实值多变量维数约简: 综述

单洪明 [1, 2] 张军平 [1, 2]

摘 要 维数约简作为机器学习的经典问题之一, 主要用于处理维数灾问题、帮助加速算法的计算效率和提高可解释性以及数据可视化. 传统的维数约简算法如主成分分析 (Principal component analysis, PCA) 和线性判别分析等只能处理无标签数据或者分类数据. 然而, 当预测变量为一元或多元连续型实值变量时, 这些处理无标签数据或分类数据的维数约简方法则不能形成有效的预测性能. 近 20 年来, 有一系列工作从多个角度对这一问题展开了研究, 并取得了系统性的研究成果. 在此背景下, 本文将综述这些面向回归问题的降维算法, 即实值多变量维数约简. 本文将介绍与实值多变量维数约简密切相关的基本概念、算法、理论, 并探讨一些潜在的研究方向.

关键词 维数约简, 维数灾难, 回归分析, 条件独立性, 互信息

引用格式 单洪明, 张军平. 实值多变量维数约简: 综述. 自动化学报, 2018, **44**(2): 193−215

DOI 10.16383/j.aas.2018.c160812

Real-valued Multivariate Dimension Reduction: A Survey

SHAN Hong-Ming[1, 2] ZHANG Jun-Ping[1, 2]

Abstract As one of the classical problems in machine learning, dimension reduction is used for dealing with the curse of dimensionality, speeding up computational efficiency of the algorithm, and improving interpretability as well as visualizing high-dimensional data. Traditional dimension reduction algorithms such as principal component analysis (PCA) and linear discriminant analysis are mainly suitable for unlabeled data or classification data. When the response variables are univariate or multivariate continuous real-valued ones, however, such dimension reduction methods cannot guarantee the effective predictive performance of the reduced subspace. In the recent two decades, researchers have been devoted to studying this issue with different viewpoints, attaining many promising and systemic achievements. Under this background, we will survey the developments of real-valued multivariate dimension reduction in detail. We will also introduce its basic concepts, algorithms and theories, and discuss some potential research directions deserving investigating.

Key words Dimension reduction, curse of dimensionality, regression analysis, conditional independence, mutual information

Citation Shan Hong-Ming, Zhang Jun-Ping. Real-valued multivariate dimension reduction: a survey. *Acta Automatica Sinica*, 2018, **44**(2): 193−215

在当今信息时代, 数据的获取变得越来越容易且经济. 由此产生的直接后果是, 数据呈现出高维度、大规模的特点. 要有效分析这些数据, 一个主要的途径是约简维数. 作为机器学习中的经典问题之一, 维数约简主要用于处理维数灾问题、帮助加速算法的计算效率和提高可解释性以及数据可视化. 在这一方向的多数研究侧重于无监督维数约简和有标签的监督维数约简, 试图发现高维数据集的内在低维结构, 而较少关注当存在连续型响应变量时, 高维数据集与响应变量 (集) 之间的关系.

当面对回归数据时, 如果直接采用无监督维数约简技术如主成分分析 (Principal component analysis, PCA)[1−2]、局部线性嵌入 (Locally linear embedding, LLE)[3]、等度规映射 (Isometric mapping, ISOMAP)[4] 等降维算法时, 则标签信息或响应变量不能得到充分利用, 性能也会随之受损. 另一方面, 监督维数约简包括线性判别分析[5−6] 以及近年来比较受关注的度量学习[7] 虽然其考虑了标签信息, 但将数据成组分到有限的类别中并不一定是合适的. 更合适的处理是考虑标签是光滑变化的响应, 即考虑响应变量处在实值多变量域 (Real-valued multivariate domain)[8]. 在这个前提下, 现有的很多监督、无监督维数约简技术都不见得是最优的, 而需要考虑回归意义下或实值多变量的维数约

收稿日期 2016-12-08 录用日期 2017-06-12
Manuscript received December 8, 2016; accepted June 12, 2017
国家自然科学基金 (61673118), 上海市浦江人才计划 (16PJD009) 资助
Supported by National Natural Science Foundation of China (61673118) and Shanghai Pujiang Program (16PJD009)
本文责任编委 朱军
Recommended by Associate Editor ZHU Jun
1. 上海智能信息处理重点实验室 上海 200433 2. 复旦大学计算机科学技术学院 上海 200433
1. Shanghai Key Laboratory of Intelligent Information Processing, Shanghai 200433 2. School of Computer Science, Fudan University, Shanghai 200433

简. 该研究方向在许多领域如生物认证中的人群计数[9−10]、人体姿态估计[8, 11]、智能交通[12]、生物信息[13−14] 和空间划分树[15] 中都有着广泛的应用.

实值多变量维数约简中的一个关键概念 — 充分维数约简 (Sufficient dimension reduction, SDR) — 正是在这一前提下发展起来的. 它与机器学习近年来的多个热点问题相互交织, 如流形学习[3−4]、回归预测[16]、多任务/多标记学习[17−18]、稀疏学习[19−20]、概率图模型[21−22] 等. 具体来讲, 实值多变量维数约简期望寻找一个条件独立的子空间, 其定义如下:

定义 1. 给定投影矩阵 B 的前提下, 使得响应变量 (集) Y 与自变量 X 条件独立, 表述如下式:

$$Y \perp\!\!\!\perp X \,|\, B^{\mathrm{T}}X \tag{1}$$

其中, 符号 "$\perp\!\!\!\perp$" 表示 Y 独立于 X. X 和 Y 均为多元的随机变量. 本文中的实值多变量维数约简是指响应变量 Y 为多元实值的情况. 当 Y 只有一元时, 对应的降维算法常被称为充分维数约简算法.

这种维数约简方法有两层含义: 1) 去掉 X 中与 Y 独立的一些特征, 也就是说这些特征对 Y 的预测不提供任何信息; 2) 从与 Y 相关的因素中寻找一个小的子空间或者子集, 不属于该子空间或者子集的特征虽然对 Y 的预测作出贡献, 但是在给定子空间或者子集后, 这部分信息不足以增加或改进对 Y 的预测性能.

通过实值多变量维数约简的定义可以看出, 与离散型 (分类) 维数约简算法最大的不同点就是响应变量为连续的实值. 若通过离散函数将连续变量离散化, 继而可使用离散型的维数约简算法如线性判别分析进行降维处理, 然而这种做法势必会导致信息的丢失. 相反, 若响应变量为离散型的标签时, 可以视为某个连续变量离散后的结果, 因此使用实值多变量维数约简进行降维处理. 由此可见, 实值多变量维数约简在维数约简中有着极其重要的研究价值. 简而言之, 当响应变量 Y 是范畴变量或分类变量时, 充分维数约简对应于分类意义下的维数约简; 当 Y 是连续值变量时, 对应于回归意义下的维数约简; 而当 Y 是多维变量时, 则可以对应于多元回归分析以及多任务和多标签学习. 由于一些近期研究的算法可以处理多元响应变量的类型, 本文将充分维数约简算法统称为实值多变量维数约简算法 (Real-valued multivariate dimension reduction).

本文将综述实值多变量维数约简的发展历史, 从技术手段和求解手段分析数十种经典的实值多变量维数约简算法, 并分析其优缺点. 最后讨论一些未来潜在的研究方向.

1 实值多变量维数约简的发展历史

实值多变量维数约简最初是针对回归问题来展开的. 回归分析是研究自变量与因变量之间的关系, 这类研究常会假定某个参数模型, 并通过标准的估计方法如最大似然或最小二乘方法来获得最优参数估计. 然而, 参数模型通常只是对真实内在结构的一个逼近, 寻找真实的、充分的模型并不容易. 当不存在这类参数模型时, 则常利用非参模型作为替代, 如考虑真实回归函数连续性或可微性的局部光滑方法. 该方法利用了一个几何直觉, 即在感兴趣点周围的样本点对空间结构的估计提供了足够的信息. 然而, 随着维数的增加, 由于维数灾的影响, 这类光滑技术要求的样本数也需要指数的增加. 因此, 这一策略在高维时不是很有效. 幸运地是, 近年来大量的机器学习相关的文献都表明, 高维数据存在低维结构. 因此, 寻找一个简单、有效且合理的低维表达是近 10 余年主要的努力方向之一. 充分维数约简正是希望在回归分析中找到一个线性变换满足式 (1) 条件独立性, 即研究在对高维数据降维后如何不影响对回归变量的预测性能. 以下, 我们将简要综述实值多变量维数约简的发展历史.

充分维数约简 (实值多变量维数约简的关键概念) 最早由 Li[23] 提出, 其采用逆切片回归技术将响应变量分块后, 再对切片后的数据进行主成分分析来获得回归约简子空间. 给定该子空间, 响应变量与高维数据集将条件独立, 相应的约简子空间称为有效维数约简子空间. Chiaromonte 等将该概念更进一步形式化, 称之为条件独立性, 相应的子空间称为中心子空间[24].

在近 20 余年中, 大量相关的工作致力于发现具有不同结构特性的维数约简子空间. 基于实值多变量维数约简算法不同的技术手段, 大部分主流算法可以归纳如图 1.

Li 等利用 Stein 引理, 有效地构造了主 Hessian 方向子空间[25]; Cook 等构造了二阶统计矩意义下的切片平均方差子空间[26]; Li 等提出了一个基于二阶矩且相对精确的方向回归子空间[27]. 由于该类算法基于前二阶矩估计量, 我们称为基于前二阶矩的降维算法.

随后, Cook 等又提出了一套基于模型的充分维数约简算法[28−30], 和前面基于统计阶矩不同的是, 该方法可以通过建立不同的统计模型使用最大似然估计子空间, 模型相对灵活. 我们称为基于模型的降维算法. 其代表算法为主拟合成分算法.

基于核独立分量分析的研究成果[31], Fukumizu 等将条件独立的概念推广至实值多变量的监督学习中[32]. 其从互信息角度出发, 构造了在重建核希尔

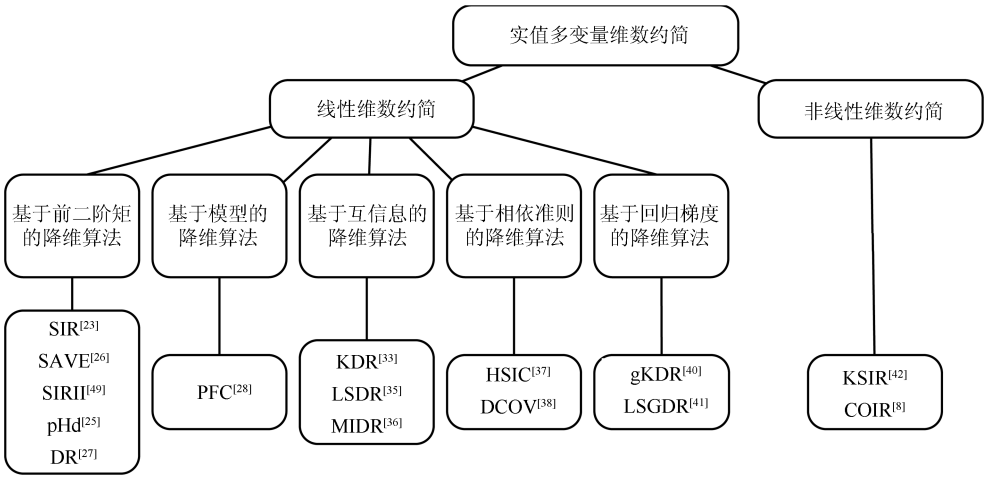


Fig. 1 实值多变量维数约简算法分类

Fig. 1 Taxonomy of real-valued multivariate dimension reduction algorithms

伯特空间的交叉协方差算子. 考虑到方法的可行性, 他们进一步将其问题的求解转化为半参意义下的优化问题. 其优势在于不需要考虑变量的边缘分布和变量间的条件概率分布. 文献 [33] 进一步从理论上对上述成果进行了更完善的证明. 随后, Suzuki 等采用最小平方互信息来度量条件独立性, 得到更为直观的条件独立性表达, 并且可以通过交叉验证得到最优参数[34]. Suzuki 等又进一步将解集从 Stiefel 流形推广到 Grassmann 流形[35]. 除此之外, Faivishevsky 等通过使用 k-近邻密度估计构造了互信息的统计量[36], 其目标函数简单、直观. 由于该类方法使用的估计量为互信息准则, 我们统称为基于互信息的降维算法.

Wang 等提出使用 Hilbert-Schmidt 独立准则进行子空间估计[37], 虽然从严格意义上来讲, 该类方法并不能实现条件独立性, 可是由于其算法方便求解, 也同样受到很多研究者的青睐. 此外, Sheng 等通过使用距离协方差独立性估计量来估计子空间[38]. 由于该类方法使用变量间的相依准则, 因此, 我们称之为基于相依准则的降维算法.

基于回归函数的梯度形式, Fukumizu 等提出了一个核维数约简的闭式解算法[39−40]. Sasaki 等通过最小平方梯度直接估计条件密度在对数形势下的梯度[41], 同样可以得到子空间估计的闭式解. 后者的一个优势在于带有交叉验证选择最优参数. 由于这类算法是考虑回归函数的梯度, 我们称之为基于回归梯度的降维算法.

前面提到的所有降维算法都是基于线性投影的, 虽然便于理解, 但是数据中一些非线性关系并不能很好的捕获. 因此, Wu 等通过核技巧 (Kernel trick) 提出了核化的切片逆回归算法[42]. Kim 等通过对重建希尔伯特空间下的协方差算子的形式进行分析, 建立了与条件均值子空间的关联, 并利用偏差方差技巧获得了闭式的条件均值子空间求解[8]. 我们称该类算法为基于非线性的降维算法.

除上面分类外, 还有一些工作是基于上述工作的推广. 譬如, 考虑到可以从特征集中提取稀疏特征, Li 等组合充分维数约简的回归技术和稀疏学习中的收缩估计策略, 来获得稀疏且精确的子空间[43]. Chen 等结合稀疏学习领域的 ℓ_1 范数技术, 实现了坐标独立的稀疏回归维数约简[44]. 值得指出的是, 这类稀疏技术主要在特征的线性组合意义下实现, 而未考虑非线性数据时如何实现. 假设数据集处在低维流形中, Nilsson 等将高维数据集首次通过拉普拉斯特征映射至一低维空间, 然后再利用核维数约简技术再次降维至具有条件独立性的更低维子空间[45]. Shyr 等通过定义时序核来获得时间序列数据中的中心子空间[46].

在众多降维算法的发展过程中, 中心子空间的概念一直贯穿其中. 其对降维算法的理解起着至关重要的作用. 接下来的一节, 我们介绍中心子空间的概念.

2 中心子空间简介

在充分维数约简发展初期, 人们只是致力于研究在回归中的维数约简和图形学. 尽管他们使用不同的模型, 但是人们对维数约简子空间 (Dimension reduction subspace, DRS) 并没有明确的定义. 所以, Cook 等[24, 47−48] 为了让人们更好地理解充分维数约简中的条件独立性, 对 DRS 进行了详细的研究.

本文中的常用数学符号定义如下: $X \in \mathbf{R}^D$ 表示 D 维的自变量; Z 表示自变量 X 的标准化; $Y \in \mathbf{R}^l$ 表示 l 维的响应变量; $B \in \mathbf{R}^{D\times d}$ 表示从高

维空间到低维空间的投影矩阵; $U = B^{\mathrm{T}}X$ 表示自变量 X 的低维表示; $\mathrm{E}(X)$ 和 $\mathrm{var}(X)$ 分别表示随机变量 X 的期望和协方差矩阵.

2.1 维数约简子空间

假定 $P_{\mathcal{S}}$ 是从 X 到维数约简子空间 $\mathcal{S}$ 的正交投影算子,

$$Y \perp\!\!\!\perp X \,|\, P_{\mathcal{S}}X \tag{2}$$

从式 (2) 可知, 原点 $\{\mathbf{0}\}$ 是 Y 在 X 上回归的 DRS, 当且仅当响应与预测向量是独立的. 在另一个极端, 任何回归允许至少一个明显的 DRS, 即空间自己. 所以, 很显然 DRS 具有不唯一性[24].

考虑早期实值多变量维数约简主要提供一个对自变量维数约简的框架来进行图形分析, Cook 提出了最小 DRS 的概念[47], 即对任意一个 DRS 空间都有 $\dim(\mathcal{S}_m) \le \dim(\mathcal{S}_{\mathrm{DRS}})$. 其中, dim 表示 DRS 的维数. 最小 DRS $\mathcal{S}_m$ 在回归图 (Regression plots) 中扮演很重要的角色, 然而, 要求最小维数并不一直足以选择唯一的子空间.

因为最小 DRS 可以提供最大的对自变量维数的约简, 因此, Cook 进一步提出中心子空间 (Central subspace, CS) 的概念[48]. Cook 认为如果全体维数约简子空间的交形成的子空间是维数约简子空间, 则该子空间为中心子空间, 简写为 $\mathcal{S}_{Y|X} = \cap\mathcal{S}_{\mathrm{DRS}}$. 显然中心子空间具有最小维数. 中心子空间与要求最小维数的条件 $\dim(\mathcal{S}_m) \le \dim(\mathcal{S}_{\mathrm{DRS}})$ 相比, 唯一性的实现用了一个更强的包含关系: $\mathcal{S}_{Y|X} \subseteq \mathcal{S}_{DRS}$. 注意, 尽管它是一个子空间, 但它不一定是一个维数约简子空间, 仅当其满足某些条件时才成立.

通过区分这三类子空间: 任意的 DRS、最小 DRS 和中心子空间, 可以更好地理解充分维数约简中的条件独立性. 在早期文献 [23] 中, 称之为有效 DRS. 有效 DRS 类似于其他三类子空间, 但没有进行显式地定义. 除此以外, 该 "模型" 表达在响应变量只有两类时会相对复杂, 一些引入的概念较难处理. 但文献 [23] 可以看成是统计维数约简研究的一个标志点. 对中心子空间的研究, 便于我们理解实值多变量维数约简算法.

在接下来一节中, 我们将介绍中心子空间的主要性质, 这会进一步加深对中心空间的理解.

2.2 中心子空间性质

在假定中心子空间存在的情况下, Chiaromonte 等研究了中心子空间的一些性质[24]. 通过这些性质, 我们可以对变量做适当的变换而不影响对中心子空间的估计.

性质 1. 如果 $\boldsymbol{a} \in \mathbf{R}^D$ 和 $A : \mathbf{R}^D \to \mathbf{R}^D$ 是一个满秩线性算子, 那么自变量的满秩仿射变换:

$$\mathcal{S}_{Y|\boldsymbol{a}+AX} = (A^{\mathrm{T}})^{-1}\mathcal{S}_{Y|X} \tag{3}$$

对自变量的满秩仿射变换不会影响中心子空间, 也不改变其结构维. 即中心子空间 $\mathcal{S}_{Y|X}$ 可以通过 $\mathcal{S}_{Y|\boldsymbol{a}+AX}$ 和线性变换得到. 最常见的应用就是对自变量 X 的标准化: 如果协方差矩阵 $\Sigma_{xx} = \mathrm{var}(X)$ 是正定的, 那么我们可以使用 $Z = \Sigma_{xx}^{-1/2}(X - \mathrm{E}(X))$ 来估计中心子空间, 并有 $\mathcal{S}_{Y|Z} = \Sigma_{xx}^{1/2}\mathcal{S}_{Y|X}$.

性质 2. 对任意的变换函数 ϕ, 响应变量的变换:

$$\mathcal{S}_{\phi(Y)|X} \subseteq \mathcal{S}_{Y|X}, \quad \phi \text{ 双射} \Rightarrow \mathcal{S}_{\phi(Y)|X} = \mathcal{S}_{Y|X} \tag{4}$$

对响应变量的任意变换, 不会导致对于中心子空间的过估计, 并在在双射变换的时候, 可以保证对中心子空间没影响. 在回归图形中, 一个很重要的推论就是使用严格单调的变换来改善绘图并不影响中心子空间. 其次, 对响应变量进行切片, 仍然可以保证子空间是包含在中心子空间中的. 因而, 可以通过多个侧面来了解中心子空间.

响应变量变换的另一个有效的应用是可以用于分析类别数据. 考虑下面一个二值变换, 给定某常数 c, 当 $Y > c$ 时, 有 $\widetilde{Y} = 1$; 否则, $\widetilde{Y} = 0$. 那么这个性质告诉我们, 可以通过研究 $\mathcal{S}_{\widetilde{Y}|X}$ 来重构 $\mathcal{S}_{Y|X}$ 的一部分. 这样做的好处就是当维度是 3 的时候, 可以通过 3D 二值响应画图 (3D binary response plots). 这个想法的扩展也非常直接, 我们把响应变量 Y 切成 h 片, 记为 $I_1, \cdots, I_h$, 并且当 $Y \in I_s$ 时, 定义其切片响应值 $\widetilde{Y} = s$. 那么我们依然有 $\mathcal{S}_{\widetilde{Y}|X} \subseteq \mathcal{S}_{Y|X}$. 值得注意的是, 基于切片响应值的方法经常被用到非图形方法中用来推断中心子空间.

性质 3. 把自变量投影到任意 DRS 上并不会影响中心子空间, 也就是说

$$Y \perp\!\!\!\perp X \,|\, P_{\mathcal{S}}X \Rightarrow \mathcal{S}_{Y|P_{\mathcal{S}}X} = \mathcal{S}_{Y|X} \tag{5}$$

这个概念就是条件独立性意义下的充分维数约简. 在这个条件下, 我们可以使用 $P_{\mathcal{S}}X$ 来取代 X 并保证不丢失信息. 这个性质还有一个深远的意义, 它为任何序列维数约简程序 (Sequential dimension reduction procedures) 提供了理论基础.

下面我们将介绍列在图 1 中的实值多变量维数约简算法.

3 基于前二阶矩的实值多变量维数约简算法

本节介绍的降维算法主要基于统计前二阶矩. 该类方法具有较高的统计可解释性, 但也需要一定

的统计假设才会成立. 下面将依次介绍 5 种经典的降维算法.

3.1 切片逆回归法

充分维数约简的研究始于 Li 提出的切片逆回归 (Sliced inverse regression, SIR)[23]. Li 认为, 当自变量的维数较高时, 在自变量 X 上对响应变量 Y 做回归会产生维数灾问题. 一种解决办法是在响应变量 Y 上对 X 做回归, 即逆回归 (Inverse regression, IR). 由于逆回归可以在 Y 上对 X 的每个维度做回归, 因而将高维问题转换成一维到一维的回归问题, 从而避免了维数灾问题. 其模型不需要进行参数的模型拟合, 具体形式如下:

$$y = f(\boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{b}_2^{\mathrm{T}}\boldsymbol{x}, \cdots, \boldsymbol{b}_d^{\mathrm{T}}\boldsymbol{x}, \epsilon) \tag{6}$$

其中, $\boldsymbol{b}_i$, $1 \le i \le d$, 是未知变量, ϵ 为与 X 无关的变量, 常表示测量误差. 其与非线性回归的最大区别是函数 $f(\cdot)$ 的形式完全未知. 要实现有效的降维, 则仅需要估计由 $\boldsymbol{b}_i$ 生成的有效维数约简 (Effective dimension reduction, EDR), 而通常的回归问题则需要估计全部的回归系数.

切片逆回归考虑当 y 变化时的逆回归曲线 $\mathrm{E}(\boldsymbol{x}|y)$. 逆回归曲线的中心在 $\mathrm{E}(\mathrm{E}(\boldsymbol{x}|y)) = \mathrm{E}(\boldsymbol{x})$ 处. 所以, 中心化的逆回归曲线 $\mathrm{E}(\boldsymbol{x}|y) - \mathrm{E}(\boldsymbol{x})$ 是空间 $\mathbf{R}^D$ 中的一条 D 维曲线. 切片逆回归算法通过条件 1 说明中心化的逆回归曲线是在一个 d 维子空间内.

条件 1. 给定任意 $\boldsymbol{b} \in \mathbf{R}^D$, 条件期望 $\mathrm{E}(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} \,|\, \boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{x}, \cdots, \boldsymbol{b}_d^{\mathrm{T}}\boldsymbol{x})$ 关于 $\boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{x}, \cdots, \boldsymbol{b}_d^{\mathrm{T}}\boldsymbol{x}$ 是线性的. 也就是说, 存在一些常数 $c_0, c_1, \cdots, c_d$, 使得 $\mathrm{E}(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} \,|\, \boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{x}, \cdots, \boldsymbol{b}_d^{\mathrm{T}}\boldsymbol{x}) = c_0 + c_1\boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{x} + \cdots + c_d\boldsymbol{b}_d^{\mathrm{T}}\boldsymbol{x}$.

这个条件成立的条件是自变量 X 服从椭圆对称分布, 例如, 高斯分布. 自变量 X 标准化至零均值和单位协方差, Li 将响应变量 Y 划分成若干个区间, 然而计算 X 的切片均值或逆回归曲线估计 $\mathrm{E}(\boldsymbol{x}|y)$[23]. 基于这些均值, 运用主成分分析来寻找最重要的 d 维子空间, 以获得了对有效维数约简方向的估计. SIR 的估计量可以写成 $\mathrm{var}(\mathrm{E}(\boldsymbol{x}|y))$.

这一方法的不足在于要保证逆回归曲线处于有效维数约简空间, 数据变量 X 需要满足椭圆对称分布 (例如高斯分布). 因而, 对于强非线性数据的非线性结构或主方向的捕捉是无效的. 另外, 尽管如何切片比多数回归方向中选择光滑参数的条件要弱, 但仍然具有一定的启发性, 对随后的性能有一定影响. 且逆切片技术则比较依赖于对目标空间的切片或聚类的过程, 为了能够形成对逆回归方差的估计, 结果使得这一类方法对目标的维数相对敏感. 其提出的维数估计准则仅能真实方向已知的情况下, 才可以进行有效评估. SIR 方法容易受到函数 f 关于 $\boldsymbol{x}$ 的对称性的影响. 举例来说, 如果 $y = x_1^2$, 这里 x_1 是 $\boldsymbol{x}$ 中的第一个坐标变量, 是关于均值对称的, 则逆回归曲线 $\mathrm{E}(\boldsymbol{x}|y)$ 会退化[1].

3.2 切片逆回归的推广

SIR 工作发表之后, 受到了许多学者的质疑. 譬如, 与经典的线性判别分析的关系、对数据椭圆对称分布的假设是否合理、在真实数据上效果如何等问题. 文献 [49] 对这些置疑进行了回复, 从直观上来看, 线性判别分析期望类内尽可能小, 类间尽可能大, 而没有考虑是否约简的子空间与原空间保留了相同的信息.

其中指出 SIR 最大的问题在于数据分布的椭圆对称分布的假设, 然而, 文献 [49] 也指出没有这个假设时, 仍然可以有合理的线性条件期望表示. 文章也指出这一算法在某些应用上取得了一定的成果, 如波士顿住房问题等[50].

为了解决 SIR 方法容易受到函数对称性的影响, 考虑函数对称的时候, 随着 y 的变化, $\mathrm{E}(\boldsymbol{x}|y)$ 恒等于 $\mathbf{0}$, 但逆方差 $\mathrm{var}(\boldsymbol{x} \,|\, y)$ 在每个切片内都是不同. 所以, 文献 [26] 提出使用二阶的逆方差函数去估计子空间的切片平均方差估计 (Sliced average variance estimation, SAVE). 假设数据 $\boldsymbol{x}$ 已经标准化, P_B 是到有效维数约子空间的投影算子. 那么, 对于二阶的逆方差存在如下等式[26]:

$$w_y I - \mathrm{var}(\boldsymbol{x}|y) = P_B\Big[w_y I - \mathrm{var}(\boldsymbol{x}|y)\Big]P_B \tag{7}$$

这里 w_y 是依赖于 X 特定椭圆对称分布的关于 Y 的函数, 例如 X 服从正态分布时, 对于所有的 Y, 有 $w_y = 1$. 所以, 文献 [26] 在每个切片内用 $I - \mathrm{var}(\boldsymbol{x}|y)$ 来作为二阶的逆方差计算, 并且从实验可以看出会带来一些更好的结果. 为了保证矩阵分解出来的特征值非负, 实际中 SAVE 使用的是估计量是 $\mathrm{E}[I - \mathrm{var}(\boldsymbol{x}|y)]^2$.

相比 SIR, 虽然 SAVE 可以对子空间提供更充分的估计, 可是它并没有 SIR 那么有效的检测线性趋势的能力[30].

回顾一下在 SIR 中, 逆回归曲线的中心点 $\mathrm{E}(\mathrm{E}(\boldsymbol{x}|y)) = \mathrm{E}(\boldsymbol{x})$ 对有效子空间的估计并不提供任何信息, 因此, 把变量 $\boldsymbol{x}$ 作标准化处理, 逆回归曲线的中心点随之变成零. 因此, 文献 [49] 中除了回复质疑和肯定 Cook 等的工作外[26], Li 提出了另一种基于二阶矩的方法, 称为 SIRII. SIRII 考虑去掉二阶矩逆回归曲线的中心点 $\mathrm{E}(\mathrm{var}(\boldsymbol{x} \,|\, y))$, 利用了方差分解公式来获得每个切片的二阶逆方差矩

[1] 因为 $y = x_1^2$, 所以有 $\mathrm{E}(x_1|y) = 0$, 这说明 y 变量本身显然没考虑到对称性这个问题, 因此会对估计逆回归曲线产生错误的导向.

$\mathrm{E}[\mathrm{var}(\boldsymbol{x}|y) - \mathrm{E}(\mathrm{var}(\boldsymbol{x}\,|\,y))]^2$.

虽然文献 [26, 49] 分别从不同角度使用二阶逆方差来解决 SIR 受函数对称的影响, 但是它们仍然需要 X 符合椭圆对称分布.

3.3 主 Hessian 方向

观察到回归函数的 Hessian 矩阵特征值可以帮助研究回归曲面的形状, Li 提出了主 Hessian 方向 (Principal Hessian direction, pHd) 来实现数据可视化和维数约简[25]. 并且发现当主 Hessian 方向在集成意义上, 可以用来定位沿回归曲面上具有最大曲率的方向, 因而可以有助于寻找 DRS 方向.

Li 观察到回归函数 $f(\boldsymbol{x}) = \mathrm{E}(y|\boldsymbol{x})$ 在任意点 $\boldsymbol{x}$ 处的 Hessian 矩阵 H_x, $[\frac{\partial^2}{\partial \boldsymbol{x}^{(i)} \partial \boldsymbol{x}^{(j)}} f(\boldsymbol{x})]$, 会沿与 B 正交的任意方向退化. 基于这一观察, Li 提出了辨识 EDR 空间的主 Hessian 方向的概念[25]. 由于 Hessian 矩阵 H_x 会随着 $\boldsymbol{x}$ 的变化而变化, 除非 $f(\boldsymbol{x})$ 曲面是二次的. 所以, 作者提出使用平均的 Hessian 矩阵, $\overline{H} = \mathrm{E}H_x$, 来表示数据整体的结构.

pHd 形成了一个可以用来可视化 y 与 $\boldsymbol{x}$ 关系的新坐标系统, 其特点是回归函数沿坐标轴的曲率是依次最大. 并且利用著名的 Stein 引理[51] 避免对随机变量函数的求导. 因此, Hessian 矩阵 $\overline{H}$ 可以通过 $\overline{H} = \Sigma_{xx}^{-1}\Sigma_{yxx}\Sigma_{xx}^{-1}$ 关联到加权协方差矩阵 $\Sigma_{yxx} = \mathrm{E}(y - \mathrm{E}y)(\boldsymbol{x} - \mathrm{E}\boldsymbol{x})(\boldsymbol{x} - \mathrm{E}\boldsymbol{x})^{\mathrm{T}}$ 来获得. EDR 的方向 $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_d$ 可以通过 $\overline{H}\Sigma_{xx}$ 的最大主成分构成, 或者通过下面广义特征值的形式

$$\Sigma_{yxx}\boldsymbol{b}_i = \lambda_i \Sigma_{xx}\boldsymbol{b}_i,\ |\lambda_1| \geq \cdots \geq |\lambda_d|,\ 1 \leq i \leq d \quad (8)$$

同时观察到对响应变量 y 加上或者减去一个关于 $\boldsymbol{x}$ 的线性函数, 并不影响 Hessian 矩阵[2] 所以, 相对于上述基于 y 的 pHd, 文献 [25] 又提出使用 y 对 $\boldsymbol{x}$ 的最小平方拟合后的残差 r 来替代 y 的基于残差的 pHd, 即使用 $\Sigma_{rxx} = \mathrm{E}r(\boldsymbol{x} - \mathrm{E}\boldsymbol{x})(\boldsymbol{x} - \mathrm{E}\boldsymbol{x})^{\mathrm{T}}$. 和前者相比, 基于残差 r 的 pHd 算法可以更有效地发现真实的成分, 因为残差 r 的方差要比 y 的方差小很多. 此外, 作者进一步将 pHd 推广到多项式的情况: 如果 $\boldsymbol{x}$ 符合正态分布, 且在 y 对 $\boldsymbol{x}$ 的最小平方拟合多项式次数大于 1 时, 拟合多项式的平均 Hessian 矩阵和 y 的平均 Hessian 矩阵是相同的.

因为 pHd 具有发现回归曲面曲率的能力, Li 等进一步提出了树形 pHd 算法来挖掘数据中的几何信息[52]. 但因为 Stein 引理成立的前提是要求数据符合正态分布, 所以, pHd 没有摆脱对数据分布的假设. 如果我们使用一个足够大的泛函空间 (例如重建核希尔伯特空间), 对于任意分布的数据, 都可以近似满足 Stein 定理, 那么 pHd 方法就可以得到更广的应用. 同时, 如果响应变量 y 的取值过于离散化, 那么 pHd 将不能获得有效的维数约简子空间. 而鉴于 pHd 可以有效获得回归曲面的几何信息, 如果我们可以找到一个很好打分函数 (Score function), 将离散变量转换成回归变量, 那么 pHd 也可以对离散变量进行有效处理. 文献 [25] 还指出, pHd 可作为一阶矩算法 SIR 的互补算法, 一方面, SIR 比 pHd 更倾向于稳定, 因为后者使用了二阶矩; 另一方面, pHd 可以解决 SIR 无法解决的对称函数问题.

3.4 方向回归算法

基于前两阶矩的统计量 SIR 和 SAVE 是被经常使用的维数约简估计量. 然而, 按照我们前面讨论的, 它们分别有自己的局限. SIR 不能处理响应变量关于原点对称的问题, SAVE 不能有效地估计单调趋势. 像上面提到的 pHd 和 SIRII 也有类似的困难. 因此, 一些作者提出通过凸组合这些前两阶矩估计量来克服这些困难, 例如 SIR 和 SIRII 的凸组合[49, 53], SIR 和 pHd 的凸组合[54], SIR 和 SAVE 的凸组合[54]. 此外, 文献 [54] 还提出一种挑选凸组合方法系数的自助法. 值得关注的是文献 [27] 介绍了一种简单自然的维数约简准则—方向回归算法 (Directional regression, DR), 它通过前两阶矩估计量合成的降维方法可以在准确率上实现本质的提升.

方向回归算法的想法主要来源于经验方向 (Empirical directions). 如果我们有来自联合分布 (X, Y) 的观测样本 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, 那么文献 [55] 把经验方向定义为一组向量的集合: $\{\boldsymbol{x}_i - \boldsymbol{x}_j : 1 \leq i < j \leq n\}$, 并且认为经验方向包含了样本中 X 的所有的方向信息. DR 可以非常有效地使用经验方向, 并且是直接用经验方向对响应变量作回归. 这种方法不仅可以实现 $\sqrt{n}$ 的收敛率, 还可以对中心子空间实现更为充分的估计. DR 的优势在于准确率高, 计算代价低, 而且只是前两阶矩估计量的组合. 下面介绍一下 DR 方法的实现过程.

给定随机变量 X, 我们使用和前面一样标准化的随机变量 Z, 即 $Z = \Sigma_{xx}^{-1/2}(X - \mathrm{E}X)$. 假设 (Z', Y') 是 (Z, Y) 的独立副本, 那么定义下式

$$A(Y, Y') = \mathrm{E}[(Z - Z')(Z - Z')^{\mathrm{T}}\,|\,Y, Y'] \quad (9)$$

式 (9) 是将 $Z - Z'$ 的方向, $(Z - Z')(Z - Z')^{\mathrm{T}}$, 和 (Y, Y') 的函数作回归. 直观上来说, 那些和中心子空间 $\mathcal{S}_{Y|Z}$ 对齐的 $Z - Z'$ 方向明显受到 Y 的影响, 而那些和子空间 $\mathcal{S}_{Y|Z}^{\perp}$ 对齐的 $Z - Z'$ 方向将不受 Y 影响. 因此可以通过那些和中心子空间 $\mathcal{S}_{Y|X}$ 相一致的 $Z - Z'$ 方向来估计 $\mathcal{S}_{Y|Z}$. 下面定理给出在适当条件下, 对任意给定的 $(Y, Y') = (y, y')$, $2I - A(Y, Y')$

[2] 因为关于 $\boldsymbol{x}$ 的任意线性组合, 其二阶偏导恒为 0.

的列空间是包含于 $\mathcal{S}_{Y|Z}$ 内的.

定理 1[27]**.** 假定任意的向量 $\boldsymbol{v} \in \mathbf{R}^D$ 且 $\boldsymbol{v} \perp\!\!\!\perp \mathcal{S}_{Y|Z}$, 如果

1) $\mathrm{E}(\boldsymbol{v}^{\mathrm{T}}Z|P_BX)$ 是关于 Z 的线性函数;

2) $\mathrm{var}(\boldsymbol{v}^{\mathrm{T}}Z|P_BX)$ 是非随机数.

那么, 对任意 (Y,Y'), $2I-A(Y,Y')$ 的列空间是包含于 $\mathcal{S}_{Y|Z}$ 内的.

上述定理表明我们可以通过目标 $\mathrm{E}[2I-A(Y,Y')]^2$ 来估计中心子空间 $\mathcal{S}_{Y|Z}$. 文献 [27] 将 DR 的估计量中的 (Z,Y) 和 (Z',Y') 合并后重写得到下式的估计量

$$\begin{aligned}&2\mathrm{E}\big[\mathrm{E}^2(ZZ^{\mathrm{T}}|Y)\big]+2\mathrm{E}^2\big[\mathrm{E}(Z|Y)\mathrm{E}(Z^{\mathrm{T}}|Y)\big]+\\&2\mathrm{E}\big[\mathrm{E}(Z^{\mathrm{T}}|Y)\mathrm{E}(Z|Y)\big]\mathrm{E}\big[\mathrm{E}(Z|Y)\mathrm{E}(Z^{\mathrm{T}}|Y)\big]-2I\end{aligned} \tag{10}$$

上式中的估计量是关于给定 Y, Z 的条件矩的非线性函数. 并且, 相对之前的凸组合方法, 该估计量只有 $\mathrm{O}(n)$ 的计算复杂度.

上面我们介绍的所有方法都是基于前二阶矩的方法, 我们统称为基于前二阶矩 (Frist two moment, F2M) 的充分维数约简降维算法[30]. 在同等温和的条件下, DR 可以比前面介绍的维数约简方法 (SIR、SAVE、SIRII、pHd 等) 更充分地估计中心子空间 $S_{Y|Z}$. 并且文献 [27] 总结了在实际应用上, 除了回归曲面非常震动被迫使用更高阶矩的统计量外, 该类方法可以实现较优异的性能.

4　基于模型的实值多变量维数约简算法

相比之前基于原始空间统计阶矩的维数约简方法, Cook 等将主成分 (Principal component, PC) 的概念引入到回归问题中, 并提出了一个基于最大似然估计的充分维数约简方法[28−30]. 和 SIR、SAVE、DR、pHd 等方法同等温和条件下相比, 基于最大似然的方法可以实现对子空间 $\mathcal{S}_{Y|X}$ 更详尽的估计. 基于最大似然的维数约简方法是通过基于模型的逆回归中获得对正回归 X 对 Y 的约简信息.

4.1　主成分回归模型

假设我们有 n 对从联合概率 (X,Y) 中独立同分布采样的观测样本 $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$, 以及假设 $X|Y\sim\mathrm{N}(\boldsymbol{\mu}_y,\Delta_y)$, 那么我们有:

$$\boldsymbol{x}=\boldsymbol{\mu}_y+\boldsymbol{\epsilon} \tag{11}$$

这里的 $\boldsymbol{\epsilon}\sim\mathrm{N}(\mathbf{0},\Delta_y)$ 且满足 $\Delta_y>0$. 那么, 我们的目标就是找到中心子空间 $\mathcal{S}_{Y|X}$ 的最大似然估计量. 如果 $d=\dim(\mathcal{S}_{Y|X})$, 以及 B 在 Stiefel 流形上[3], 即满足 $B^{\mathrm{T}}B=I_d$, 并且 B 的列向量是组成中心子空间的一组基, 显然我们有 $Y|X\sim Y|B^{\mathrm{T}}X$. 基于最大似然的方法就是在 $\boldsymbol{\mu}_y$ 和 Δ_y 不同的结构下来估计中心子空间 $\mathcal{S}_{Y|X}=\mathrm{span}(B)$.

在介绍基于最大似然的充分维数约简前, 我们先介绍文献 [28] 是如何将主成分引入到回归问题中的. 首先, 考虑下面逆回归模型

$$\boldsymbol{x}_y=\boldsymbol{\mu}+B\boldsymbol{v}_y+\delta\boldsymbol{\epsilon} \tag{12}$$

这里使用下标 y 是为了表示其是条件模型. 坐标向量 $\boldsymbol{v}_y\in\mathbf{R}^d$ 是一个关于 y 函数, 并且假设 $\boldsymbol{v}_y$ 有正定协方差矩阵, 和满足均值为 $\mathbf{0}$, 即 $\sum_y\boldsymbol{v}_y=\mathbf{0}$. 对 $\boldsymbol{v}_y$ 的中心化只是为了后面描述方便, 并不是必须满足的约束. 误差向量满足 $\boldsymbol{\epsilon}\sim\mathrm{N}(\mathbf{0},I)$. 模型 (12) 描述的是通过截距 $\boldsymbol{\mu}$ 变换后, 条件均值会落在由 B 列向量展开的 d 维子空间内. 坐标向量 $\boldsymbol{v}_y$ 则是变换后的条件均值 $\mathrm{E}(\boldsymbol{x}_y)-\boldsymbol{\mu}$ 在 B 下的坐标. 即便 d 非常小, 哪怕只有 3 或者 4, 这里的均值函数相当灵活[28]. 在模型 (12) 的右端, 除了下标 y 外, 其他都是未知的.

下面的命题将逆回归模型 (12) 和 X 对 Y 的正回归联系起来:

命题 1[28]**.** 在逆回归模型 (12) 下, 对 X 的所有取值, $Y|X$ 和 $Y|B^{\mathrm{T}}X$ 的分布一样.

所以, 按照上述命题, 我们可以使用 X 的充分约简变量 $B^{\mathrm{T}}X$ 来取代 X 本身, 并且不会对 Y 的回归预测造成信息损失. 对模型 (12) 中的中心子空间估计需要借助最大似然.

在进行最大似然估计前, 我们先研究一下这个子空间的性质: 1) 半正交矩阵 B 是在 Stiefel 流形上, 即满足约束 $B^{\mathrm{T}}B=I_d$; 2) 对任意的正交变换 $O\in\mathbf{R}^{d\times d}$, 子空间保持不变. 即 $(BO)^{\mathrm{T}}(BO)=O^{\mathrm{T}}O=I_d$.

虽然性质 2 表明子空间不变, 但是坐标向量会改变, 即 $B\boldsymbol{v}_y=(BO)(O^{\mathrm{T}}\boldsymbol{v}_y)$. 因此, 我们不可以同时对 B 和 $\boldsymbol{v}_y$ 进行估计. 所以, 模型 (12) 的最大似然估计第一步是固定 B 和 δ^2, 对 $\boldsymbol{v}_y$ 和 $\boldsymbol{\mu}$ 估计; 第二步是把估计得到的 $\boldsymbol{v}_y$ 和 $\boldsymbol{\mu}$ 代入, 估计 B 和 δ^2.

首先, 我们把 (B,δ^2) 固定在 (G,s^2) 处时, 模型 (12) 变成

$$\boldsymbol{x}_y=\boldsymbol{\mu}+G\boldsymbol{v}_y+s\boldsymbol{\epsilon}\sim\mathrm{N}\Big(\boldsymbol{\mu},\,G\boldsymbol{v}_y\boldsymbol{v}_y^{\mathrm{T}}G^{\mathrm{T}}+s^2I\Big) \tag{13}$$

对上述概率分布进行最大似然估计后, 可以得到 $\widehat{\boldsymbol{\mu}}=\bar{\boldsymbol{x}}$ 和在 y 缺失的情况下, $\boldsymbol{v}_y(G,s^2)=G^{\mathrm{T}}(\boldsymbol{x}_y-\bar{\boldsymbol{x}})$ 是一个关于 (G,s^2) 的函数. 其次, 把估

[3]注: 原文中设定为 Grassmann 流形. 当假定矩阵 B 满足 $B^{\mathrm{T}}B=I_d$ 时, B 展成的子空间和 Grassmann 流形上定义的子空间等价. 由于后文中的算法都是基于 Stiefel 流形讨论, 所以, 在不影响问题描述的情况下, 本文统一成 Stiefel 流形.

计得到的 $\widehat{\boldsymbol{\mu}}$ 和 $\boldsymbol{v}_y(G, s^2)$ 代入模型 (12), 得到下式

$$\boldsymbol{x}_y = \bar{\boldsymbol{x}} + GG^{\mathrm{T}}(\boldsymbol{x}_y - \bar{\boldsymbol{x}}) + s\boldsymbol{\epsilon} \tag{14}$$

整理可得

$$\boldsymbol{x}_y = \bar{\boldsymbol{x}} + s(I - GG^{\mathrm{T}})^{-1}\boldsymbol{\epsilon} \sim \mathrm{N}\Big(\bar{\boldsymbol{x}}, s^2(I - GG^{\mathrm{T}})^{-1}\Big) \tag{15}$$

对上式的对数最大似然估计可得:

$$\mathcal{L}(G, s^2) = -\frac{1}{2}nD\log(s^2) - \frac{n}{2s^2}\mathrm{tr}(\Sigma_{xx}Q_G) + \mathrm{const} \tag{16}$$

其中, $P_G = GG^{\mathrm{T}}$, $Q_G = I - P_G$. 虽然这里我们用 G 作为参数, 但是函数本身只依赖于 $\mathrm{span}(G)$, 并且其最大值是在 Stiefel 流形上面. 假设 $(\lambda_i, \boldsymbol{\gamma}_i)$ 分别是 Σ_{xx} 的特征值和特征向量, 并且满足 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$. 因此, 最大似然估计得到的中心子空间 $\mathcal{S}_{Y|X} = \mathrm{span}(\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_d)$, 并且对 σ^2 的估计是 $\sigma^2 = \frac{1}{D}\sum_{j=d+1}^{D}\lambda_j$. 模型 (12) 被称为**主成分回归模型**. 在响应变量 Y 缺失的情况下, 模型中的 Y 是扮演一个非显示条件参数的形式. 而一旦我们有响应变量的, 我们就可以为响应变量 "定制" 主成分.

4.2 主拟合成分

模型 (12) 的不足之处在于没有充分考虑响应变量信息, 而且最求解和主成分分析等价. 模型 (12) 和概率主成分分析模型[56] 本质上是一样, 只是模型 (12) 的出发点是考虑的回归问题, 而不是无监督的维数约简.

当我们观测到响应变量 Y 后, 可以通过对 $\boldsymbol{v}_y$ 的构建不同的模型来将主成分适配响应变量, 这就是我们接下来介绍的**主拟合成分模型**. 假设 $\boldsymbol{v}_y = \boldsymbol{\beta}\{\boldsymbol{f}_y - \mathrm{E}\boldsymbol{f}_y\}$, 其中的 $\boldsymbol{f}_y \in \mathbf{R}^r$ 为已知的关于 Y 的向量, $\beta \in \mathbf{R}^{d\times r}$, 显然坐标向量 $\boldsymbol{v}_y$ 是由 $\boldsymbol{f}_y$ 的多元线性回归得到. 因此, 主成分拟合模型可以写成

$$\boldsymbol{x}_y = \boldsymbol{\mu} + B\beta\{\boldsymbol{f}_y - \mathrm{E}\boldsymbol{f}_y\} + \delta\boldsymbol{\epsilon} \tag{17}$$

Cook 等提到了三种构造 $\boldsymbol{f}_y$ 的方法[28−29]:

1) 多项式法

$$\boldsymbol{f}_y = [y, y^2, \cdots, y^r]^{\mathrm{T}} \in \mathbf{R}^r$$

2) 傅里叶变换法

$$\boldsymbol{f}_y = [\cos(2\pi y), \sin(2\pi y), \cdots, \cos(2\pi k y), \sin(2\pi k y)]^{\mathrm{T}} \in \mathbf{R}^{2k} \tag{18}$$

3) 切片法

按照 y 的数值切成 r 块, 分别为 $I_1, \cdots, I_r$, 那么 $\boldsymbol{f}_y$ 中的第 k 个元素值为

$$f_{yk} = \delta_k(y) - \frac{n_k}{n}$$

其中, $\delta_h(y)$ 为指示 y 是否在 I_k 里面. 如果 $y \in I_k$, 那么 $\delta_k(y) = 1$; 否则, $\delta_k(y) = 0$. n_k 表示切片 I_k 里面的元素个数. 显然 $\mathrm{E}\boldsymbol{f}_y = \mathbf{0}$.

下面介绍如何使用最大似然估计模型 (17) 中的中心子空间 $\mathcal{S}_{Y|X} = \mathrm{span}(B)$. 为了方便讨论, 我们定义 X 为 $n \times D$ 的矩阵, 每行为 $(\boldsymbol{x}_y - \bar{\boldsymbol{x}})^{\mathrm{T}}$. 定义 F 为 $n \times r$ 的矩阵, 每行为 $(\boldsymbol{f}_y - \mathrm{E}\boldsymbol{f}_y)^{\mathrm{T}}$. 那么, 当我们固定 (B, σ) 在 (G, s^2) 处时, 对 $\boldsymbol{\mu}$ 和 β 最大似然估计得到 $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$, $\widehat{\beta} = G^{\mathrm{T}}X^{\mathrm{T}}F(F^{\mathrm{T}}F)^{-1}$. 把估计得到的 $\widehat{\boldsymbol{\mu}}$ 和 $\widehat{\beta}$ 代入模型 (17), 可以得到如下对数最大似然估计

$$\mathcal{L}(G, s^2) = -\frac{nD}{2}\log(s^2) - \frac{n}{2s^2}\{\mathrm{tr}[\Sigma_{xx}] - \mathrm{tr}[P_G\Sigma_{fit}]\} + \mathrm{const} \tag{19}$$

定义矩阵 $\widehat{X}$ 是从 $\boldsymbol{f}_y$ 到 X 的中心化的多元拟和值, $\widehat{X} = P_F X$, 那么 Σ_{fit} 是多元拟和值的协方差矩阵, 即 $\Sigma_{fit} = \frac{1}{n}\widehat{X}^{\mathrm{T}}\widehat{X}$. 同样的, 最大似然函数 (19) 只依赖于子空间 $\mathrm{span}(G)$. 这里的 $\mathrm{rank}(\Sigma_{fit}) \leq r$, 通常等号是成立的. 并且假设 $\mathrm{rank}(\Sigma_{fit}) \geq d$. $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_d$ 是矩阵 Σ_{fit} 的最大的 d 的特征值 λ_i 对应的特征向量, 那么最大似然函数 (19) 的最大值是在 $\mathrm{span}(G) = \{\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_d\}$ 处取得.

文章把 $\boldsymbol{\phi}_1^{\mathrm{T}}X, \cdots, \boldsymbol{\phi}_D^{\mathrm{T}}X$ 称为主拟合成分. 和模型 (12) 不同的是, 模型 (17) 中估计子空间的协方差矩阵是通过 $\boldsymbol{f}_y$ 对 X 拟合出来的向量构成的, 因此, 该模型通过主拟合成分充分考虑和响应变量之间的关系. 除此之外, 文献 [28−29] 还分析了和 SIR、PLS、OLS 之间的关系. 通过对噪音协方差矩阵强加不同的结构, 可以使 PFC 模型更灵活. 文献 [57] 提供了一个基于最大似然估计的充分维数约简软件包, 可以供读者进行实验.

虽然该类方法可以用不同的方式建模, 灵活多变, 可如果建立的模型与数据真实分布差异较大, 那么会导致对中心子空间的错误估计.

5 基于互信息的实值多变量维数约简算法

条件独立性的假设 (1) 是建立在空间的概率分布上, 即有下式[32]:

$$p(Y|X) = p(Y|P_B X) \tag{20}$$

其中, P_B 是从 $\mathbf{R}^D$ 到中心子空间 $\mathcal{S}_{Y|X}$ 的正交投影.

若令 (B, C) 是 D-维正交矩阵使得 B 的列向量张成子空间 $\mathcal{S}_{Y|X}$, 即 $\mathcal{S}_{Y|X} = \mathrm{span}(B)$, 定义

$U = B^{\mathrm{T}}X$ 和 $V = C^{\mathrm{T}}X$. 因为 (B, C) 是正交矩阵, 所以有以下概率密度函数[32]:

$$p(X) = p(U, V), \quad p(X, Y) = p(U, V, Y) \tag{21}$$

同时, 条件独立性可等价地表示为[32]

$$p(Y|U, V) = p(Y|U) \tag{22}$$

其式表示中心子空间 $\mathcal{S}_{Y|X}$ 是一个给定 U 后, 随机变量 Y 和 V 条件独立的空间.

互信息提供了关于条件独立性和中心子空间存在性之间等价性的另一种视角[32]:

$$\begin{aligned}\mathbb{I}(Y; X) = \mathbb{I}(Y; U, V) = \mathrm{E}_{X,Y}\left[\log\frac{p(U, V, Y)}{p(U, V)p(Y)}\right] = \\ \mathbb{I}(Y; U) + \mathrm{E}_{Y,V}\left[\log\frac{p(Y, V|U)}{p(Y|U)p(V|U)}\right] = \\ \mathbb{I}(Y; U) + \mathrm{E}_U\left[\mathbb{I}(Y|U; V|U)\right]\end{aligned} \tag{23}$$

其中, 互信息 $\mathbb{I}(X; Y)$ 定义为[32]

$$\mathbb{I}(X; Y) = \iint p(X, Y)\log\frac{p(X, Y)}{p(X)p(Y)})\mathrm{d}X\mathrm{d}Y \tag{24}$$

如果满足条件独立性, 那么就有 $\mathbb{I}(Y; X) = \mathbb{I}(Y; U)$ 或者 $\mathrm{E}_U\left[\mathbb{I}(Y|U; V|U)\right] = 0$. 因此, 条件独立性的问题可以转化为最优化的问题, 也就说最大化公式 (23) 右端的第一项或者最小化右端的第二项. 本节介绍三种基于互信息的充分维数约简算法, 方法总结见表 1.

5.1 核维数约简

基于在核独立分量分析的研究成果[31], Fukumizu 等将条件独立的概念推广到监督学习中, 并通过构造在重建核希尔伯特空间上的协方差算子, 以形成条件独立性的非参形式化描述, 从而将该问题转化为优化问题, 称为核维数约简 (Kernel dimension reduction, KDR)[32]. 与其他监督学习的维数约简方法不同, 提出的方法既不要求 X 的边缘分布假设, 也没有 Y 的条件分布的参数模型假设.

Fukumizu 等[32] 使用重建核希尔伯特空间 (Reproducing kernel Hilbert spaces, RKHSs) 的交叉协方差算子 (Corss-covariance operator) 来获得维数约简的目标函数. Fukumizu 首先将定义在文献 [58] 中的交叉协方差算子从希尔伯特空间推广到重建核希尔伯特空间. 如果我们用 $(\mathcal{H}_x, \kappa_x)$ 和 $(\mathcal{H}_y, \kappa_y)$ 分别表示在测度空间 $(\Omega_x, \mathcal{B}_x)$ 和 $(\Omega_y, \mathcal{B}_y)$ 上由核函数 κ_x 和 κ_y 诱导出的重建核希尔伯特空间. 重建性质的意思是对于任意函数 $f \in \mathcal{H}_x$, 我们有 $\langle f, \kappa(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}_x} = f(\boldsymbol{x})$. 那么对于在 $\Omega_x \times \Omega_y$ 上的随机向量 (X, Y), 从 RKHS $\mathcal{H}_x$ 到 $\mathcal{H}_y$ 的交叉协方差算子定义如下[32]:

$$\begin{aligned}\left\langle g, \Sigma_{YX} f\right\rangle_{\mathcal{H}_y} =& \mathrm{E}_{XY}\left[f(X)g(Y)\right] - \\ & \mathrm{E}_X\left[f(X)\right]\mathrm{E}_Y\left[g(Y)\right]\end{aligned} \tag{25}$$

其中, 所有的 $f \in \mathcal{H}_x$, $g \in \mathcal{H}_y$. 式 (25) 表明 $f(X)$ 和 $g(Y)$ 的协方差是可以通过线性算子 Σ_{YX} 的内积得到. 并且该协方差算子提供了一个讨论概率分布和条件独立性的有效框架. 如果 Σ_{XX} 可逆, 那么条件期望和交叉协方差算子之间的关系可以写成下式[32]:

$$\mathrm{E}\left[g(Y) \,|\, X = \cdot\right] = \Sigma_{XX}^{-1}\Sigma_{XY}g, \quad g \in \mathcal{H}_y \tag{26}$$

对于变量 $Y \in \mathbf{R}^{\ell}$, $U \in \mathbf{R}^{d}$ 和 $V \in \mathbf{R}^{D-d}$, 分别对应着重建核希尔伯特空间 $(\mathcal{H}_y, \kappa_y)$, $(\mathcal{H}_u, \kappa_u)$ 和 $(\mathcal{H}_v, \kappa_v)$, 每个空间都是由高斯径向基核函数 $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2/2\sigma^2)$ 诱导生成. 那么, 条件协方差算子 $\Sigma_{YY|U}$ 就可以由交叉协方差算子 Σ_{YY}, Σ_{YU} 和 Σ_{UU} 由下式得到[32]:

$$\Sigma_{YY|U} = \Sigma_{YY} - \Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY} \tag{27}$$

并且, Fukuzimu 等[32] 给出定理说明: 1) 对于任意的 U, $\Sigma_{YY|U} \geq \Sigma_{YY|X}$[4]; 2) $\Sigma_{YY|U} - \Sigma_{YY|X} = 0 \Leftrightarrow Y \perp\!\!\!\perp V \,|\, U$.

因此, 维数约简子空间 $\mathcal{Q}$ 可以通过下式最小化问题得到[32]:

$$\min_{\mathcal{Q}} \Sigma_{YY|U}, \quad \text{s.t.} \quad U = B^{\mathrm{T}}X \tag{28}$$

为了得到基于样本的目标函数, 我们需要从数据中估计条件协方差算子, 并选择特定的方式来估计自伴算子. 首先, 文献 [32] 介绍了协方差算子可以通过中心化的 Gram 矩阵得到, 即

$$\widehat{\Sigma}_{YU} = \overline{K}_Y\overline{K}_U \tag{29}$$

其中, 中心化的 Gram 矩阵的定义为 $\overline{K}_Y = HK_YH$, H 被称为中心化矩阵, 其定义为 $[I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}]$, $\mathbf{1}_n$ 为全是 1 长度为 n 的列向量. Gram 矩阵 K_Y 定义为 $[K_Y]_{ij} = \kappa_y(\boldsymbol{y}_i, \boldsymbol{y}_j)$, $1 \leq i, j \leq n$. $\overline{K}_U$ 的定义类似.

可以按照同样的方式定义协方差矩阵 $\widehat{\Sigma}_{YY}$ 和 $\widehat{\Sigma}_{UU}$. 结合文献 [31] 的正则化技术, 经验条件协方差算子 $\widehat{\Sigma}_{YY|U}$ 可以定义为

$$\begin{aligned}\widehat{\Sigma}_{YY|U} = \widehat{\Sigma}_{YY} - \widehat{\Sigma}_{YU}\widehat{\Sigma}_{UU}^{-1}\widehat{\Sigma}_{UY} = \\ (\overline{K}_Y + \epsilon I_n)^2 - \overline{K}_Y\overline{K}_U(\overline{K}_U + \epsilon I_n)^{-2}\overline{K}_U\overline{K}_Y\end{aligned} \tag{30}$$

[4] 这里的不等式为自伴算子的偏序.

表 1 基于互信息的降维算法对比

Table 1 Comparison of mutual information-based dimension reduction algorithms

目标函数	条件独立性	代表方法	解释
$\min_B \mathrm{E}_U\left[\mathbb{I}(Y\|U;V\|U)\right]$	$Y \perp\!\!\!\perp V \,\|\, U$	KDR[32−33]	去掉 X 中与 Y 独立的一些特征, 也就是说这些特征对 Y 的预测不提供任何信息
$\max_B \mathbb{I}(Y;U)$	$Y \perp\!\!\!\perp X \,\|\, U$	LSDR[34−35]、MIDR[36]	从与 Y 相关的因素中寻找一个小的子空间或者子集, 不属于该子空间或者子集的特征, 虽然对 Y 的预测作出贡献, 但是在给定子空间或者子集后, 这部分信息不足以增加或改进对 Y 的预测能力

其中, $\epsilon > 0$ 是正则项常数.

文献 [33] 中考虑将 B 放在 Stiefel 流形上 $\mathrm{St}(d,D) = \{B \in \mathbf{R}^{D\times d} \,|\, B^{\mathrm{T}}B = I_d\}$[59], 并且采用矩阵迹来重新刻画了目标函数, 得到如下目标函数

$$\min_B \mathrm{tr}\left[\overline{K}_Y(\overline{K}_U + n\epsilon I_n)^{-1}\right], \quad \mathrm{s.t.}\ U = B^{\mathrm{T}}X \tag{31}$$

因为目标函数是非凸, 可以通过非线性优化技术如基于线性查找的梯度下降法来获得最小值. 迭代求解可能求得的是局部极小值, 为缓和这问题, 可以在迭代优化过程中用逐步减小高斯核带宽的连续方法[33].

他们将这一问题看成是半参数问题, 即估计一个有限维的参数 (一个由列张成有效子空间的矩阵), 同时对 X 的分布和给定 X、Y 的条件分布不作假设. 由于通过使用核函数, 我们是在更大的空间中搜索到的最优解, 实验结果也比之前介绍的方法要好很多. 并且可作为任何监督学习器的预处理器. 这种方法与某类神经网络有些类似, 如两层神经网络. 它假设第一层是一个线性变换, 第二层是非线性变换, 从而可以看成是基于某个特定假设对回归子 $p(Y|P_BX)$ 的有效子空间估计.

但是对于一些由复杂分布或流形产生的数据, 使用线性维数约简, 并不能很好地获得数据的内部结构. 对于大样本, 核矩阵就会变得非常大, 不利于迭代求解. 虽然可以使用不完全 Cholesky 分解[31] 或 Nyström 方法[60] 来近似核矩阵, 但是 KDR 的性能还会受到核函数的选择和规则化参数的选取的影响.

国外一些学者在核维数约简基础上进行了改进和推广. 假设高维数据在于一个低维流形上面, Nilsson 等首先使用拉普拉斯特征映射 (Laplacian eigenmap) 来求得一个中间子空间[45], 然后再应用核维数约简来获得最后的一个线性子空间, 但是这个过程可能不适合预测新样本. Wang 等通过考虑将响应变量 Y 替换为 X, 从而将 KDR 推广到无监督情况[37]. 进一步将响应变量 Y 通过非线性变换, 然后再与解释变量 X 用 KDR 来实现非线性维数约简, 但这种方法缺少一定的可解释性.

5.2 最小平方维数约简

从式 (23) 可以看出, 充分维数约简的另外一种求解方式是寻求变量 X 的子空间, 使得其与响应变量之间的互信息量保持最高. 这里介绍两种通过最大化互信息的方式来获得最优中心子空间的充分维数约简方法[34, 36].

Sugiyama 等一直从事密度比率 (Density ratio) 的相关研究[61]. 他们采用一个互信息的平方损失互信息 (Squared-loss mutual information, SMI) 来进行独立性度量, 然后应用近似 SMI (Least-squares mutual information, LSMI) 进行求解, 并且称这一方法为最小平方维数约简 (Least-squared dimension reduction, LSDR)[34].

文章以平方损失互信息 (SMI) 作为度量的标准, 其定义如下[34]:

$$\mathbb{I}(Y;X) = \frac{1}{2}\iint \left(\frac{p(Y,X)}{p(Y)p(X)} - 1\right)^2 p(Y)p(X)\mathrm{d}Y\,\mathrm{d}X \tag{32}$$

这里需要说明的是 SMI 可以来估计两个随机变量之间的独立性是因为 $\mathbb{I}(Y;X)$ 等于 0 当且仅当 $p(Y,X) = p(Y)p(X)$. 而且满足数据预处理不等式 $\mathbb{I}(Y;X) - \mathbb{I}(Y;U) \geq 0$, 等号成立的条件是满足条件独立性, 即 $p(X,Y|U) = p(X|U)p(Y|U)$.

但是由于 $p(Y,U)$, $p(Y)$ 和 $p(U)$ 均是未知的, 我们不能直接使用 SMI 来估计条件独立性. 为了避免密度估计, 文章采用密度率估计来近似 SMI. 通过凸对偶, SMI 可以表示成如下[34]:

$$\mathbb{I}(Y;U) = -\inf_g J(g) - \frac{1}{2} \tag{33}$$

其中

$$J(g) = \frac{1}{2}\iint g(Y,U)^2 p(Y)p(U)\mathrm{d}Y\,\mathrm{d}U - \int g(Y,U)p(Y,U)\mathrm{d}Y\,\mathrm{d}U \tag{34}$$

这样, 计算 $\mathbb{I}(Y;U)$ 就变成寻找 $J(g)$ 的最小值 g^*, 而且 $g^*(Y,U) = \frac{p(Y,U)}{p(Y)p(U)}$[62], 这样对于 $\mathbb{I}(Y;U)$ 的估计就变成对于 $g^*(Y,U)$ 的估计.

Suzuki 等指出直接计算 $J(g)$ 会遇到两个困难, 一个是搜索空间太大, 二是密度率中真实分布未知[34]. 针对第一个问题, 我们将空间限制在一个 b 维子空间上 $\mathcal{G} := \{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{y},\boldsymbol{u})|\boldsymbol{\alpha}^{\mathrm{T}} \in \mathbf{R}^b\}$, $\boldsymbol{\alpha}$ 是从样本中学习得到, b 维子空间基函数 $\boldsymbol{\varphi}(\boldsymbol{y},\boldsymbol{u}) := [\varphi_1(\boldsymbol{y},\boldsymbol{u}),\cdots,\varphi_b(\boldsymbol{y},\boldsymbol{u})]^{\mathrm{T}} \geq \mathbf{0}_b$, $\boldsymbol{\varphi}(\boldsymbol{y},\boldsymbol{u})$ 表示样本到核空间的映射函数. 对于第二个问题, 可以使用经验分布来近似, 因此有[34]:

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbf{R}^b}\left[\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\widehat{H}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^{\mathrm{T}}R\boldsymbol{\alpha}\right] \quad (35)$$

这里的 $\frac{\lambda}{2}\boldsymbol{\alpha}^{\mathrm{T}}R\boldsymbol{\alpha}$ 表示正则化项, R 为某些正定矩阵, $\widehat{H} = \frac{1}{n^2}\sum_{i,j=1}^{n}\boldsymbol{\varphi}(\boldsymbol{y}_i,\boldsymbol{u}_j)\boldsymbol{\varphi}(\boldsymbol{y}_i,\boldsymbol{u}_j)^{\mathrm{T}}$ 和 $\widehat{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}(\boldsymbol{y}_i,\boldsymbol{u}_i)$.

对式 (35) 求导很容易得到其解析解为 $\widehat{\boldsymbol{\alpha}} = (\widehat{H} + \lambda R)^{-1}\widehat{\boldsymbol{h}}$. 这样我们就得到最小平方互信息 LSMI 的估计量[34]

$$\mathbb{I}(Y;U) := \widehat{\boldsymbol{h}}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}} - \frac{1}{2}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\widehat{H}\widehat{\boldsymbol{\alpha}} - \frac{1}{2} \quad (36)$$

因为这里 B 也是位于 Stiefel 流形上 $\mathrm{St}(d,D) = \{B \in \mathbf{R}^{D\times d} \,|\, B^{\mathrm{T}}B = I_d\}$, 所以, 可以通过梯度下降算法来寻找让 LSMI 估计量最大的变换矩阵 B. 文章中从理论上分析了最小平方维数约简的渐近收敛性, 而且对于径向基函数 $\boldsymbol{\varphi}(\boldsymbol{y},\boldsymbol{u})$ 中的参数和规则化参数 λ 可以使用交叉验证进行选择.

Suzuki 等进一步将 LSDR 推广, 得到在 Grassmann 流形上的最小平方维数约简[35]. Grassmann 流形为所有在 D 维空间的 d 维子空间的集合, 记 $\mathrm{Gr}(d,D) = \{B \in \mathbf{R}^{D\times d} \,|\, B^{\mathrm{T}}B = I_d\}/\sim$, 其中 $\sim$ 表示两个矩阵展开相同的子空间的等价关系. 文献 [63] 指出 KDR 和 LSDR 的两大限制, 一是因为使用梯度下降, 所以不适用于大样本, 二是初始值的选取, 可能导致的是局部最优值. 为了加速算法的效率, Yamada 等通过在 LSDR 每次迭代的时候, 加入基于 Epanechnikov 核的平方损失互信息估计的依赖最大化过程, 可以有效地改善算法的效率[63].

5.3　基于非参互信息的维数约简

另外一种基于 k-近邻的非参互信息估计量是由 Faivishevsky 等提出, 并且已经被成功应用到独立成分分析, 聚类和降维[36, 64−65]. 原文从概率密度微分的角度来描述基于 k-近邻的互信息, 这里我们提供一种更简便的理解. 首先, Loftsgaarden 等在 1965 年就提出了基于 k-近邻的密度估计量[66]. 给定样本集合 $\{\boldsymbol{x}_i\}_{i=1}^n$, 那么 X 的概率密度估计量定义如下:

$$p_k(\boldsymbol{x}_i) = \frac{k}{(n-1)\ c\ [\rho_k(i)]^D} \quad (37)$$

其中, 常数 c 为 D 维空间里面单位超立方体体积, $\rho_k(i)$ 表示样本 $\boldsymbol{x}_i$ 与其第 k 个近邻之间的欧氏距离. 很显然, 如果 $\rho_k(i)$ 越大, 那么说明样本 $\boldsymbol{x}_i$ 附近越稀疏, 进而密度会越小. 反之亦然.

那么, 基于 k-近邻密度估计的信息熵可以写成

$$\begin{aligned}\mathbb{H}_k(X) =& \frac{1}{n}\sum_{i=1}^{n} -\log p_k(\boldsymbol{x}_i) = \\ & \frac{1}{n}\sum_{i=1}^{n}\left[D\log\rho_k(i) + \log\frac{(n-1)c}{k}\right] = \\ & \frac{D}{n}\sum_{i=1}^{n}\log\rho_k(i) + \mathrm{const}\end{aligned} \quad (38)$$

给定 k 的取值, 那么式 (38) 会给出样本的信息熵. 然而, 对于基于 k-近邻的非参估计量, 在一些应用中它的梯度是很难得到[5]. 同时, 由于使用 k-近邻, 该估计量缺乏一定的平滑性.

因此, Faivishevsky 等提出使用平均的基于 k-近邻的估计量 (忽略常数项)

$$\begin{aligned}\mathbb{H}(X) =& \frac{1}{n-1}\sum_{k=1}^{n-1}\mathbb{H}_k(X) = \\ & \frac{D}{n(n-1)}\sum_{i\neq j}\log\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\end{aligned} \quad (39)$$

显然, 式 (39) 是一个没有任何参数且平滑的信息熵估计量. 当我们同时也有响应变量 Y 的时候, 那么可以把随机变量 X 和 Y 间的互信息通过联合和边际熵表示出来, 如果使用上式的估计量, 那么有:

$$\begin{aligned}\mathbb{I}(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) = & \\ \frac{D}{n(n-1)}\sum_{i\neq j}\log\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 + & \\ \frac{1}{n(n-1)}\sum_{i\neq j}\log\|y_i - y_j\|^2 - & \\ \frac{D+1}{n(n-1)}\sum_{i\neq j}\log(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 + & \\ \|y_i - y_j\|^2) &\end{aligned} \quad (40)$$

这里假设 y 是一维的响应变量. 当给定观测样本时, 可以用上式估计中心子空间, 相应的目标函数就写

[5] 每次投影都可能会导致近邻的变化, 因此, 不适合利用梯度下降来求解降维问题.

成下式

$$\max_{B}\left(\frac{d}{n(n-1)}\sum_{i\neq j}\log\left\|B^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_j)\right\|^2-\frac{d+1}{n(n-1)}\sum_{i\neq j}\log(\left\|B^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_j)\right\|^2+\|y_i-y_j\|^2)\right) \tag{41}$$

该方法称为基于非参互信息的维数约简 (Mutual information-based dimension reduction, MIDR). 这里的投影矩阵 B 是在 Stiefel 流形上. 虽然估计量 (41) 比前面提到的 KDR、LSDR 等简单, 而且没有参数, 可是在实际使用时, 问题相对比较多. 因为一个前提假设是样本中不能有重复样本, 其次, 使用梯度下降求解的时候, 不同样本可能在投影后变得一样, 即会导致梯度消失. 此外, 近邻的平均并不能带来很好的性能提升, 因为 Loftsgaarden 早在 1965 年就指出基于 k- 近邻的密度收敛率随着样本 n 增加, 需要满足 $\lim_{n\to\infty}k=\infty$ 和 $\lim_{n\to\infty}n/k=\infty$[66]. 显然, 当 $k=1$ 或者 $k=n-1$ 时, 密度估计的收敛率是无法保证的.

6 基于相依准则的实值多变量维数约简算法

独立性准则几乎与条件独立性是平行发展的, Gretton 等在重建核希尔伯特空间中提出了受约束的协方差标准去描述独立性[67], 随后他们通过用希尔伯特–施密特范数来取代原来的 ℓ_2 范数, 得到了希尔伯特–施密特独立性准则 (Hilbert-Schmidt independence criterion, HSIC)[68]. 与此同时, Székely 等提出一种基于样本距离而且非常简单的独立性准则估计量, 称为距离协方差 (Distance covariance, DCOV)[69]. 这两种独立性准则的估计量分别被先后引入实值多变量维数约简算法[37−38]. 下面我们先介绍随机变量的独立性, 两个随机变量的独立性可以直接写成

$$Y \perp\!\!\!\perp X \tag{42}$$

表示随机变量 X 和 Y 互相独立. HSIC 和 DCOV 均为估计两个随机变量的独立性准则. 独立的对立面是相依性 (Dependence)[6]. 因此, 本节介绍的两个算法分别基于 HSIC 和 DCOV 最大化随机变量 $B^{\mathrm{T}}X$ 和 Y 之间的相依性来获得中心子空间. 值得注意的是, 相依性 (独立性) 和条件独立性之间存在着本质的区别, 所以, 从严格意义上来讲, 基于相依准则的降维算法不属于充分维数约简范畴. 但是, 由于其目标函数有时候比一些标准的实值多变量维数约简算法更为简单, 方便优化, 同样也吸引了一些研究人的关注. 本章不作严格区分, 把基于相依准则的降维算法也列为充分维数约简算法中. 下面将着重介绍这两个独立性准则.

6.1 希尔伯特–施密特独立性准则

式 (25) 中定义的交叉协方差算子可以看成是相关系数的非线性推广, 即对任意的非线性函数 f 和 g, 如果交叉协方差仍然为 0, 那么可以说明这两个随机变量是独立的. 因此, Gretton 等提出使用希尔伯特–施密特范数来度量交叉协方差算子[68].

希尔伯特–施密特范数是对矩阵的 Frobenius 范数的推广. 因此, HSIC 的估计量其实就是算子 $\Sigma_{XY}\Sigma_{XY}^{\mathrm{T}}$ 的迹[7]. 因此, 我们用平方的希尔伯特–施密特范数来作为独立性的准则. Gretton 等证明了 HSIC 可以通过核函数表示如下[68]:

$$\begin{aligned}\mathrm{HSIC}(X,Y)=\|\Sigma_{XY}\|_{HS}^2=&\\ \mathrm{E}_{XX'YY'}\left[\kappa_x(X,X')\kappa_y(Y,Y')\right]+&\\ \mathrm{E}_{XX'}\left[\kappa_x(X,X')\right]\mathrm{E}_{YY'}\left[\kappa_y(Y,Y')\right]-&\\ 2\mathrm{E}_{XY}\left[\mathrm{E}_{X'}[\kappa_x(X,X')]\mathrm{E}_{Y'}[\kappa_y(Y,Y')]\right]&\end{aligned} \tag{43}$$

其中, (X',Y') 是 (X,Y) 的独立拷贝. 上述 HSIC 为零当且仅当变量 X 和 Y 互相独立. 因此, 最大化 HSIC 即为最大化两个随机变量之间的相依关系.

给定观测样本 $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$, HSIC 的样本估计量可以写成

$$\begin{aligned}\mathrm{HSIC}(X,Y)=&\frac{1}{(n-1)^2}\mathrm{tr}\left[K_XHK_YH\right]=\\ &\frac{1}{(n-1)^2}\mathrm{tr}\left[\overline{K}_X\overline{K}_Y\right]\end{aligned} \tag{44}$$

其中, K_X 和 K_Y 分别是 X 和 Y 对应的 Gram 矩阵, H 为上一节定义的中心化矩阵.

Wang 等通过最大化 HSIC 估计量来让 $U=B^{\mathrm{T}}X$ 和 Y 之间的相依关系最大, 即寻找一个低维空间, 使得 U 和 Y 之间的信息保留的最多[37]. 其对应的优化目标函数为

$$\max_{B}\mathrm{HSIC}(U,Y)=\tfrac{1}{(n-1)^2}\mathrm{tr}\left[K_UHK_YH\right] \tag{45}$$

相对于 KDR 的目标函数 (31), HSIC 目标函数的优势在于避免了矩阵求逆. 虽然, Wang 等证明了 HISC 和 KDR 的渐近等价性[37], 但 Suzuki 等强调了 HISC 是估计两个随机变量之间的独立性, 而并非条件独立性[34].

[6] 统计中对单词 “Dependence” 没有很明确的翻译, 有的地方翻译成 “相关”, 有的地方直译成 “依赖”, 本文中, 我们采取第三种翻译 “相依”, 表示两个随机变量相互依赖.

[7] 对于任意矩阵 $A\in\mathbf{R}^{n\times m}$, 其 Frobenius 范数 $\|A\|_F^2=\sum_{i=1}^n\sum_{j=1}^m[A]_{ij}^2=\mathrm{tr}\left[AA^{\mathrm{T}}\right]$.

值得一提的是, 最大化 HSIC 来度量变量之间的相依关系这个策略不止被用在充分维数约简降维算法中, 还被广泛用在其他应用中, 譬如, 多标签降维[70]、聚类分析[71−72]、特征选择[73]、目标匹配[74−75]、因果推演[76] 等.

6.2　距离协方差

另外一种独立性估计量是由 Szekely 等[69] 在 2007 年提出的基于样本距离而且非常简单的估计量, 称为距离协方差 (Distance covariance, DCOV). 由于其不包含参数, 易计算, 而被应用到特征抽取[77], 特征提取[38, 78] 等应用.

依据文献 [69], 距离协方差的定义如下:

定义 2.　对于给定有限一阶矩的两个随机变量 $X \in \mathbf{R}^D$ 和 $Y \in \mathbf{R}^l$, 那么距离协方差的定义如下:

$$\mathrm{DCOV}(X,Y) = \int_{\mathbf{R}^{D+l}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)\mathrm{d}t\mathrm{d}s \quad (46)$$

其中, f_X 和 f_Y 分别是 X 和 Y 的特征函数, $f_{X,Y}$ 是联合特征函数, $w(t,s)$ 是一个权重函数, 定义为

$$w(t,s) = \Big(C(D,\alpha)C(l,\alpha)\|t\|_D^{\alpha+D}\|s\|_l^{\alpha+l}\Big)^{-1} \quad (47)$$

这里的 $C(D,\alpha) = \frac{2\pi^{D/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma((\alpha+D)/2)}$.

这里的权重函数 $w(t,s)$ 需要保证式 (46) 的积分存在. 在文献 [69, 79] 中给出了 DCOV 的另一种等价形式:

$$\begin{aligned}\mathrm{DCOV}(X,Y) = &\\ &\mathrm{E}_{XX'YY'}\|X-X'\|\|Y-Y'\| + \\ &\mathrm{E}_{XX'}\|X-X'\|\mathrm{E}_{YY'}\|Y-Y'\| - \\ &2\mathrm{E}_{XY}\big[\mathrm{E}_{X'}\|X-X'\|\mathrm{E}_{Y'}\|Y-Y'\|\big]\end{aligned} \quad (48)$$

其中, (X',Y') 是 (X,Y) 的独立拷贝. 上述距离协方差为零当且仅当随机变量 X 和 Y 相互独立. 同样的, 最大化 DCOV 对应的就是相依关系. 显然地, 如果用欧氏距离替换 HSIC (43) 中的核函数, 那么 HSIC 和 DCOV 是等价的. 此外, Sejdinovic 等已经证明 DCOV 只是 HSIC 的一个特例[80].

给定样本集合 $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$, 距离协方差可以通过计算两两样本间的欧氏距离得到, 即其估计量为

$$\begin{aligned}\mathrm{DCOV}(X,Y) =& \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n [\overline{D}_X]_{ij}[\overline{D}_Y]_{ij} = \\ &\frac{1}{n^2}\mathrm{tr}\Big[\overline{D}_X\overline{D}_Y\Big]\end{aligned} \quad (49)$$

其中, $\overline{D}_X$ 为变量 X 中心化的距离矩阵, 也就是说, $\overline{D}_X = HD_XH$, $[D_X]_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$. 类似的可以定义中心化的距离矩阵 $\overline{D}_Y$.

因此, Sheng 等考虑通过最大化上述 DCOV 估计量来估计中心子空间[38], 构造目标函数为

$$\max_B \mathrm{DCOV}(U,Y) = \frac{1}{n^2}\mathrm{tr}\Big[\overline{D}_U\overline{D}_Y\Big] \quad (50)$$

相对于 HSIC, DCOV 的优势在于它的估计量中没有任何参数, 弱势在于, 估计量中采用的欧氏距离无法处理复杂的非线性相依关系.

除了这两项经典工作外, 还有一些其他变体的推广. 像 Fukumizu 等提出了一个规范化的协方差算子 $V_{XY}: \Sigma_{XY} = \Sigma_{XY}^{1/2}V_{XY}\Sigma_{YY}^{1/2}$ 在极限情况下是核函数无关的, 因此具有更好的刻画独立性关系的能力[81]. 此外, Fukumizu 等还证明了在一定假设下, 规范化的协方差算子的希尔伯特–施密特范数和式 (32) 等价[81]. 然而该估计量在基于样本情况下, 依然需要选择核函数. Poczos 等考虑使用最大均值差异 (Maximum mean discrepancy, MMD) 来衡量 Copula 的联合分布[82], 从而实现独立性的估计, 而且这种方法满足所有依赖性公理[83], 但是因为使用了序关系, 无法用于对中心子空间的估计.

7　基于回归梯度的实值多变量维数约简算法

前面提到的方法虽然可以统计高阶信息, 可是非凸的目标函数导致优化代价昂贵. 因此, 即可以包含高阶信息, 又可以快速求解, 这就使得算法更实用. 文献 [39−41] 从回归模型中条件期望的梯度入手, 提出了既可以包含高阶统计信息, 又有闭式解的两种降维方法. 我们首先介绍一下如何通过梯度来估计中心子空间.

考虑响应变量 $Y \in \mathbf{R}$, 那么我们有:

$$\begin{aligned}\frac{\partial \mathrm{E}[Y|X=\boldsymbol{x}]}{\partial \boldsymbol{x}} = \frac{\partial}{\partial \boldsymbol{x}}\int yp(y|\boldsymbol{x})\mathrm{d}y = &\\ \int y\frac{\partial}{\partial \boldsymbol{x}}p(y|B^{\mathrm{T}}\boldsymbol{x})\mathrm{d}y = &\\ B\int y\frac{\partial}{\partial \boldsymbol{u}}p(y|\boldsymbol{u})|_{\boldsymbol{u}=B^{\mathrm{T}}\boldsymbol{x}}\mathrm{d}y &\end{aligned} \quad (51)$$

其中第二步到第三步, 使用了条件独立性 $p(Y|X) = p(Y|B^{\mathrm{T}}X)$. 从式 (51) 可以看出, 对任意的 $\boldsymbol{x}$, 梯度 $\frac{\partial \mathrm{E}[Y|X=\boldsymbol{x}]}{\partial \boldsymbol{x}}$ 是包含在中心子空间里面的. 因此, Fukumizu 等使用条件协方差算子的梯度形式进而估计中心子空间, 该方法称为基于梯度的核维数约简 (Gradient-based kernel dimension reduction, gKDR)[39−40]. Sasaki 等使用最小平方对数密度的梯度来直接估计回归函数的梯度, 进而估计中心子空间, 该方法称为最小平方梯度维数约简

(Least-squares gradients for dimension reduction, LSGDR)[41].

7.1 基于梯度的核维数约简

首先, Steinwart 等已经证明, 如果在欧氏空间给定一个定义在开集上的正定核 $\kappa(\boldsymbol{x},\boldsymbol{y})$ 对 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 是连续可微的, 那么在对应 RKHS 空间的任意函数 f 也是连续可微的[84]. 更进一步, 如果 $\frac{\partial}{\partial \boldsymbol{x}}\kappa(\cdot,\boldsymbol{x})\in\mathcal{H}_x$, 我们有:

$$\frac{\partial f}{\partial \boldsymbol{x}}=\Big\langle f,\frac{\partial}{\partial \boldsymbol{x}}\kappa(\cdot,\boldsymbol{x})\Big\rangle_{\mathcal{H}_x} \tag{52}$$

其次, 式 (26) 已经给出了用交叉协方差算子表示条件期望. 假设 Σ_{XX} 是双射, $\kappa_x(\boldsymbol{x},\tilde{\boldsymbol{x}})$ 是连续可微的, 对任意 $g\in\mathcal{H}_y$, $\mathrm{E}[g(Y)\,|\,X=\cdot]\in\mathcal{H}_x$, 结合式 (52) 和式 (26), 我们有:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{x}}\mathrm{E}[g(Y)|X=\boldsymbol{x}]=&\\ \Big\langle\Sigma_{XX}^{-1}\Sigma_{XY}g,&\frac{\partial\kappa_x(\cdot,\boldsymbol{x})}{\partial \boldsymbol{x}}\Big\rangle_{\mathcal{H}_x}=\\ \Big\langle g,\Sigma_{YX}\Sigma_{XX}^{-1}&\frac{\partial\kappa_x(\cdot,\boldsymbol{x})}{\partial \boldsymbol{x}}\Big\rangle_{\mathcal{H}_y}\end{aligned} \tag{53}$$

此时, 令函数 $g=\kappa_y(\cdot,y)$, 那么我们得到如下核空间回归函数的梯度形式

$$\frac{\partial}{\partial \boldsymbol{x}}\mathrm{E}[\kappa_y(\cdot,y)\,|\,X=\boldsymbol{x}]=\Sigma_{YX}\Sigma_{XX}^{-1}\frac{\partial\kappa_x(\cdot,\boldsymbol{x})}{\partial \boldsymbol{x}} \tag{54}$$

定义矩阵 $M(\boldsymbol{x})$ 如下:

$$\begin{aligned}M(\boldsymbol{x}):=&\\ \Big\langle\Sigma_{YX}\Sigma_{XX}^{-1}&\frac{\partial\kappa_x(\cdot,\boldsymbol{x})}{\partial \boldsymbol{x}},\Sigma_{YX}\Sigma_{XX}^{-1}\frac{\partial\kappa_x(\cdot,\boldsymbol{x})}{\partial \boldsymbol{x}}\Big\rangle_{\mathcal{H}_y}=\\ \Big\langle\frac{\partial\kappa_x(\cdot,\boldsymbol{x})}{\partial \boldsymbol{x}},&\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-1}\frac{\partial\kappa_x(\cdot,\boldsymbol{x})}{\partial \boldsymbol{x}}\Big\rangle_{\mathcal{H}_x}\end{aligned} \tag{55}$$

按照式 (51), 显然可知 $D\times D$ 矩阵 $M(\boldsymbol{x})$ 的非零特征值对应的特征向量是包含在中心子空间内的. 和传统方法不一定的地方是, 这个方法包含的是高维 (或者无穷维) 的回归函数 $\mathrm{E}[\kappa_y(\cdot,y)\,|\,X=\boldsymbol{x}]$.

当给定观测样本 $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$, 基于经验协方差算子和正则化, 矩阵 $M(\boldsymbol{x})$ 可以用下式估计

$$\begin{aligned}&\widehat{M}(\boldsymbol{x})=\\ &\nabla\boldsymbol{k}_X^{\mathrm{T}}(\boldsymbol{x})\big(K_X+n\epsilon I\big)^{-1}K_Y\big(\boldsymbol{k}_X+n\epsilon I\big)^{-1}\nabla\boldsymbol{k}_X(\boldsymbol{x})\end{aligned} \tag{56}$$

其中

$$\nabla\boldsymbol{k}_X(\boldsymbol{x})=\Big[\frac{\partial\kappa_x(\boldsymbol{x}_1,\boldsymbol{x})}{\partial \boldsymbol{x}},\cdots,\frac{\partial\kappa_x(\boldsymbol{x}_n,\boldsymbol{x})}{\partial \boldsymbol{x}}\Big]^{\mathrm{T}}\in\mathbf{R}^{n\times D} \tag{57}$$

因为对任意的 $\boldsymbol{x}$, 矩阵 $M(\boldsymbol{x})$ 的特征向量都包含在中心子空间内, 因此文献 [39−40] 提出使用矩阵 $M(\boldsymbol{x})$ 的平均来构造 gKDR 的目标矩阵, 即

$$\widehat{M}:=\frac{1}{n}\sum_{i=1}^n\widehat{M}(\boldsymbol{x}_i) \tag{58}$$

和之前基于二阶矩的方法相比, gKDR 中包含更高阶的统计信息, 可以处理任意类型的 Y, 包括多模态或结构变量, 并且对 X 的分布不作假设. 同时, 如果选择某些分类或者回归的方法, 核函数的参数可以通过交叉验证的方式选取. gKDR 的时间复杂度是 $\mathrm{O}(n^3)$, 主要耗时的地方是矩阵求逆和特征值分解. 为了适应大数据的处理, 文章还提供了两种变体: 1) 通过使用不完全 Cholesky 分解; 2) 通过采样的方法, 把数据切成 h 份, 然后在每一份数据上求一个投影方向, 再用 h 个投影方向求得最终平均的投影方向.

gKDR 的性能取决于正则化参数 ϵ 和核函数 κ_x, κ_y 中的带宽参数. 并且没有系统的模型选取方法用来挑选这些参数. 因此, Sasaki 等提出了一个带有交叉验证的基于梯度的降维方法[41].

7.2 最小平方梯度维数约简

和 gKDR 出发点不一样的是, LSGDR 旨在估计条件密度的梯度, 即下式:

$$\frac{\partial}{\partial \boldsymbol{x}}p(y|\boldsymbol{x})=B\frac{\partial}{\partial \boldsymbol{u}}p(y|\boldsymbol{u}) \tag{59}$$

因此, 只要估计出条件密度的梯度, 我们就可以用外积构造矩阵, 进而估计中心子空间.

因此, Sasaki 等提出了一个对数条件密度梯度的估计量[41]:

$$\partial_j\log p(y|\boldsymbol{x})=\frac{\partial_j p(y|\boldsymbol{x})}{p(y|\boldsymbol{x})}\frac{\partial_j p(y,\boldsymbol{x})}{p(y,\boldsymbol{x})}-\frac{\partial_j p(\boldsymbol{x})}{p(\boldsymbol{x})} \tag{60}$$

其中, $\partial_j=\frac{\partial}{\partial \boldsymbol{x}^{(j)}}$. 文章的核心思想是直接用梯度模型, $\boldsymbol{g}(y,\boldsymbol{x})=[g^{(1)}(y,\boldsymbol{x}),\cdots,g^{(D)}(y,\boldsymbol{x})]$ 去拟合其真实的对数条件密度梯度. 具体过程和 LSDR 方法类似, 此处不作展开.

经过一些操作变换, $g^{(j)}$ 的经验近似损失可以写成

$$\begin{aligned}\tilde{J}(g^{(j)})=&\frac{1}{n}\sum_{i=1}^n\Big\{g^{(j)}(y_i,\boldsymbol{x}_i)\Big\}^2+\\ &2\Big\{\partial_j g^{(j)}(y_i,\boldsymbol{x}_i)+g^{(j)}(y_i,\boldsymbol{x}_i)\widehat{\gamma}^{(j)}(\boldsymbol{x}_i)\Big\}\end{aligned} \tag{61}$$

为了估计 $\partial_j \log p(y|\boldsymbol{x})$, 我们使用如下线性参数模型:

$$g^{(j)}(y,\boldsymbol{x}) = \sum_{i=1}^{n}\theta_{ij}\underbrace{\exp\Big(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2(\sigma_x^{(j)})^2}-\frac{(y-y_i)^2}{2\sigma_y^2}\Big)}_{\psi_i^{(j)}(y,\boldsymbol{x})} = \boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{\psi}_j(y,\boldsymbol{x}) \tag{62}$$

$$\partial_j g^{(j)}(y,\boldsymbol{x}) = \sum_{i=1}^{n}\theta_{ij}\underbrace{\partial_j\psi_j^{(i)}(y,\boldsymbol{x})}_{\phi_i^{(j)}(y,\boldsymbol{x})} = \boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{\phi}_j(y,\boldsymbol{x}) \tag{63}$$

其中, $\psi_i^{(j)}(y,\boldsymbol{x})$ 为基函数. 通过代入模型, 并且加入 ℓ_2 正则项, 我们可以得到闭式解

$$\widehat{\boldsymbol{\theta}}_j = \arg\min_{\boldsymbol{\theta}_j}\Big[\boldsymbol{\theta}_j^{\mathrm{T}}G_j\boldsymbol{\theta}_j + 2\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{h}_j + \lambda^{(j)}\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{\theta}_j\Big] = -(G_j+\lambda^{(j)}I)^{-1}h_j \tag{64}$$

其中, $\lambda^{(j)}$ 为正则项系数,

$$G_j = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_j(y_i,\boldsymbol{x}_i)\boldsymbol{\psi}_j^{\mathrm{T}}(y_i,\boldsymbol{x}_i) \tag{65}$$

$$\boldsymbol{h}_j = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}_j(y_i,\boldsymbol{x}_i) + \widehat{\gamma}^{(j)}(\boldsymbol{x}_i)\boldsymbol{\psi}_j(y_i,\boldsymbol{x}_i) \tag{66}$$

最后, 我们可以得到一个最小平方对数条件密度梯度 (Least-squares logarithmic conditional density gradient, LSLCG) 估计量, 即

$$\widehat{g}^{(j)}(y,\boldsymbol{x}) = \widehat{\boldsymbol{\theta}}_j^{\mathrm{T}}\boldsymbol{\psi}_j(y,\boldsymbol{x}) \tag{67}$$

为了估计中心子空间, 和 gKDR 类似的定义外积矩阵

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{g}}(y_i,\boldsymbol{x}_i)\widehat{\boldsymbol{g}}^{\mathrm{T}}(y_i,\boldsymbol{x}_i) \tag{68}$$

其最大特征值对应的特征向量构成我们的中心子空间, 该方法称为最小平方梯度维数约简. 相对于 gKDR, 该方法具有交叉验证, 便于模型选取.

8 基于非线性的充分维数约简算法

前面提到的所有方法都是线性维数约简算法. 而对于数据中的非线性结构, 线性往往不能满足需求. 因此, 一些研究人员开始研究非线性的实值多变量维数约简算法. 非线性降维算法主要考虑由核函数 κ_x 诱导得到的重建核希尔伯特空间 $\mathcal{H}_x$, 那么任意样本 $\boldsymbol{x}$ 都可以通过隐式的非线性特征变换映射到重建核希尔伯特空间, 即 $\boldsymbol{\phi}(\boldsymbol{x}) = \kappa_x(\cdot,\boldsymbol{x}) \in \mathcal{H}_x$, 本节将介绍两种非线性方法, 希望寻找一组非线性的投影向量 $\boldsymbol{b}_l(\cdot) \in \mathcal{H}_x$, $l = 1,\cdots,d$, 组成非线性的中心子空间, 方法分别为核切片逆回归 (Kernel sliced inverse regression, KSIR)[42] 和协方差算子逆回归 (Covariance operator inverse regression, COIR)[8, 11, 85].

8.1 核切片逆回归

KSIR 是 SIR 的核化算法. 在第 3.1 节中, 我们已经介绍了 SIR 通过逆回归曲线 $\mathrm{E}[\boldsymbol{x}\,|\,y]$ 来估计中心子空间, 并且其估计量可以写成 $\mathrm{var}(\mathrm{E}[\boldsymbol{x}\,|\,y])$. 而在 KSIR 中, 作者直接将 $\boldsymbol{x}$ 换成 $\boldsymbol{\phi}(\boldsymbol{x})$ 实现非线性. 即 KSIR 的估计量为 $\mathrm{var}(\mathrm{E}[\boldsymbol{\phi}(\boldsymbol{x})\,|\,y])$.

和 SIR 类似, KSIR 首先将数据按照 Y 的数值切成 h 份, 分别为 $I_1,\cdots,I_h$. 令 $\Phi_x = [\boldsymbol{\phi}(\boldsymbol{x}_1),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_n)]$, C 表示 $n\times h$ 的 0/1 指示矩阵, 每一行只有一个元素是 1 (表示该样本属于某一切片), 其余为 0. 此外, 矩阵 P 为一个 $h\times h$ 的对角矩阵, 其对角线元素表示该切片内样本占全部样本的比例, 即 $[P]_{jj} = |I_j|/n$. 那么, KSIR 按照以下步骤估计中心子空间

1) 把 $\{y_i\}_{i=1}^n$ 切成 h 份, 计算切片内均值

$$M = [\boldsymbol{m}_1,\cdots,\boldsymbol{m}_h] = \frac{1}{n}\Phi_x CP^{-1} \tag{69}$$

2) 估计切片内样本的协方差矩阵, 即

$$V = MPM^{\mathrm{T}} \tag{70}$$

我们把矩阵 V 的前 d 个特征向量记成 $\{\boldsymbol{v}_l\}_{l=1}^d$.

3) 中心子空间的方向可以通过 $\boldsymbol{b}_l = \Sigma_{XX}^{-1}\boldsymbol{v}_l$, $l = 1,\cdots,d$ 获得.

虽然上面的步骤不能显示的计算, 但是我们可以通过核技巧实现. 更具体点, 在步骤 2), 为了求解特征值分解 $V\boldsymbol{v} = \lambda\boldsymbol{v}$, 我们把 $\boldsymbol{v}$ 表示成 $\{\boldsymbol{\phi}(\boldsymbol{x})_i\}_{i=1}^n$ 的线性组合, 也就是说, $\boldsymbol{v} = \sum_{i=1}^n\alpha_i\boldsymbol{\phi}(\boldsymbol{x}_i) = \Phi_x\boldsymbol{\alpha}$, 其中 $\boldsymbol{\alpha} = [\alpha_1,\cdots,\alpha_n]^{\mathrm{T}}$. 在 $V\boldsymbol{v} = \lambda\boldsymbol{v}$ 的两侧同乘 Φ_x^{T} 后, 可以得到下面对偶方程:

$$\frac{1}{n^2}K_X CP^{-1}C^{\mathrm{T}}K_X\boldsymbol{\alpha} = \lambda K_X\boldsymbol{\alpha} \tag{71}$$

步骤 3) 中的投影向量 $\boldsymbol{b}$ 可以用类似的方式从 $\boldsymbol{v}$ 得到. 为了求解 $\Sigma_{XX}\boldsymbol{b} = \boldsymbol{v}$, 我们把样本估计量 $\frac{1}{n}\Phi_x\Phi_x^{\mathrm{T}}$ 来替换协方差算子 Σ_{XX}. 令 $\boldsymbol{b} = \sum_{i=1}^n\beta_i\boldsymbol{\phi}(\boldsymbol{x}_i) = \Phi_x\boldsymbol{\beta}$, 其中 $\boldsymbol{\beta} = [\beta_1,\cdots,\beta_n]^{\mathrm{T}}$, 并且两侧同乘以 Φ_x^{T}, 即可以得到 $\boldsymbol{\beta}$ 的闭式解:

$$\boldsymbol{\beta} = nK_X^{-1}\boldsymbol{\alpha} \tag{72}$$

非线性中心子空间就可以通过 d 个从 $\{\boldsymbol{\beta}_l\}_{l=1}^d$ 构成的基函数 $\{\boldsymbol{b}_l\}_{l=1}^d$ 表示. 如果给定一个新的

样本 $\boldsymbol{x}^*$, 它的低维表示 $\boldsymbol{z}^* \in \mathbf{R}^d$ 可以通过将 $\boldsymbol{\phi}(\boldsymbol{x}^*)$ 投影中心子空间得到, 也就是说, $\boldsymbol{z}^*$ 的第 l 个元素是: $\boldsymbol{b}_l^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}^*) = \boldsymbol{k}_*^{\mathrm{T}}\boldsymbol{\beta}^l$, $l = 1, \cdots, d$, 其中 $\boldsymbol{k}_* = [\kappa_x(\boldsymbol{x}_1, \boldsymbol{x}^*), \cdots, \kappa_x(\boldsymbol{x}_n, \boldsymbol{x}^*)]^{\mathrm{T}}$.

核化的切片逆回归方法解决了 SIR 的一些问题. 它允许一个对边际分布有较少约束的非线性子空间. 然后, KSIR 基于切片的估计量 $\mathrm{var}(\mathrm{E}[\boldsymbol{\phi}(\boldsymbol{x}) \,|\, y])$ 并不适用于高维的变量 Y. 这样使得 KSIR 只能用于一元变量回归或者分类. 为了避免因为切片带来的问题, Kim 等提出了一个新的协方差算子逆回归[8, 11, 85].

8.2 协方差算子逆回归

前面已经提到交叉协方差算子可以通过协方差算子定义, 即

$$\Sigma_{XX|Y} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \tag{73}$$

并且在一些温和条件下, 条件协方差算子 $\Sigma_{XX|Y}$ 和给定 $\boldsymbol{y}$, $\boldsymbol{\phi}(\boldsymbol{x})$ 的期望条件方差等价. 具体描述如下面定理.

定理 2. 对任意的函数 $f \in \mathcal{H}_x$, 如果存在一个函数 $g \in \mathcal{H}_y$, 使得对几乎所有的 $\boldsymbol{y}$, 有 $\mathrm{E}[f(\boldsymbol{x}) \,|\, \boldsymbol{y}] = g(\boldsymbol{y})$. 那么 $\Sigma_{XX|Y} = \mathrm{E}[\mathrm{var}[\boldsymbol{\phi}(\boldsymbol{x}) \,|\, \boldsymbol{y}]]$.

使用方差公式, 我们有:

$$\mathrm{var}[\mathrm{E}[\boldsymbol{\phi}(\boldsymbol{x})|\boldsymbol{y}]] = \mathrm{var}(\boldsymbol{\phi}(\boldsymbol{x})) - \mathrm{E}[\mathrm{var}[\boldsymbol{\phi}(\boldsymbol{x})|\boldsymbol{y}]] \tag{74}$$

通过定理 2 和式 (73), 式 (74) 可以写成

$$\mathrm{var}[\mathrm{E}[\boldsymbol{\phi}(\boldsymbol{x}) \,|\, \boldsymbol{y}]] = \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \tag{75}$$

注意 $\mathrm{var}(\boldsymbol{\phi}(\boldsymbol{x})) = \Sigma_{XX}$.

COIR 需要对上式进行特征值分解, 给定观测样本 $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, 再令 $\Phi_y = [\boldsymbol{\phi}(\boldsymbol{y}_1), \cdots, \boldsymbol{\phi}(\boldsymbol{y}_n)]$, 那么式 (75) 的估计量可以写成

$$\begin{aligned}&\mathrm{var}(\mathrm{E}[\boldsymbol{\phi}(\boldsymbol{x}) \,|\, \boldsymbol{y}]) = \\ &\Big(\frac{1}{n}\Phi_x\Phi_y^{\mathrm{T}}\Big)\Big(\frac{1}{n}(\Phi_y\Phi_y^{\mathrm{T}} + n\epsilon I)\Big)^{-1}\Big(\frac{1}{n}\Phi_y\Phi_x^{\mathrm{T}}\Big) = \\ &\frac{1}{n}\Phi_x^{\mathrm{T}}\Phi_y(\Phi_y\Phi_y^{\mathrm{T}} + n\epsilon I)^{-1}\Phi_y\Phi_x^{\mathrm{T}} = \\ &\frac{1}{n}\Phi_x K_Y(K_Y + n\epsilon I)^{-1}\Phi_x^{\mathrm{T}}\end{aligned} \tag{76}$$

同样的核技巧, 特征值分解 $\mathrm{var}(\mathrm{E}[\boldsymbol{\phi}(\boldsymbol{x}) \,|\, \boldsymbol{y}])\boldsymbol{v} = \lambda\boldsymbol{v}$ 可以写成下式,

$$\frac{1}{n}K_Y(K_Y + n\epsilon I)^{-1}K_X\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \tag{77}$$

因此, 和 KSIR 类似的, 通过式 (72) 得到最终的非线性子空间. 和 KSIR 相比, COIR 中不需要对响应变量 Y 切片, 可以处理多元变量的 Y, 并且该方法可以轻易推广到半监督情况[85].

9 实值多变量维数约简算法求解

首先, 本节给出对本文介绍的所有实值多变量维数约简算法的全面对比. 对比主要从响应变量类型、算法求解类型、分布假设、算法中使用的统计矩、是否带有交叉验证、算法中的回归方式、算法中的参数个数以及参数列表, 详见表 2. 需要指出的是, 表 2 中的统计矩表示模型中使用变量 X 的阶矩; 无穷表示使用高斯核或者 log 操作可以把变量映射到一个无穷维的空间. 模型选取中的 "-" 表示模型无参数, 不需要模型选取; "无" 表示该模型无系统的选取方法, 需结合其他回归算法作交叉验证; "交叉验证" 表示该方法可直接进行交叉验证.

表 2 实值多变量维数约简算法总结

Table 2 Summary of real-valued multivariate dimension reduction algorithms

算法	响应变量	求解类型	分布假设	统计矩	模型选取	回归方式	# 参数	参数列表
SIR[23]	一元	解析解	椭圆对称	一阶	无	逆回归	1	切片数 h
SAVE[26]	一元	解析解	椭圆对称	二阶	无	逆回归	1	h
SIRII[49]	一元	解析解	椭圆对称	二阶	无	逆回归	1	h
pHd[25]	一元	解析解	正态分布	二阶	−	正回归	0	−
DR[27]	一元	解析解	渐进正态	二阶	无	逆回归	1	h
PFC[28]	一元	解析解	模型假设	模型相关	无	逆回归	1	次数 r
KDR[33]	多元	非凸优化	无	无穷	无	正回归	3	带宽 $\sigma_x, \sigma_y, \epsilon$
LSDR[35]	多元	非凸优化	无	无穷	交叉验证	正回归	4	$\sigma_x, \sigma_y, \lambda, b$
MIDR[36]	多元	非凸优化	无	无穷	−	正回归	0	−
HSIC[37]	多元	非凸优化	无	无穷	无	正回归	2	σ_x, σ_y
DCOV[38]	多元	非凸优化	无	二阶	−	正回归	0	−
gKDR[40]	多元	解析解	无	无穷	无	正回归	3	$\sigma_x, \sigma_y, \epsilon$
LSGDR[41]	多元	解析解	无	无穷	交叉验证	正回归	$2D+1$	$\{\sigma_x^{(j)}, \lambda^{(j)}\}_{j=1}^D, \sigma_y$
KSIR[42]	一元	解析解	无	无穷	无	逆回归	2	σ_x, h
COIR[8]	多元	解析解	无	无穷	无	逆回归	3	$\sigma_x, \sigma_y, \epsilon$

至此, 我们已经对降维算法有一定的了解, 剩下的就是如何求解这些算法. 本章所介绍的降维算法可以分为两种求解类型: 基于广义特征值分解的降维算法和基于矩阵流形优化求解的降维算法. 本节最后会介绍一种内在维度的估计算法.

9.1 基于广义特征值求解的实值多变量维数约简算法

表 3 中列出算法可以转换给广义特征值求解. 需要特别指出的是, 除了 KSIR 和 COIR 为非线性降维外, 其余均为线性降维.

表 3　基于广义特征值求解的实值多变量维约简算法估计量

Table 3　Estimation of generalized eigen-decomposition-based multivariate dimension reduction algorithms

方法	(样本) 估计量 M
SIR[23]	$\mathrm{E}\Big[\mathrm{E}(Z\|Y)\mathrm{E}(Z\|Y)^{\mathrm{T}}\Big]$
SAVE[26]	$\mathrm{E}\Big[I-\mathrm{var}(Z\|Y)\Big]^2$
SIRII[49]	$\mathrm{E}\Big[\mathrm{var}(Z\|Y)-\mathrm{E}(\mathrm{var}(Z\|Y))\Big]^2$
pHd[25]	$\mathrm{E}\Big[(Y-EY)ZZ^{\mathrm{T}}\Big]^2$
DR[27]	$2\mathrm{E}\Big[E^2(ZZ^{\mathrm{T}}\|Y)\Big]+2E^2\Big[\mathrm{E}(Z\|Y)\mathrm{E}(Z^{\mathrm{T}}\|Y)\Big]+$ $2\mathrm{E}\Big[\mathrm{E}(Z^{\mathrm{T}}\|Y)\mathrm{E}(Z\|Y)\Big]\mathrm{E}\Big[\mathrm{E}(Z\|Y)\mathrm{E}(Z^{\mathrm{T}}\|Y)\Big]-2I$
PFC[28]	$\frac{1}{n}\widehat{X}^{\mathrm{T}}\widehat{X}$
gKDR[40]	$\Big\langle \frac{\partial \kappa_x(\cdot,\boldsymbol{x})}{\partial \boldsymbol{x}}, \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-1}\frac{\partial \kappa_x(\cdot,\boldsymbol{x})}{\partial \boldsymbol{x}}\Big\rangle_{\mathcal{H}_x}$
LSGDR[41]	$\mathrm{E}(\boldsymbol{g}(y,\boldsymbol{x})\boldsymbol{g}(y,\boldsymbol{x})^{\mathrm{T}})$
KSIR[42]	$\mathrm{var}(\mathrm{E}[\boldsymbol{\phi}(\boldsymbol{x})\,\|\,y])$
COIR[8]	$\mathrm{var}(\mathrm{E}[\boldsymbol{\phi}(\boldsymbol{x})\,\|\,y])$

在线性降维基础上, 维数约简成分都是原来预测量的线性组合得到. 这样就会很难对约简后的成分进行解释. 例如, 我们有一个回归模型 $Y=\exp(-0.5\boldsymbol{b}^{\mathrm{T}}X)+0.5\epsilon$, 其中 X 是 6 维向量, 并且噪音变量 ϵ 是独立的标准正态分布. 那么中心子空间 $\mathcal{S}_{Y|X}=\mathrm{span}(\boldsymbol{b})$, 并且这里的 $\boldsymbol{b}$ 设为 $(1,-1,0,0,0,0)^{\mathrm{T}}/\sqrt{2}$. 拿 SIR 为例, 通过 100 个仿真样本, 得到的估计量为 $\widehat{\boldsymbol{b}}=(0.651,-0.745,-0.063,0.134,0.014,0.003)^{\mathrm{T}}$. 虽然最后 4 个系数相对比较小, 但是 6 个维度的信息都被包含在里面了, 这样会导致对真实中心子空间的估计出现偏差. 所以, 文献 [43−44] 组合了充分维数约简中的回归技术和稀疏学习中的收缩估计策略, 对子空间进行了稀疏化表示.

稀疏维数约简是对现有算法的推广, 本节简要介绍两种稀疏策略. 1) 文献 [43] 借助稀疏主成分分析的思想[86], 将广义特征值问题转换为优化问题, 并使用 LASSO (Least absolute shrinkage and select operator) 算法求解. 期望使得投影矩阵 B 中的一些元素收缩至 0, 这样可以有更好的解释性. 2) 考虑到上面方法中的惩罚项是坐标依赖的, 即对于基的正交变换, ℓ_1 的惩罚项不是保持不变的. 因此, 文献 [44] 提出一个坐标独立的惩罚项, 并且考虑对投影矩阵 B 的行向量进行惩罚, 这样在实现稀疏学习的同时还可以进行变量选取.

9.2 基于矩阵流形优化的实值多变量维数约简算法

由于实值多变量维数约简算法中基于矩阵流形优化的投影矩阵 B 都是在 Stiefel 流形 $\mathrm{St}(D,d)=\{B\in\mathbf{R}^{D\times d}\ :\ B^{\mathrm{T}}B=I_d\}$, 其中 I_d 是 $d\times d$ 的单位矩阵. 为了方便, 我们简记成 $\mathcal{O}^{D\times d}$. 那么 $\mathcal{O}^{D\times d}$ 是一个嵌入在 $\mathbf{R}^{D\times d}$ 的子流形.

这里我们提供矩阵流行优化最基本的介绍, 主要是基于流形的一阶投影梯度优化[59, 87]. 直觉上, 流形投影梯度方法是迭代优化路线, 首先需要理解在约束集上的搜索方向, 我们称为切空间 (Tangent space). 给定一个目标函数 f, 梯度 $-\nabla_B f$ 是在全空间上计算得到的, 即空间 $\mathbf{R}^{D\times d}$. 这些梯度都被投影到切空间上. 在线性切空间的任意非零步长都会导致离开非线性约束集. 所以, 最后我们都需要使用回缩 (Retraction) 把步长投影到约束集合上. 使用上面三步, 一个标准的一阶迭代求解方法就可以用来求解矩阵流行优化问题了. 详见图 2.

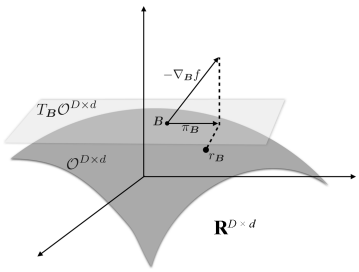


图 2　矩阵流形优化示意图

Fig. 2　Illustration of matrix manifold optimization

切空间 $T_B\mathcal{O}^{D\times d}$: 切空间对于理解任意流形的几何结构是至关重要的, 它是在某一个点对流形的线性近似. 为了定义该空间, 我们首先定义流形 $\mathcal{O}^{D\times d}$ 上的一条平滑曲线 $\gamma(\cdot):\mathbf{R}\to\mathcal{O}^{D\times d}$. 那么切空间定义如下:

$$T_B\mathcal{O}^{D\times d}=\big\{\gamma'(0):\gamma(\cdot)\text{ 是流形 }\mathcal{O}^{D\times d}\text{ 上的曲线,}\\ \text{并且满足 }\gamma(0)=B\big\}\tag{78}$$

这里的 γ' 表示微分 $\frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)$. 简单而言, 切空间 $T_B\mathcal{O}^{D\times d}$ 是在点 B 处所有沿着流形方向的空间. 然而上式是非常一般的表达形式, 而且非常抽象. 为了方便, 下面给出 Stiefel 流形的简洁等价形式:

命题 2. 下面三个集合是互相等价的

$$T_B\mathcal{O}^{D\times d}=\big\{\gamma'(0):\gamma(\cdot)\text{ 是流形 }\mathcal{O}^{D\times d}\text{ 上的曲线,}\\ \text{并且满足 }\gamma(0)=B\big\}\tag{79}$$

$$T_1=\big\{X\in\mathbf{R}^{D\times d}:B^{\mathrm{T}}X+X^{\mathrm{T}}B=0\big\}\tag{80}$$

$$T_2=\big\{BA+(I-BB^{\mathrm{T}})C\colon A=-A^{\mathrm{T}},C\in\mathbf{R}^{D\times d}\big\}\tag{81}$$

式 (81) 的定义是非常有用的, 因为它是构造性的, 对于优化而言是非常必须的.

投影 $\pi_B:\mathbf{R}^{D\times d}\to T_B\mathcal{O}^{D\times d}$ 因为 $\mathcal{O}^{D\times d}$ 是嵌入在欧氏空间的 $\mathbf{R}^{D\times d}$ 的一个子流形, 所以自然想到使用欧氏空间的度量: 标准的内积运算 $\langle P,N\rangle=\mathrm{tr}(P^{\mathrm{T}}N)$ 以及诱导的 Frobenius 范数 $\|\cdot\|_F$. 通过这种度量, Stiefel 流形是欧氏空间的一个黎曼子流形 (Riemannian submanifold). 这可以直接使我们考虑一个可以把任意矩阵 $Z\in\mathbf{R}^{D\times d}$ 投影到切空间 $T_B\mathcal{O}^{D\times d}$ 上的投影, 换句话说,

$$\begin{aligned}&\pi(Z)=\arg\min_{X\in T_B\mathcal{O}^{D\times d}}\|Z-X\|_F=\\&\arg\min\|Z-(BA+(I-BB^{\mathrm{T}})C)\|_F=\\&\arg\min\|(BB^{\mathrm{T}}Z-BA)+(I-BB^{\mathrm{T}})(Z-C)\|_F=\\&\arg\min\|B(B^{\mathrm{T}}Z-A)\|_F+\|(I-BB^{\mathrm{T}})(Z-C)\|_F=\\&\arg\min\|(B^{\mathrm{T}}Z-A)\|_F+\|(I-BB^{\mathrm{T}})(Z-C)\|_F\end{aligned}\tag{82}$$

最后一个等式是由于 Frobenious 范数的酉不变性. 上述的最小值可以通过设置 $C=Z$ 和 A 为 $B^{\mathrm{T}}Z$ 的斜对称部分, 即 $A=\mathrm{skew}(B^{\mathrm{T}}Z)=\frac{1}{2}(B^{\mathrm{T}}Z-Z^{\mathrm{T}}B)$[88]. 所以, 投影为

$$\pi_B(Z)=B\mathrm{skew}(B^{\mathrm{T}}Z)+(I-BB^{\mathrm{T}})Z\tag{83}$$

回缩 $r_B:T_B\mathcal{O}^{D\times d}\to\mathcal{O}^{D\times d}$ 投影梯度方法需要沿着流形最陡梯度方向迭代步长, 也就是说 $B+\beta\pi_B(-\nabla_Bf)$. 任意非零步长 β, 这个迭代都会离开 Stiefel 流形. 所以, 回缩是需要把它再投影到流形上来. 现在有很多不同的投影回缩方法[89], 这里我们定义一个在点 B 处步长 Z 回缩到流形的方法:

$$r_B(Z)=\arg\min_{N\in\mathcal{O}^{D\times d}}\|N-(B+Z)\|_F\tag{84}$$

即在流形上离迭代 $B+Z$ 最近的点. 由于是酉不变范数, 一个经典的结果就是 $r_B(Z)=UV^{\mathrm{T}}$, 其中 $(B+Z)=USV^{\mathrm{T}}$ 是奇异特征值分解[88]. 或者等价的 $r_B(Z)=W$, 其中 $(B+Z)=WP$ 是极分解 (Polar decomposition). 更方便地, 当 $Z\in T_B\mathcal{O}^{D\times d}$, 那么回缩是有闭式解的 $r_B(Z)=(B+Z)(I+Z^{\mathrm{T}}Z)^{-1/2}$[89].

算法 1. 在 Stiefel 流形上的梯度下降 (包含线性查找和回缩)

初始化 B
while $f(B)$ 未收敛 **do**
  计算 $\nabla_Bf\in\mathbf{R}^{D\times d}$
  计算 $\pi_B(-\nabla_Bf)\in T_B\mathcal{O}^{D\times d}$
  while $f(r_B(\beta\pi_B(-\nabla_Bf)))$ 没有充分小于 $f(B)$ **do**
    调整步长 β
  end while
  $B\leftarrow r_B(\beta\pi_B(-\nabla_Bf))$
end while
return f 的 (局部) 最小值 B^*

介绍完矩阵流形优化里面的三个重要概念后, 就可以利用传统的非凸优化方法进行优化.

算法 1 给出通过线性查找的进行梯度下降优化的伪代码.

因此, 只需要知道非凸目标函数在欧氏空间的梯度, 即可通过算法 1 进行优化求解. 基于矩阵流形优化的实值多变量维数约简算法及其欧氏空间梯度见表 4. 注意表 4 中的优化目标函数均已转化成求最小值. 其中, $L=D-W$, D 为对角矩阵, 对角线元素为 W 每行元素和; LSDR 方法虽然没有解析形式的, 但是其梯度可逐元素计算: $[\nabla_Bf]_{ij}=\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\frac{\partial\widehat{H}}{\partial B_{ij}}(\frac{3}{2}\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}})-\frac{\partial\widehat{\boldsymbol{h}}^{\mathrm{T}}}{\partial B_{ij}}(2\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}})+\lambda\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\frac{\partial R}{B_{ij}}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\beta}}),\widehat{\boldsymbol{\beta}}=(\widehat{H}+\lambda R)^{-1}\widehat{H}\widehat{\boldsymbol{\alpha}}$; MIDR 在计算权值矩阵 W 时的除法为矩阵除法, 即逐元素相除, 实际计算中对角线元素强行置为零. 此外, 符号 $\odot$ 表示矩阵 Hadamard 乘积.

9.3 中心子空间维度估计

前面我们在讨论中心子空间的时候, 都是假设子空间维度 d 已知. 然而, 实际情况下, 我们无法知道其真实的子空间维度. 文献 [23, 25] 提出基于卡方统计量的序列检验. 文献 [90] 提出了使用排列检验来估计子空间的维度. 使用自助法来估计子空间维度被广泛使用[38, 54, 91]. 因此, 本节主要介绍使用自助法进行子空间维度估计.

首先, 我们介绍一个距离度量[55]:

$$\Delta(\mathcal{S}_1,\mathcal{S}_2)=\|P_{\mathcal{S}_1}-P_{\mathcal{S}_2}\|_2\tag{85}$$

这里的, $\mathcal{S}_1$ 和 $\mathcal{S}_2$ 表示 D 维空间的两个 d 维子空间, $P_{\mathcal{S}_1}$ 和 $P_{\mathcal{S}_2}$ 为到这两个子空间的正交投影, 这里的 $\|\cdot\|_2$ 为矩阵 2-范数, 即矩阵的最大特征值. 因此 Δ 越小, 两个子空间越近.

自助法的基本想法如下: 对所有可能的维度

$1 \le k \le D-1$, 我们可以获得一个估计的子空间 $\widehat{\mathcal{S}}_k$, 然后我们可放回的采样和原数据等大小的数据 T 次, 我们通过自助样本, 我们可以得到 T 个估计的子空间 $\widehat{\mathcal{S}}_k^t, t=1,\cdots,T$. 因此, 对于每一个 k, 我们可以计算出 $\widehat{\mathcal{S}}_k$ 和 $\widehat{\mathcal{S}}_k^t$, 记为 $\Delta(\widehat{\mathcal{S}}_k, \widehat{\mathcal{S}}_k^t), t=1,\cdots,T$. 最后我们使用 $\Delta(\widehat{\mathcal{S}}_k, \widehat{\mathcal{S}}_k^t), t=1,\cdots,T$ 的均值来表示每一个 k 的变异性的度量, 那么最小变异度对应的 k 就是我们需要估计的 d.

下面分类讨论一下该方法的合理性:

1) 当 $k<d$ 时, 那么我们估计的 k 维子空间是中心子空间的子空间. 理论上来说, 该中心子空间的 k 维子空间可以是任意的. 因此在自助法的情况下, 变异度会增加.

2) 当 $k=d$ 时, 中心子空间总是可以被估计出来.

3) 当 $k>d$ 时, 有 $k-d$ 个方向是会和中心子空间垂直, 而且是随机得到. 因此, 变异度会显著增加.

10　典型应用

本文探讨的实值多变量维数约简算法适用于一般意义上的回归问题中的维数约简, 涵盖一维和多维响应变量的情况. 为了便于读者了解文本核心内容所针对的研究方向, 本节将介绍行人计数[10]、人体姿态估计[8] 和空间划分树[15] 这三个经典应用.

1) 行人计数, 是指通过计算机技术对视频中正在行走的行人进行数量的估计. 在当今社会对行人计数的需要越来越突出, 尤其在公共安全和智能交通等领域中, 行人计数扮演着重要的角色, 如卖场的人流管理、地铁人流引导、火车站人流管控、集会临时通行口设置等[92−94]. 随着众多学者的深入研究, 越来越多的学者对行人计数问题提出了有效的行人数量、密度特征来对行人进行预估, 例如, 文献[9−10] 中应用的行人特征达 129 维. 因此, 通过维数约简算法对数据预处理、消除冗余、挖掘特征间的相关性就显得很有必要. Zhang 等使用核维数约简算法进行行人特征提取, 与未降维特征和主成分分析降维后特征相比, 核维数约简算法可以实现更好的特征提取[10]. 此应用中的响应变量为一维的情况, 即每张图像中的行人数量.

2) 人体姿态估计, 是指通过人体的轮廓图来估计人体三维空间的姿态. Kim 等在实验中, 使用一个 160×100 的轮廓图, 来对人体 59 个三维关节角的进行预测[8, 11]. 应用中的自变量 X 为 16 000 维, 而响应变量为 177 维的向量, 即本文所介绍的多元响应变量的情况. Kim 等使用提出的协方差算子逆回归算法对输入空间进行降维, 并进行可视化和模型预测. 可视化结果表明, 协方差算子逆回归算法可以有效挖掘人体姿态的内在规律. 预测结果表明, 协方差算子逆回归算法可以提高预测精度.

3) 空间划分树是指迭代使用切平面对数据空间进行划分, 在向量量化和最近邻查找过程中起重要的作用[15, 95]. 其中, 切平面的表达学习对空间划分树的构建起到决定性的作用. Shan 等通过使用无监督的核维数约简算法来学习数据空间的一维投影方向[15], 即切平面. 较之前的空间划分树算法, Shan 等提出的空间划分树在向量量化和最近邻查找中实现更好的性能, 并且具有发现数据内在维度的潜力. 在该应用中, 虽然是无监督的设置, 实质上是把自变量 X 的拷贝视为响应变量 Y[37], 即为响应变量为多元的情况.

11　总结与展望

本文对实值多变量维数约简维数约简算法作了综述, 并介绍了实值多变量维数约简发展历史. 按照不同的研究技术, 把现有算法划分成基于前两阶矩、模型、互信息、相依准则、回归梯度和非线性的实值多变量维数约简, 并分析其代表算法的优缺点. 其次, 通过求解手段的不同, 算法可以划分成基于特征值求解和矩阵流形优化求解.

表 4　通过矩阵流形优化的降维算法的总结

Table 4　Summary of matrix manifold optimization-based dimension reduction algorithms

算法	目标函数 $f(B)$	欧式空间梯度 $\nabla_B f$	对应的权值矩阵 W
KDR[33]	$\mathrm{tr}[\overline{K}_Y(\overline{K}_U+n\epsilon I_n)^{-1}]$	$\frac{2}{\sigma^2}XL_{\mathrm{KDR}}X^{\mathrm{T}}B$	$W_{\mathrm{KDR}}=H\Omega H\odot K_U$ $\Omega=M^{-1}HK_YHM^{-1}$ $M=HK_UH+n\epsilon I$
LSDR[35]	$\frac{1}{2}\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\hat{H}\hat{\boldsymbol{\alpha}}-\hat{\boldsymbol{h}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}$	不存在解析形式梯度	无拉普拉斯形式
MIDR[36]	$\frac{d+1}{n(n-1)}\sum_{i\neq j}\log(\Vert(\boldsymbol{x}_i-\boldsymbol{x}_j)B\Vert^2+\Vert y_i-y_j\Vert^2)-$ $\frac{d}{n(n-1)}\sum_{i\neq j}\log\Vert(\boldsymbol{x}_i-\boldsymbol{x}_j)B\Vert^2$	$-\frac{4}{n(n-1)}XL_{\mathrm{MIDR}}X^{\mathrm{T}}B$	$W_{\mathrm{MIDR}}=\frac{d}{D_X\odot D_X}-$ $\frac{d+1}{(D_X\odot D_X+D_Y\odot D_Y)}$
HSIC[37]	$-\frac{1}{(n-1)^2}\mathrm{tr}[K_UHK_YH]$	$\frac{2}{\sigma^2(n-1)^2}XL_{\mathrm{HSIC}}X^{\mathrm{T}}B$	$W_{\mathrm{HSIC}}=HK_YH\odot K_U$
DCOV[38]	$-\frac{1}{n^2}\mathrm{tr}[\overline{D}_U\overline{D}_Y]$	$-\frac{2}{n^2}XL_{\mathrm{DCOV}}X^{\mathrm{T}}B$	$W_{\mathrm{DCOV}}=\frac{HD_YH}{D_U}$

从第 1 节介绍的关于实值多变量维数约简的研究现状看, 存在以下值得进一步研究的问题和潜在的方向:

1) 实值多变量维数约简方法侧重于考虑响应变量集与自变量集之间的条件独立性等统计特性, 而较少考虑数据集中的拓扑结构保持. 多数研究工作集中考虑了获得的统计子空间是否仍属于中央子空间的问题, 如中央均值子空间、二阶方差子空间等. 这些工作强调了不同统计性质的线性组合不会改变其属于中央子空间的性质. 但这些统计性质不能有效地反映数据集中的拓扑结构. 从近年来的研究趋势来看, 对数据的分析不仅要求预测性能好, 也需要有好的可解释性. 因此, 在低维约简时保持好相应的拓扑结构就十分重要了.

2) 实值多变量维数约简方法多数假定数据集具有线性和对称分布, 如椭圆分布. 当数据是由低维流形嵌套在高维空间的表达时, 这类方法可能在约简过程中, 不能正确的恢复低维流形, 因而导致预测性能降低. 在生物认证中的某些数据集, 如姿态等, 在降维时可能呈现光滑的沿流形的变化, 如果在降维时失去低维流形的正确或近似表达, 则会降低对姿态的预测. 尽管在文献 [45] 中涉及了流形学习与实值多变量维数约简的融合, 但其处理方法 (先利用拉普拉斯降维, 再采用核维数约简做进一步降维) 未能很好的同时保持这两者共有的优势.

3) 实值多变量维数约简通常假定了数据与时间序列无关, 而时序数据集的特点是数据之间具有时间的依赖性, 或上下文关系. 如智能交通领域中对行人的计数, 在进入采集区域的人群数量显然具有时间序列的关联性; 在生物认证中, 步态序列里帧与帧之间也具有时间的依赖性. 如果忽视这层关系, 则在降维过程中会导致预测响应变量时产生错误的时序表达, 从而影响对数据分析的全局判断. 因此, 时序数据的实值多变量维数约简方法研究有待于研究.

4) 实值多变量维数约简主要考虑了单组自变量集和响应变量集之间的关系, 并试图构造按条件独立性来简化高维自变量集和响应变量集的结构. 当待分析的数据呈现为复杂关系时, 如常见的贝叶斯网络结构时, 现有的实值多变量维数约简方法则不能很好地获取相应的结构. 而研究这类复杂网络关系或概率图模型在机器学习领域是比较重要的一个分支, 包括对其的建模、推断和学习.

5) 面对大数据时代的到来, 更对维数约简提出了很大的挑战, 同样也显示出降维的必要性. 然而, 如何将实值多变量维数约简应用到规模庞大的数据中并且可以对数据进行高效的处理就成为一大难题. 核矩阵在大数据中就会变得异常大, 因此基于核函数的那些方法就更显得束手无策. 但是大数据或许会有些统计特性来放松对线性条件的约束.

6) 虽然文献 [43−44] 中提出了对充分维数约简的稀疏化学习, 但是局限于特征值分解的情况. 因此, 有必要对基于矩阵流形优化的实值多变量维数约简方法进行稀疏化学习, 并且可以考虑为实值多变量维数约简建立统一的稀疏化学习框架, 这样更有利于加强实值多变量维数约简的应用前景.

References

1 Jolliffe I. *Principal Component Analysis*. New York: Wiley Online Library, 2002.

2 Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 1993, **24**(6): 417−441

3 Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, **290**(5500): 2323−2326

4 Tenenbaum J B, Silva V D, Langford J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, **290**(5500): 2319−2323

5 Fisher R A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936, **7**(2): 179−188

6 Mika S, Ratsch G, Weston J, Schölkopft B, Mullert K R. Fisher discriminant analysis with kernels. In: Proceeding of the 1999 IEEE Signal Processing Society on Neural Networks for Signal Processing IX. Madison, WI, USA, USA: IEEE, 1999. 41−48

7 Kulis B. Metric learning: a survey. *Foundations and Trends in Machine Learning*, 2012, **5**(4): 287−364

8 Kim M, Pavlovic V. Central subspace dimensionality reduction using covariance operators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(4): 657−670

9 Tan B, Zhang J P, Wang L. Semi-supervised elastic net for pedestrian counting. *Pattern Recognition*, 2011, **44**(10−11): 2297−2304

10 Zhang J P, Tan B, Sha F, He L. Predicting pedestrian counts in crowded scenes with rich and high-dimensional features. *IEEE Transactions on Intelligent Transportation Systems*, 2011, **12**(4): 1037−1046

11 Kim M Y, Pavlovic V. Dimensionality reduction using covariance operator inverse regression. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA: IEEE, 2008. 1−8

12 Zhang J P, Wang F Y, Wang K F, Lin W H, Xu X, Chen C. Data-driven intelligent transportation systems: a survey. *IEEE Transactions on Intelligent Transportation Systems*, 2011, **12**(4): 1037−1046

13 Zhu H J, Li L X. Biological pathway selection through nonlinear dimension reduction. *Biostatistics*, 2011, **12**(3): 429−444

14 Zhu H J, Li L X, Zhou H. Nonlinear dimension reduction with Wright-Fisher kernel for genotype aggregation and association mapping. *Bioinformatics*, 2012, **28**(18): i375−i381

15 Shan H M, Zhang J P, Kruger U. Learning linear representation of space partitioning trees based on unsupervised kernel dimension reduction. *IEEE Transactions on Cybernetics*, 2016, **46**(12): 3427−3438

16 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

17 Caruana R. Multitask learning. *Machine Learning*, 1997, **28**(1): 41−75

18 Tsoumakas G, Katakis I. Multi-label classification: an overview. *International Journal of Data Warehousing and Mining*, 2007, **3**(3): 1−13

19 Tipping M E. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 2001, **1**: 211−244

20 Wang Wei-Wei, Li Xiao-Ping, Feng Xiang-Chu, Wang Si-Qi. A survey on sparse subspace clustering. *Acta Automatica Sinica*, 2015, **41**(8): 1373−1384
(王卫卫, 李小平, 冯象初, 王斯琪. 稀疏子空间聚类综述. 自动化学报, 2015, **41**(8): 1373−1384)

21 Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*. Massachusetts: MIT Press, 2009.

22 Liu Jian-Wei, Li Hai-En, Luo Xiong-Lin. Learning technique of probabilistic graphical models: a review. *Acta Automatica Sinica*, 2014, **40**(6): 1025−1044
(刘建伟, 黎海恩, 罗雄麟. 概率图模型学习技术研究进展. 自动化学报, 2014, **40**(6): 1025−1044)

23 Li K C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 1991, **86**(414): 316−327

24 Chiaromonte F, Cook R D. Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics*, 2002, **54**(4): 768−795

25 Li K C. On principal Hessian directions for data visualization and dimension reduction: another application of Stein′s lemma. *Journal of the American Statistical Association*, 1992, **87**(420): 1025−1039

26 Cook R D, Weisberg S. Sliced inverse regression for dimension reduction: comment. *Journal of the American Statistical Association*, 1991, **86**(414): 328−332

27 Li B, Wang S L. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 2007, **102**(479): 997−1008

28 Cook R D. Fisher lecture: dimension reduction in regression. *Statistical Science*, 2007, **22**(1): 1−26

29 Cook R D, Forzani L. Principal fitted components for dimension reduction in regression. *Statistical Science*, 2008, **23**(4): 485−501

30 Cook R D, Forzani L. Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 2009, **104**(485): 197−208

31 Bach F R, Jordan M I. Kernel independent component analysis. *Journal of Machine Learning Research*, 2002, **3**: 1−48

32 Fukumizu K, Bach F R, Jordan M I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 2004, **5**: 73−99

33 Fukumizu K, Bach F R, Jordan M I. Kernel dimension reduction in regression. *The Annals of Statistics*, 2009, **37**(4): 1871−1905

34 Suzuki T, Sugiyama M. Sufficient dimension reduction via squared-loss mutual information estimation. In Proceedings of International Conference on Artificial Intelligence and Statistics, Chia Laguna Resort, Sardinia, Italy, 2010, **9**: 804−811

35 Suzuki T, Sugiyama M. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*, 2013, **25**(3): 725−758

36 Faivishevsky L, Goldberger J. Dimensionality reduction based on non-parametric mutual information. *Neurocomputing*, 2012, **80**: 31−37

37 Wang M H, Sha F, Jordan M I. Unsupervised kernel dimension reduction. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2010. 2379−2387

38 Sheng W H, Yin X R. Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 2016, **25**(1): 91−104

39 Fukumizu K, Leng C L. Gradient-based kernel method for feature extraction and variable selection. In: Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, 2012. 2114−2122

40 Fukumizu K, Leng C L. Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, 2014, **109**(505): 359−370

41 Sasaki H, Tangkaratt V, Sugiyama M. Sufficient dimension reduction via direct estimation of the gradients of logarithmic conditional densities. In: Proceedings of the 7th Asian Conference on Machine Learning. Hong Kong, China: PMLR, 2015. 33−48

42 Wu H M. Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 2012, **17**(3): 590−610

43 Li L X. Sparse sufficient dimension reduction. *Biometrika*, 2007, **94**(3): 603−613

44 Chen X, Zou C L, Cook R D. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 2010, **38**(6): 3696−723

45 Nilsson J, Sha F, Jordan M I. Regression on manifolds using kernel dimension reduction. In: Proceedings of the 24th International Conference on Machine Learning. Corvalis, Oregon, USA: ACM, 2007. 697−704

46 Shyr A, Urtasun R, Jordan M I. Sufficient dimension reduction for visual sequence classification. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, California, USA: IEEE, 2010. 3610−3617

47 Cook R D. On the interpretation of regression plots. *Journal of the American Statistical Association*, 1994, **89**(425): 177−189

48 Cook R D. Using dimension-reduction subspaces to identify important inputs in models of physical systems. In: Proceedings of the Section on Physical and Engineering Sciences. 1994. 18−25

49 Li K C. Sliced inverse regression for dimension reduction: rejoinde. *Journal of the American Statistical Association*, 1991, **86**(414): 337−342

50 Harrison D, Rubinfeld D L. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 1978, **5**(1): 81−102

51 Stein C M. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 1981, **9**(6): 1135−1151

52 Li K C, Lue H H, Chen C H. Interactive tree-structured regression via principal Hessian directions. *Journal of the American Statistical Association*, 2000, **95**(450): 547−560

53 Gannoun A, Saracco J. An asymptotic theory for SIR α method. *Statistica Sinica*, 2003, **13**: 297−310

54 Ye Z S, Weiss R E. Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 2003, **98**(464): 968−979

55 Li B, Zha H Y, Chiaromonte F. Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, 2005, **33**(4): 1580−1616

56 Tipping M E, Bishop C M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1999, **61**(3): 611−622

57 Cook R D, Forzani L M, Tomassi D R. LDR: a package for likelihood-based sufficient dimension reduction. *Journal of Statistical Software*, 2011, **39**(3): 1−20

58 Baker C R. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 1973, **186**: 273−289

59 Absil P A, Mahony R, Sepulchre R. *Optimization Algorithms on Matrix Manifolds*. Princeton: Princeton University Press, 2009.

60 Zhang K, Tsang I W, Kwok J T. Improved Nyström low-rank approximation and error analysis. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008. 1232−1239

61 Sugiyama M, Suzuki T, Kanamori T. *Density Ratio Estimation in Machine Learning*. Cambridge: Cambridge University Press, 2012.

62 Nguyen X, Wainwright M J, Jordan M I. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: ACM, 2008. 1089−1096

63 Yamada M, Niu G, Takagi G, Sugiyama M. Computationally efficient sufficient dimension reduction via squared-loss mutual information. In: Proceedings of the 3rd Asian Conference on Machine Learning. Canberra, Australia: ACML, 2011. 247−262

64 Faivishevsky L, Goldberger J. ICA based on a smooth estimation of the differential entropy. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada: ACM, 2009. 433−440

65 Faivishevsky L, Goldberger J. Nonparametric information theoretic clustering algorithm. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: ICML, 2010. 351−358

66 Loftsgaarden D O, Quesenberry C P. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 1965, **36**(3): 1049−1051

67 Gretton A, Smola A J, Bousquet O, Herbrich R, Belitski A, Augath M, Murayama Y, Pauls J, Schölkopf B, Logothetis N K. Kernel constrained covariance for dependence measurement. In: Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics. Barbados, 2005. 1−8

68 Gretton A, Bousquet O, Smola A, Schölkopf B. Measuring statistical dependence with Hilbert-Schmidt norms. In: Proceedings of the 16th International Conference on Algorithmic Learning Theory. Berlin, Heidelberg: Springer-Verlag, 2005. 63−77

69 Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 2007, **35**(6): 2769−2794

70 Zhang Y, Zhou Z H. Multi-label dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data*, 2010, **4**(3): Article No. 14

71 Song L, Smola A, Gretton A, Borgwardt K M. A dependence maximization view of clustering. In: Proceedings of the 24th International Conference on Machine Learning. Corvalis, Oregon, USA: ACM, 2007. 815−822

72 Blaschko M, Gretton A. Learning taxonomies by dependence maximization. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, B. C., Canada, 2009. 153−160

73 Song L, Smola A, Gretton A, Bedo J, Borgwardt K. Feature selection via dependence maximization. *The Journal of Machine Learning Research*, 2012, **13**(1): 1393−1434

74 Quadrianto N, Song L, Smola A J. Kernelized sorting. In: Proceedings of the Advances in neural information processing systems. Vancouver, B. C., Canada, 2009. 1289−1296

75 Quadrianto N, Smola A J, Song L, Tuytelaars T. Kernelized sorting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(10): 1809−1821

76 Sun X H, Janzing D, Schölkopf B, Fukumizu K. A kernel-based causal learning algorithm. In: Proceedings of the 24th International Conference on Machine Learning. Corvalis, Oregon, USA: ACM, 2007. 855−862

77 Li R Z, Zhong W, Zhu L P. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 2012, **107**(499): 1129−1139

78 Vepakomma P, Tonde C, Elgammal A. Supervised dimensionality reduction via distance correlation maximization. arXiv: 1601.00236, 2016.

79 Székely G J, Rizzo M L. Brownian distance covariance. *The Annals of Applied Statistics*, 2009, **3**(4): 1236−1265

80 Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 2013, **41**(5): 2263−2291

81 Fukumizu K, Gretton A, Sun X H, Schölkopf B. Kernel measures of conditional dependence. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada: Curran Associates Inc., 2007. 489−496

82 Póczos B, Ghahramani Z, Schneider J G. Copula-based kernel dependencymeasures. In: Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland: ICML, 2012. 1635−1642

83 Schweizer B, Wolff E F. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 1981, **9**(4): 879−885

84 Steinwart I, Christmann A. *Support Vector Machines*. Berlin, Heidelberg: Springer Science & Business Media, 2008.

85 Kim M Y, Pavlovic V. Covariance operator based dimensionality reduction with extension to semi-supervised settings. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Varna, Bulgaria: Springe, 2009. 280−287

86 Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 2006, **15**(2): 265−286

87 Cunningham J P, Ghahramani Z. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 2015, **16**: 2859−2900

88 Fan K, Hoffman A J. Some metric inequalities in the space of matrices. *Proceedings of the American Mathematical Society*, 1995, **6**(1): 111−116

89 Kaneko T, Fiori S, Tanaka T. Empirical arithmetic averaging over the compactStiefel manifold. *IEEE Transactions on Signal Processing*, 2013, **61**(4): 883−894

90 Cook R D, Yin X R. Theory & methods: special invited paper: dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, 2001, **43**(2): 147−199

91 Zhu Y, Zeng P. Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, 2006, **101**(476): 1638−1651

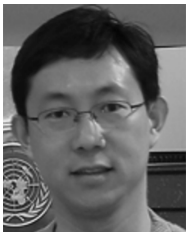
92 Davies A C, Yin J H, Velastin S A. Crowd monitoring using image processing. *Electronics & Communication Engineering Journal*, 1995, **7**(1): 37−47

93 Cho S Y, Chow T W, Leung C T. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 1999, **29**(4): 535−541

94 Shi Zeng-Lin, Ye Yang-Dong, Wu Yun-Peng, Lou Zheng-Zheng. Crowd counting using rank-based spatial pyramid pooling network. *Acta Automatica Sinica*, 2016, **42**(6): 866−874
(时增林, 叶阳东, 吴云鹏, 娄铮铮. 基于序的空间金字塔池化网络的人群计数方法. 自动化学报, 2016, **42**(6): 866−874)

95 Dasgupta S, Freund Y. Random projection trees and low dimensional manifolds. In: Proceedings of the 40th Annual ACM Symposium on Theory of Computing. Victoria, British Columbia, Canada: ACM, 2008. 537−546

单洪明 复旦大学计算机学院博士研究生. 主要研究方向为维数约简, 随机特征, 深度学习.
E-mail: hmshan@fudan.edu.cn
(**SHAN Hong-Ming** Ph. D. candidate at the School of Computer Science, Fudan University. His research interest covers dimension reduction, random feature, and deep learning.)



张军平 复旦大学计算机科学技术学院教授. 主要研究方向为机器学习, 智能交通, 生物认证, 图像识别. 本文通信作者.
E-mail: jpzhang@fudan.edu.cn
(**ZHANG Jun-Ping** Professor at the School of Computer Science, Fudan University. His research interest covers machine learning, intelligent transportation systems, biometric authentication, and image processing. Corresponding author of this paper.)