ 分别表示 V 和 W 的边际分布函数, $F_{V,W}(v,w)$ 表示 (V,W) 的联合分布函数, 则 V 与 W 之间的 BKR 相关系数定义为

$$\rho^{\mathrm{CI}}=\int_{\mathbb{R}^1}\int_{\mathbb{R}^1}\{F_{V,W}(v,w)-F_V(v)F_W(w)\}^2dF_V(v)dF_W(w).$$

从定义可以看出, ρ^{CI} 是非负的, 其等于 0 当且仅当 V 与 W 相互独立. 在对 X 或 Y 进行单调变换之后, ρ^{CI} 值仍保持不变. 文献 [100] 提出的检验统计量在原假设下的渐近分布不依赖于 X、Y 或 $\boldsymbol{z}$ 的分布, 并且表达式中不涉及任何的调节参数, 能够快速准确地完成条件独立性检验.

6.4 基于相互独立的条件相依性度量

文献 [9] 建立了条件独立与 3 个随机变量之间相互独立之间的等价关系. 文献 [9] 关注的是 X、Y 和 Z 为连续型一维随机变量的设定. 令 $U=F_{X|Z}(X\mid Z)$, $V=F_{Y|Z}(Y\mid Z)$ 和 $W=F_Z(Z)$. 假设一维随机变量 X 和 Y 对于每一个给定 Z 的值都有连续的条件分布函数, Z 也是一个连续的一维变量随机变量, 则 $X\perp\!\!\!\perp Y\mid Z$ 当且仅当 U、V 和 W 相互独立 (mutually independent).

鉴于此等价关系, 文献 [9] 开始从特征函数的角度出发研究 3 个随机变量的相互独立性, 即

$$\begin{aligned}\rho^M(X,Y\mid Z)=\ & c_0\mathrm{E}\{(\mathrm{e}^{-|U_1-U_2|}+\mathrm{e}^{-U_1}+\mathrm{e}^{U_1-1}+\mathrm{e}^{-U_2}+\mathrm{e}^{U_2-1}+2\mathrm{e}^{-1}-4)\\&\times(\mathrm{e}^{-|V_1-V_2|}+\mathrm{e}^{-V_1}+\mathrm{e}^{V_1-1}+\mathrm{e}^{-V_2}+\mathrm{e}^{V_2-1}+2\mathrm{e}^{-1}-4)\mathrm{e}^{-|W_1-W_2|}\},\end{aligned}$$

其中 $c_0=(13^{\mathrm{e}}-3-40\mathrm{e}^{-2}+13\mathrm{e}^{-1})^{-1}$. 由文献 [9, 定理 2] 可知, $\rho^M(X,Y\mid Z)=0$ 当且仅当 $X\perp\!\!\!\perp Y\mid Z$. 度量是基于秩来构造的, 则对于任意单调变换 $m_1(\cdot)$、$m_2(\cdot)$ 和 $m_3(\cdot)$, 都有

$$\rho^M(X,Y\mid Z)=\rho^M\{m_1(X),m_2(Y)\mid m_3(Z)\}.$$

6.5 基于投影的条件相依性度量

令 $F_{\boldsymbol{y}|\boldsymbol{z}}(\boldsymbol{y}\mid\boldsymbol{z})$ 和 $F_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z}}(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{z})$ 分别为给定 $\boldsymbol{z}$ 和 $(\boldsymbol{x},\boldsymbol{z})$ 时 $\boldsymbol{y}$ 的条件分布函数. 定义误差过程 $\varepsilon(\boldsymbol{y})=I(\boldsymbol{y}\leqslant\boldsymbol{y})-F_{\boldsymbol{y}\mid\boldsymbol{z}}(\boldsymbol{y}\mid\boldsymbol{z})$, 其中 $I(\cdot)$ 表示示性函数. 条件独立性相当于 $F_{\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{z}}(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{z})=F_{\boldsymbol{y}\mid\boldsymbol{z}}(\boldsymbol{y}\mid\boldsymbol{z})$ 对所有的 $\boldsymbol{y}\in\mathbb{R}^q$ 几乎处处成立, 也意味着

$$\mathrm{pr}[\mathrm{E}\{\varepsilon(\boldsymbol{y})\mid\boldsymbol{x},\boldsymbol{z}\}=0]=1,\quad \text{对所有的 } \boldsymbol{y}\in\mathbb{R}^q \text{ 成立}.$$

文献 [102] 将上式中的 $(\boldsymbol{x},\boldsymbol{z})$ 投影到低维空间, 即 $\mathrm{E}\{\varepsilon(\boldsymbol{y})\mid\boldsymbol{x},\boldsymbol{z}\}=0$ 等价于研究

$$\mathrm{E}\{\varepsilon(\boldsymbol{y})I(\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{z}\leqslant v)\}=0,\quad \text{对所有的 } \boldsymbol{y}\in\mathbb{R}^q,\boldsymbol{\alpha}_1\in\mathbb{R}^p,\boldsymbol{\alpha}_2\in\mathbb{R}^d \text{ 和 } v\in\mathbb{R}^1 \text{ 成立}.$$

文献 [17] 也讨论了存在控制变量时, 充分降维的投影方向估计问题. 对所有可能的投影方向进行积分, 将正态分布密度函数作为权函数, 可以得到显式表达

$$\mathrm{E}\{\varepsilon(\boldsymbol{y})\widetilde{\varepsilon}(\boldsymbol{y})A(\boldsymbol{x},\widetilde{\boldsymbol{x}},\boldsymbol{z},\widetilde{\boldsymbol{z}})\}=0,\quad \text{对所有的 } \boldsymbol{y}\in\mathbb{R}^q \text{ 成立},$$

其中 $(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}},\widetilde{\boldsymbol{z}})$ 是一组 $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ 独立复制, $\widetilde{\varepsilon}(\boldsymbol{y})=I(\widetilde{\boldsymbol{y}}\leqslant\boldsymbol{y})-F_{\boldsymbol{y}\mid\boldsymbol{z}}(\boldsymbol{y}\mid\widetilde{\boldsymbol{z}})$, 并且

$$A(\boldsymbol{x},\widetilde{\boldsymbol{x}},\boldsymbol{z},\widetilde{\boldsymbol{z}})=\arcsin\left(\frac{1+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{x}}+\boldsymbol{z}^{\mathrm{T}}\widetilde{\boldsymbol{z}}}{\sqrt{1+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}}\sqrt{1+\widetilde{\boldsymbol{x}}^{\mathrm{T}}\widetilde{\boldsymbol{x}}+\widetilde{\boldsymbol{z}}^{\mathrm{T}}\widetilde{\boldsymbol{z}}}}\right).$$

基于此, 文献 [102] 采用了如下的准则度量条件独立性的偏离:

$$\mathrm{E}\{\varepsilon(\widetilde{\widetilde{\boldsymbol{y}}})\widetilde{\varepsilon}(\widetilde{\widetilde{\boldsymbol{y}}})A(\boldsymbol{x},\widetilde{\boldsymbol{x}},\boldsymbol{z},\widetilde{\boldsymbol{z}})\}.$$

7 关联度量的应用

7.1 高维数据变量筛选

随着科技的飞速发展, 数据收集与存储的能力大幅提升, 许多科学领域都需要从收集的大规模超高维数据中提取有效信息. 稀疏性假设认为大量的协变量, 往往只有极少数对响应变量有预测作用. 令 $\boldsymbol{x}=(X_1,\ldots,X_p)\in\mathbb{R}^p$ 是超高维协变量, Y 是响应变量. 一般而言, 高维数据特征筛选的目标是在尽可能剔除噪声协变量的同时, 保留所有重要的协变量.

高维特征筛选最早由文献 [19] 提出, 其基本思想为: 将每个协变量与响应变量的边际相依关系作为 p 个解释变量的边际效用, 将其从大到小排序, 效用最大的前 $|\mathcal{M}|$ 个解释变量被选入模型, 后 $p-|\mathcal{M}|$ 个协变量被剔除, 模型维数由 p 下降到 $|\mathcal{M}|$ ($|\mathcal{M}|$ 通常小于样本量).

不同相依性度量的选择产生了性质各不相同的特征筛选方法. 从最初基于 Pearson 相关系数筛选 [19], 到采用边际的秩相关系数 [41], 再到将条件相关系数应用到变系数模型筛选 [48] 中. 在不同的模型假设下, 采用特定的相依性度量, 但同时限制了方法的应用范围. 因此, 不依赖模型假设的特征筛选方法应运而生, 如基于距离相关系数 [44]、投影相关系数 [49,84]、MBKR 相关系数 [103]、球相关系数 [58]、鞅差相关系数 [66,83]、累积差异 [97] 和分位数相关系数 [50] 等. 在同时考虑控制变量 $\boldsymbol{z}$ 的作用时, 文献 [76,99,100] 等提供了有效的条件特征筛选方法.

7.2 模型检验

在数据分析中, 许多统计方法都是在特定的模型假设下完成的. 在实际应用时, 验证数据是否满足相应的模型假设必不可少. 非线性相依度量可以帮助验证模型假设的合理性, 其基本思想为: 在给定协变量 $\boldsymbol{x}$ 时, 研究在特定模型形式下估计出的残差均值是否为 0. 本质上, 这依然是一个关于均值的检验问题. 只不过需要获得有效的残差估计.

文献 [16] 基于投影提出了相合的模型检验方法, 文献 [85] 基于累积投影差异设计了单指标模型的检验方法. 在分位数回归框架下基于鞅差偏离系数, 文献 [81] 开发了模型检验的新方法. 在具有回归效应的参数模型下, 文献 [80, 82] 研究了残差和协变量间样本距离协方差及基于投影累积协方差的理论分布, 并基于该分布提出了新的模型检验方法和异方差检验方法. 文献 [93] 则是检验了是否某些 $\boldsymbol{x}$ 的线性组合足够刻画 Y 的条件分布.

8 结论

分析与研究数据之间相互影响、相互关联的关系具有重要的理论和应用意义. 本文介绍了度量和检验数据关联关系的一些前沿理论与方法. 在大数据时代, 随着数据的结构愈加复杂, 度量关联关系的方法必将更加丰富, 应用场景也将更加广阔.

参考文献

1 Anderson N H, Hall P, Titterington D M. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. J Multivariate Anal, 1994, 50: 41–54
2 Anderson T W. On the distribution of the two-sample Cramér-von Mises criterion. Ann of Math Stud, 1962, 33: 1148–1159
3 Baringhaus L, Franz C. On a new multivariate two-sample test. J Multivariate Anal, 2004, 88: 190–206
4 Bergsma W, Dassios A. A consistent test of independence based on a sign covariance related to Kendall's tau. Bernoulli, 2014, 20: 1006–1028
5 Berrett T B, Samworth R J. Nonparametric independence testing via mutual information. Biometrika, 2019, 106: 547–566
6 Biswas M, Ghosh A K. A nonparametric two-sample test applicable to high dimensional data. J Multivariate Anal, 2014, 123: 160–171
7 Biswas M, Mukhopadhyay M, Ghosh A K. A distribution-free two-sample run test applicable to high-dimensional data. Biometrika, 2014, 101: 913–926
8 Blum J R, Kiefer J, Rosenblatt M. Distribution free tests of independence based on the sample distribution function. Ann of Math Stud, 1961, 32: 485–498
9 Cai Z, Li, R, Zhang Y L. A distribution free conditional independence test with applications to causal discovery. J Mach Learn Res, 2022, 23: 3701–3741
10 Chatterjee S. A new coefficient of correlation. J Amer Statist Assoc, 2021, 116: 2009–2022
11 Chatterjee S. A survey of some recent developments in measures of association. Prob Stoch Process, 2024, in press
12 Cui H. Average projection type weighted Cramér-von Mises statistics for testing some distributions. Sci China Ser A, 2002, 45: 562–577
13 Deb N, Sen B. Multivariate rank-based distribution-free nonparametric testing using measure transportation. J Amer Statist Assoc, 2023, 118: 192–207
14 Dette H, Siburg K F, Stoimenov P A. A copula-based non-parametric measure of regression dependence. Scand J Stat, 2013, 40: 21–41
15 Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat, 2007, 1: 107–129
16 Escanciano J C. A consistent diagnostic test for regression models using projections. Econom Theory, 2006, 22: 1030–1051

17 Fan G, Zhu L. Sufficient dimension reduction in the presence of controlling variables. Sci China Math, 2022, 65: 1975–1996
18 Fan J, Feng Y, Xia L. A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. J Econometrics, 2020, 218: 119–139
19 Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol, 2008, 70: 849–911
20 Friedman J H, Rafsky L C. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann Statist, 1979, 7: 697–717
21 Gao L, Fan Y, Lv J, et al. Asymptotic distributions of high-dimensional distance correlation inference. Ann Statist, 2021, 49: 1999–2020
22 Gretton A, Borgwardt K M, Rasch M J, et al. A kernel two-sample test. J Mach Learn Res, 2012, 13: 723–773
23 Gretton A, Fukumizu K, Teo C H, et al. A kernel statistical test of independence. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2008, 585–592
24 Guo X, Li R, Liu W, et al. Stable correlation and robust feature screening. Sci China Math, 2022, 65: 153–168
25 Hall P, Marron J S, Neeman A. Geometric representation of high dimension, low sample size data. J R Stat Soc Ser B Stat Methodol, 2005, 67: 427–444
26 Hall P, Tajvidi N. Permutation tests for equality of distributions in high-dimensional settings. Biometrika, 2002, 89: 359–374
27 He D, Cheng J, Xu K. High-dimensional variable screening through kernel-based conditional mean dependence. J Statist Plann Inference, 2023, 224: 27–41
28 Heller R, Heller Y, Gorfine M. A consistent multivariate test of association based on ranks of distances. Biometrika, 2013, 100: 503–510
29 Hoeffding W. A non-parametric test of independence. Ann of Math Stud, 1948, 19: 546–557
30 Hotelling H. Relations between two sets of variates. Biometrika, 1936, 28: 321–377
31 Kendall M, Gibbons J D. Rank Correlation Methods. New York: Oxford Univ Press, 1990
32 Kendall M G. A new measure of rank correlation. Biometrika, 1938, 30: 81–93
33 Kim I, Balakrishnan S, Wasserman L. Robust multivariate nonparametric tests via projection averaging. Ann Statist, 2020, 48: 3417–3441
34 Kinney J B, Atwal G S. Equitability, mutual information, and the maximal information coefficient. Proc Natl Acad Sci USA, 2014, 111: 3354–3359
35 Kong E, Xia Y, Zhong W. Composite coefficient of determination and its application in ultrahigh dimensional variable screening. J Amer Statist Assoc, 2019, 114: 1740–1751
36 Kong Y, Li D, Fan Y, et al. Interaction pursuit in high-dimensional multi-response regression via distance correlation. Ann Statist, 2017, 45: 897–922
37 Lai T, Zhang Z, Wang Y. A kernel-based measure for conditional mean dependence. Comput Statist Data Anal, 2021, 160: 107246
38 Lawrance A. On conditional and partial correlation. Amer Statist, 1976, 30: 146–149
39 Lee C E, Zhang X, Shao X. Testing conditional mean independence for functional data. Biometrika, 2020, 107: 331–346
40 Li G, Li Y, Tsai C L. Quantile correlations and quantile autoregressive modeling. J Amer Statist Assoc, 2015, 110: 246–261
41 Li G, Peng H, Zhang J, et al. Robust rank correlation based screening. Ann Statist, 2012, 40: 1846–1877
42 Li L, Ke C, Yin X, et al. Generalized martingale difference divergence: Detecting conditional mean independence with applications in variable screening. Comput Statist Data Anal, 2023, 180: 107618
43 Li R Z, Xu K, Zhou Y Q, et al. Testing the effects of high-dimensional covariates via aggregating cumulative covariances. J Amer Statist Assoc, 2023, 118: 2184–2194
44 Li R Z, Zhong W, Zhu L P. Feature screening via distance correlation learning. J Amer Statist Assoc, 2012, 107: 1129–1139
45 Lin Z, Han F. Limit theorems of Chatterjee's rank correlation. arXiv:2204.08031, 2022
46 Lin Z, Han F. On boosting the power of Chatterjee's rank correlation. Biometrika, 2023, 110: 283–299
47 Lin Z, Han F. On the failure of the bootstrap for Chatterjee's rank correlation. Biometrika, 2024, in press
48 Liu J, Li R, Wu R. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. J Amer Statist Assoc, 2014, 109: 266–274
49 Liu W, Ke Y, Liu J, et al. Model-free feature screening and FDR control with knockoff features. J Amer Statist

Assoc, 2022, 117: 428–443
50 Ma S, Li R, Tsai C L. Variable screening via quantile partial correlation. J Amer Statist Assoc, 2017, 112: 650–663
51 Matteson D S, Tsay R S. Independent component analysis via distance covariance. J Amer Statist Assoc, 2017, 112: 623–637
52 Miao W, Liu C C, Geng Z. Statistical approaches for causal inference (in Chinese). Sci Sin Math, 2018, 48: 1753–1778 [苗旺, 刘春辰, 耿直. 因果推断的统计方法. 中国科学: 数学, 2018, 48: 1753–1778]
53 Mondal P K, Biswas M, Ghosh A K. On high dimensional two-sample tests based on nearest neighbors. J Multivariate Anal, 2015, 141: 168–178
54 Moon H, Chen K. Interpoint-ranking sign covariance for the test of independence. Biometrika, 2022, 109: 165–179
55 Moon Y I, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators. Phys Rev E, 1995, 52: 2318–2321
56 Newton M A, Quintana F A, den Boon J A, et al. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat, 2007, 1: 85–106
57 Pan W, Tian Y, Wang X, et al. Ball divergence: Nonparametric two sample test. Ann Statist, 2018, 46: 1109–1137
58 Pan W, Wang X, Xiao W, et al. A generic sure independence screening procedure. J Amer Statist Assoc, 2019, 114: 928–937
59 Pan W, Wang X, Zhang H, et al. Ball covariance: A generic measure of dependence in Banach space. J Amer Statist Assoc, 2020, 115: 307–317
60 Pearson K. Notes on regression and inheritance in the case of two parents. Proc R Soc Lond, 1895, 58: 240–242
61 Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London Edinburgh Dublin Philos Mag J Sci, 1900, 50: 157–175
62 Pearson K. Notes on the history of correlation. Biometrika, 1920, 13: 25–45
63 Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets. Science, 2011, 334: 1518–1524
64 Rosenblatt M. Limit theorems associated with variants of the von Mises statistic. Ann of Math Stud, 1952, 23: 617–623
65 Sejdinovic D, Sriperumbudur B, Gretton A, et al. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Ann Statist, 2013, 41: 2263–2291
66 Shao X, Zhang J. Martingale difference correlation and its use in high-dimensional variable screening. J Amer Statist Assoc, 2014, 109: 1302–1318
67 Shi H, Drton M, Han F. On the power of Chatterjee's rank correlation. Biometrika, 2022, 109: 317–333
68 Shi H, Drton M, Han F. Distribution-free consistent independence tests via center-outward ranks and signs. J Amer Statist Assoc, 2022, 117: 395–410
69 Shi H, Hallin M, Drton M, et al. On universally consistent and fully distribution-free rank tests of vector independence. Ann Statist, 2022, 50: 1933–1959
70 Smirnov N V. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Moscow Univ Math Bull, 1939, 2: 3–14
71 Spearman C. The proof and measurement of association between two things. Am J Psychol, 1904, 15: 72–101
72 Speed T. A correlation for the 21st century. Science, 2011, 334: 1502–1503
73 Strong S P, Koberle R, de Ruyter van Steveninck R R, et al. Entropy and information in neural spike trains. Phys Rev Lett, 1998, 80: 197–200
74 Subramanian A, Tamayo P, Mootha V K, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, 102: 15545–15550
75 Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. Ann Statist, 2007, 35: 2769–2794
76 Tong Z, Cai Z, Yang S, et al. Model-free conditional feature screening with FDR control. J Amer Statist Assoc, 2023, 118: 2575–2587
77 Wang X, Pan W, Hu W, et al. Conditional distance correlation. J Amer Statist Assoc, 2015, 110: 1726–1734
78 Wang X, Zhu J, Pan W, et al. Nonparametric statistical inference via metric distribution function in metric spaces. J Amer Statist Assoc, 2024, in press
79 Wilks S S. On the independence of k sets of normally distributed statistical variables. Econometrica, 1935, 3: 309–326
80 Xu K, Cao M. Distance-covariance-based tests for heteroscedasticity in nonlinear regressions. Sci China Math, 2021, 64: 2327–2356

81 Xu K, Chen F. Martingale-difference-divergence-based tests for goodness-of-fit in quantile models. J Statist Plann Inference, 2020, 207: 138–154

82 Xu K, He D. Omnibus model checks of linear assumptions through distance covariance. Statist Sinica, 2021, 31: 1055–1079

83 Xu K, Huang X. Conditional-quantile screening for ultrahigh-dimensional survival data via martingale difference correlation. Sci China Math, 2018, 61: 1907–1922

84 Xu K, Shen Z, Huang X, et al. Projection correlation between scalar and vector variables and its use in feature screening with multi-response data. J Stat Comput Simul, 2020, 90: 1923–1942

85 Xu K, Zhou Y Q. Projection-averaging-based cumulative covariance and its use in goodness-of-fit testing for single-index models. Comput Statist Data Anal, 2021, 164: 107301

86 Xu K, Zhou Y Q. Maximum-type tests for high-dimensional regression coefficients using Wilcoxon scores. J Statist Plann Inference, 2021, 211: 221–240

87 Xu K, Zhu L P. Nonparametric two-sample tests for equality of distributions using projections (in Chinese). Sci Sin Math, 2022, 52: 1183–1202 [许凯, 朱利平. 基于投影的两样本分布相等的非参数检验. 中国科学: 数学, 2022, 52: 1183–1202]

88 Xu K, Zhu L P. Power analysis of projection-pursuit independence tests. Statist Sinica, 2022, 32: 417–433

89 Yao S, Zhang X, Shao X. Testing mutual independence in high dimension via distance covariance. J R Stat Soc Ser B Stat Methodol, 2018, 80: 455–480

90 Ying C, Yu Z. Fréchet sufficient dimension reduction for random objects. Biometrika, 2022, 109: 975–992

91 Zhang X, Yao S, Shao X. Conditional mean and quantile dependence testing in high dimension. Ann Statist, 2018, 46: 219–246

92 Zhang Y L, Chen C Y, Zhu L P. Sliced independence test. Statist Sinica, 2022, 32: 2477–2496

93 Zhang Y L, Zhou Y Q, Zhu L P. A post-screening diagnostic study for ultrahigh dimensional data. J Econometrics, 2024, 239: 105354

94 Zhang Y L, Zhu L P. Projective independence tests in high dimensions: The curses and the cures. Biometrika, 2024, in press

95 Zhong P S, Chen S X. Tests for high-dimensional regression coefficients with factorial designs. J Amer Statist Assoc, 2011, 106: 260–274

96 Zhong W, Li Z, Guo W, et al. Semi-distance correlation and its applications. J Amer Statist Assoc, 2024, in press

97 Zhou T Y, Zhu L P, Xu C, et al. Model-free forward screening via cumulative divergence. J Amer Statist Assoc, 2020, 115: 1393–1405

98 Zhou W X, Zheng C, Zhang Z. Two-sample smooth tests for the equality of distributions. Bernoulli, 2017, 23: 951–989

99 Zhou Y Q, Liu J, Hao Z, et al. Model-free conditional feature screening with exposure variables. Stat Interface, 2019, 12: 239–251

100 Zhou Y Q, Liu J, Zhu L P. Test for conditional independence with application to conditional screening. J Multivariate Anal, 2020, 175: 104557

101 Zhou Y Q, Xu K, Zhu L P, et al. Rank-based indices for testing independence between two high-dimensional vectors. Ann Statist, 2024, 52: 184–206

102 Zhou Y Q, Zhang Y L, Zhu L P. A projective approach to conditional independence test for dependent processes. J Bus Econom Statist, 2022, 40: 398–407

103 Zhou Y Q, Zhu L P. Model-free feature screening for ultrahigh dimensional data through a modified Blum-Kiefer-Rosenblatt correlation. Statist Sinica, 2018, 28: 1351–1370

104 Zhu C B, Zhang X Y, Yao S, et al. Distance-based and RKHS-based dependence metrics in high dimension. Ann Statist, 2020, 48: 3366–3394

105 Zhu L P, Xu K, Li R, et al. Projection correlation between two random vectors. Biometrika, 2017, 104: 829–843

106 Zhu L P, Zhang Y W, Xu K. Measuring and testing for interval quantile dependence. Ann Statist, 2018, 46: 2683–2710

107 Zhu L X, Fang K T, Bhatti M I. On estimated projection pursuit-type Crámer-von Mises statistics. J Multivariate Anal, 1997, 63: 1–14

Association analysis for nonlinearly dependent data

Yeqing Zhou, Kai Xu & Liping Zhu

Abstract Measuring and testing the nonlinear dependence of complex data is a fundamental problem in statistics. Within the last century, important progress has been made in the methods and theories for analyzing nonlinear dependence. In this paper, we discuss three different types of dependency relationships: distributional dependence, mean dependence, and quantile dependence. We present important research advances achieved under these dependency relationships, extending from one-dimensional to multi-dimensional, and to high-dimensional settings. Furthermore, we discuss two applications of these metrics in high-dimensional feature screening and model checking.

Keywords **association analysis, correlation, nonlinearity, independence, conditional independence**

MSC(2020) **62H20, 62M10**

doi: 10.1360/SSM-2023-0175