



充分降维的研究进展

骆威

浙江大学数据科学研究中心, 杭州 310058

E-mail: weiluo@zju.edu.cn

收稿日期: 2024-02-18; 接受日期: 2024-10-28; 网络出版日期: 2024-11-25

国家自然科学基金 (批准号: 12222117 和 12131006) 资助项目

摘要 充分降维是统计学研究中一类主流的数据降维理论与方法, 致力于将监督学习中的自变量 X 替换为低维线性变换 $\beta_0^T X$, 而不降低对响应变量 Y 的预测能力. 本文简要介绍充分降维研究的发展历史, 包括充分降维的基本理论及逆回归方法、最小平均方差估计方法、估计方程形式的半参数方法等常见的充分降维方法, 以及高维数据情形下的稀疏和非稀疏逆回归方法、充分降维在因果推断中的应用、非线性充分降维等前沿进展.

关键词 充分降维 因果推断 监督学习

MSC (2020) 主题分类 62B05

1 引言

数据分析的目的常常在于利用一组自变量来总结或预测研究者感兴趣的某个响应变量, 例如, 利用基因组信息预测某疾病, 利用过往经济指标预测未来经济发展趋势等. 直观上, 自变量组里的变量越多, 就能包含更多关于响应变量的信息, 从而使预测愈加精确. 因此, 随着数据采集技术的不断发展, 数据中常出现自变量为高维的现象. 这样的高维数据已在医学、经济学等许多领域普及. 由于数据中观测点的数量 (样本量) 往往相对有限, 传统的统计方法会因为遭受“维数灾难”而无法适用. 因此, 数据降维在最近三十多年成为了统计学研究中的重要问题, 其中充分降维是监督学习中降维方法的主流.

记 X 为 p 维自变量, Y 为一维或多维响应变量, 充分降维假设存在低维线性变换 $\beta_0^T X$, 使得 X 只通过 $\beta_0^T X$ 作用于 Y , 即

$$Y \perp\!\!\!\perp X \mid \beta_0^T X, \quad (1.1)$$

这里 $\perp\!\!\!\perp$ 表示两个随机变量独立. 在以文献 [28, 39, 42, 46] 等为代表的研究中, 线性变换 $\beta_0^T X$ 被替代为更广义的低维函数 $R(X)$ (详见第 6 节). 因此, (1.1) 也被称为线性充分降维. 由于 (1.1) 对 β_0 的任意可逆列变换具有不变性, 出于可识别性考虑, 通常将满足 (1.1) 的 β_0 的列所生成的线性空间看作 (1.1)

英文引用格式: Luo W. Some recent advances in sufficient dimension reduction (in Chinese). Sci Sin Math, 2025, 55: 1935–1948, doi: 10.1360/SSM-2024-0034

所定义的参数, 而将充分降维的目标参数定义为 $\mathbb{R}^p$ 中这些线性子空间的维数最小者. 这一目标参数被称为中心降维子空间, 并记为 $\mathcal{S}_{Y|X}$. 中心降维子空间的概念最早由 Cook^[8] 在 1998 年提出, 其唯一存在性的条件则由 Yin 等^[93] 在 2008 年放宽, 从而对一般非奇异数据成立. 这些工作都晚于充分降维首次被提出的时间 (1989 年, 参见文献 [50]). 充分降维的目标是对 $Y | \beta_0^T X$ 模型自由的前提下, 估计中心降维子空间 $\mathcal{S}_{Y|X}$. 这里所描述的模型自由是指不对 $Y | \beta_0^T X$ 的条件分布作任何参数模型假设, 从而不对基于低维数据 $(Y, \beta_0^T X)$ 的后续统计分析作任何限制.

在实际应用中, 研究者有时只关注回归分析, 即估计 $Y | X$ 的条件均值函数. 此时, 充分降维假设 (1.1) 可调整为 $E(Y | X)$ 的低维结构假设, 也就是存在低维线性变换 $\beta_0^T X$, 使得 $E(Y | X)$ 是 $\beta_0^T X$ 的可测函数. 由于 $E(\cdot)$ 是线性算子, 这等价于假设

$$E(Y | X) = E(Y | \beta_0^T X). \quad (1.2)$$

类似 $\mathcal{S}_{Y|X}$, (1.2) 所定义的目标参数被称为中心均值降维子空间 $\mathcal{S}_{E(Y|X)}$, 其任意基矩阵均满足 (1.2) 且具有最低维数^[13]. 除回归分析以外, 研究者也可能关注 $Y | X$ 的其他方面, 例如, 在金融学常见的波动分析中研究 $Y | X$ 的方差, 当 Y 富含异常值时研究 $Y | X$ 的分位数等. 对于前者, 文献 [107] 提出了中心方差降维子空间 $\mathcal{S}_{\text{Var}(Y|X)}$; 对于后者, 文献 [38] 提出了相关的充分降维理论. 推而广之, 对一般的泛函 $T(Y | X)$, 充分降维假设也可以作相应调整, 使其对应的目标参数为中心 T 降维子空间 $\mathcal{S}_{T(Y|X)}$ (参见文献 [64]). 为叙述方便, 我们用中心降维子空间 $\mathcal{S}_{Y|X}$ 统一指代不同场景下的充分降维目标参数, 而只在特定时候区分 $\mathcal{S}_{Y|X}$ 和 $\mathcal{S}_{E(Y|X)}$ 等的区别, 希望不会引起歧义. 上述充分降维假设适用于线性回归模型、广义线性回归模型、单指标回归模型、多指标回归模型、线性判别分析和二项式判别分析等常见的统计模型, 因此具有广泛的现实基础.

由于目标参数 $\mathcal{S}_{Y|X}$ 的非欧性, 为了估计方便, 充分降维方法往往引入一个矩阵型参数 M 作为中间参数, 其列空间 $\mathcal{S}(M)$ 在适当条件下等于中心降维子空间 $\mathcal{S}_{Y|X}$. 这里的 M 不但可以是 (1.1) 或者 (1.2) 中的 β_0 , 也可以具有更广泛的形式, 例如, 由 $X | Y$ 的条件矩所生成; 后者是充分降维研究的主流思路之一, 相应的方法被称为逆回归方法. 第 2 节将详细介绍这类方法. 由此展开, 第 3 节将介绍其他基于损失函数或者估计方程的充分降维方法, 第 4 节将介绍逆回归方法在高维数据情形下的拓展, 第 5 节将介绍充分降维对因果推断这一类热门统计学研究方法的改进. 第 6 节将介绍本节一开始提到的非线性充分降维方法.

最后, 所有充分降维方法均涉及对 $\mathcal{S}_{Y|X}$ 或类似目标参数 ($\mathcal{S}_{E(Y|X)}$ 和 $\mathcal{S}_{T(Y|X)}$ 等) 的维数估计. 这类问题被称为定阶问题^[62]. 相应的定阶方法通常基于对上述矩阵型中间参数 M 或相关损失函数的相合估计. 每一节将对不同类型的充分降维方法介绍相应的定阶方法. 为叙述方便, 用 d 统一表示 $\mathcal{S}_{Y|X}$ 等目标参数的维数.

2 逆回归方法

如第 1 节中所述, 逆回归方法作为一类主流的充分降维方法, 利用 $X | Y$ (而非 $Y | X$) 的条件矩来构造矩阵型中间参数 M , 使得 $\mathcal{S}(M)$ 在适当条件下等于中心降维子空间 $\mathcal{S}_{Y|X}$. 通常, M 被称为候选矩阵. 在确定 M 及其相合估计 $\widehat{M}$ 后, 只需要提取 $\widehat{M}$ 的前 d 个左奇异向量或特征向量作为基矩阵, 即可生成 $\mathcal{S}_{Y|X}$ 的相合估计.

常见的逆回归方法包括切片逆回归方法^[49]、切片平均方差估计方法^[15]、方向性回归方法^[47] 和三阶逆回归方法^[89, 90] 等. 例如, 假设 $E(X) = 0$, 记 X 的协方差矩阵为 Σ , 并且对任意矩阵 B 记 BB^T

为 $B^{\otimes 2}$, 则切片逆回归方法构造的 M 具有如下形式:

$$M_{\text{SIR}} = \Sigma^{-1} \{E(X | Y_S = 1), \dots, E(X | Y_S = H)\}^{\otimes 2} \Sigma^{-1}. \quad (2.1)$$

这里的 Y_S 为 Y 的切片变换: 当 Y 为离散型变量或分类变量时, 令 $Y_S = Y$; 当 Y 为一维连续型变量时, 对于 $i = 1, \dots, H$, 取 $Y_S = i$ 当且仅当 Y 大于等于它的第 $\{(i-1)/H\} \times 100\%$ 分位数并且小于它的第 $(i/H) \times 100\%$ 分位数; 当 Y 为多维连续型变量时, 取 Y_S 为利用 K -means 方法 (其中 $K = H$) 对 Y 作聚类分析所得结果 (参见文献 [79]). 用 Y_S 替代 Y 的技术被称为切片策略 [49], 其中关于 H 的选择可参见文献 [55] 等. 为避免对 H 调参, 切片策略也可调整为先取不同的二分切片再融合相应的 M_{SIR} 集合 [106]. 这些策略旨在提高相应 M 的可估性, 因此被逆回归方法普遍采用. 例如, 对于切片逆回归方法, 有 $\widehat{M}_{\text{SIR}} = \widehat{\Sigma}^{-1} \{\widehat{E}(X | Y_S = 1), \dots, \widehat{E}(X | Y_S = H)\}^{\otimes 2} \widehat{\Sigma}^{-1}$, 其中 $\widehat{\Sigma}$ 和 $\widehat{E}(X | Y_S = i)$ 分别为 X 的样本协方差矩阵和在第 i 个切片里的样本均值. 由此得到的矩阵估计 $\widehat{M}$ 具有理想的渐近性质, 包括 $\sqrt{n}$ 相合性和渐近正态性等.

当充分降维的目标参数为 $\mathcal{S}_{E(Y|X)}$ 时, 同样可引入矩阵型参数 M , 如最小二乘估计 [50]、主 Hessian 方向估计 [15] 和迭代 Hessian 变换 [13] 等. 这些矩阵只涉及 $E(XY)$ 或 $E(X^{\otimes 2}Y)$ 等边际矩, 而无需估计 $E(X | Y)$ 等条件矩. 为了叙述方便, 我们仍将它们统一归入逆回归方法的集合.

由于充分降维方法不对 $Y | X$ 假设参数模型, 因此, 逆回归方法用简单的矩来刻画复杂的条件独立性 (1.1), 其代价只能是对 X 的分布作相应限制. 对于切片逆回归方法, M 张成 $\mathcal{S}_{Y|X}$ 的子空间的一个充分条件是 X 满足线性条件:

$$E(X | \beta_0^T X) = P^T(\Sigma, \beta_0)X, \quad (2.2)$$

其中 β_0 为 $\mathcal{S}_{Y|X}$ 的任意基矩阵, $P(\Sigma, \beta_0) = \beta_0(\beta_0^T \Sigma \beta_0)^{-1} \beta_0^T \Sigma$ 是基于内积 $\langle u, v \rangle_{\Sigma} = u^T \Sigma v$ 的投影矩阵. 除线性条件 (2.2) 外, 切片平均方差估计方法以及方向性回归方法等涉及更高阶矩的逆回归方法还要求 X 满足常数方差条件:

$$\text{Var}(X | \beta_0^T X) = \Sigma P(\Sigma, \beta_0), \quad (2.3)$$

或将其放宽至 $\text{Var}(X | \beta_0^T X)$ 为 $\beta_0^T X$ 的二阶多项式 [60], 而三阶逆回归方法另外要求 X 满足对称条件 (参见文献 [89]). 这些条件也可改为对 $X | Y$ 而非 X 的参数模型假设 (详见文献 [10, 14]). 在线性条件 (2.2) 和常数方差条件 (2.3) 下, 如何选择逆回归方法取决于数据量以及研究者是否接受相关正则化假设等. 一方面, 切片逆回归方法需要 $X | Y$ 具有不对称性, 以及当 Y 为离散变量或分类变量时其可能结果的数量大于 $\mathcal{S}_{Y|X}$ 的维数 d . 而切片平均方差估计方法等高阶逆回归方法由于利用了数据里的更多信息, 因此不需要这些正则化假设, 从而更有可能完整地估计 $\mathcal{S}_{Y|X}$. 另一方面, 高阶逆回归方法由于涉及更多参数, 因此除了额外要求常数方差条件 (2.3) 外, 对样本量的要求也更高. 此外, 不同的逆回归方法也可结合使用 (参见文献 [87]). 更多细节可参见文献 [7, 41].

因为线性条件 (2.2) 和常数方差条件 (2.3) 等对 X 的参数假设, 逆回归方法在 $Y | X$ 模型自由的同时依然拥有简单易算、估计精确等优点. 然而, 这些假设同样给逆回归方法带来了应用范围上的争议. 一方面, 当 X 满足适当的正则化假设且维数 p 趋于无穷时, 文献 [17, 33] 在 Bayes 统计学框架下证明了 (2.2) 和 (2.3) 对 $\mathbb{R}^p$ 中的任意低维子空间 $\mathcal{S}(\beta_0)$ 几乎一致近似成立. 大量应用案例表明, 逆回归方法也对这些条件具有良好的稳健性. 因此, 在高维数据情形下, 对 X 的参数模型假设不对逆回归方法的应用能力构成限制. 另一方面, 如果将 (2.2) 和 (2.3) 加强为对所有低维矩阵 β 均成立, 则 (2.2)

要求 X 满足椭球分布, 而 (2.2) 和 (2.3) 共同要求 X 满足多元正态分布. 此外, 在高维数据情形下, 研究者常常假设 $\mathcal{S}_{Y|X}$ 或者 β_0 的稀疏性 (详见第 4 节). 该假设与文献 [17, 33] 中关于 β_0 的连续型先验分布假设相悖, 使得文献 [17, 33] 的 Bayes 近似理论无法适用. 鉴于这一争论, 文献 [21, 44] 将线性条件 (2.2) 推广为对 $E(X | \beta_0^T X)$ 的一般化参数模型假设或半参数假设, 而另一类主流的充分降维方法致力于对 $Y | X$ 和 X 均实现模型自由. 后者被称为最小平均方差估计 (minimum average variance estimation) 类方法, 详见第 3 节.

除去关于 X 的分布假设的争议, 逆回归方法在高维数据等复杂情形下依然是最主要的充分降维方法. 这是由于它们在估计精度和计算方面的优势, 尤其是与线性回归分析在形式上的相似性 (参见文献 [5, 12, 51, 81, 97]). 作为代表, 切片逆回归方法在超高维稀疏数据情形 [5, 51, 56, 57, 81, 97] 和高维非稀疏数据情形 [12, 26, 56, 66] 下均得到了丰富的拓展. 第 4 节将详细介绍这些工作.

最后, 基于逆回归方法的候选矩阵 M , 对 $\mathcal{S}_{Y|X}$ 的定阶问题等价于利用 $\widehat{M}$ 估计 M 的秩. 传统定阶方法研究 $\widehat{M}$ 的非显著特征值 (真值为零) 的渐近性质, 利用它们的渐近分布构造序贯假设检验 (参见文献 [4, 47, 55]), 或者利用 $\widehat{M}$ 的所有特征值由大到小排序所生成的碎石图, 构造 Bayes 信息准则 (参见文献 [70, 105]). 最新的定阶方法 [62, 63] 则基于自助法或引入白噪声变量等, 利用 $\widehat{M}$ 的特征向量刻画 M 的秩, 并结合 $\widehat{M}$ 的特征值构造在 d 处取最小值的目标函数. 由于后者同时利用了 $\widehat{M}$ 的特征值和特征向量所包含的定阶信息, 它们在理论上比传统方法更为精确, 同时避免了非退化的第一类错误和调参稳健性等问题 (参见文献 [62]).

3 最小平均方差估计方法及估计方程方法

由 (1.2) 可知, 在回归分析场景中, 中心均值降维子空间 $\mathcal{S}_{E(Y|X)}$ 的任意基矩阵 β_0 可等价定义为 $E\{Y - E(Y | \beta_0^T X)\}^2$ 的最小值点. 对后者的非参数估计自然生成了对 β_0 或 $\mathcal{S}_{E(Y|X)}$ 的模型自由估计方法. 这一方法被称为最小平均方差估计 [85]. 记 $K_h(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ 为窗宽为 h 的一维核密度函数, 如 Gauss 核 $(2\pi)^{-1/2} h^{-1} \exp\{-x^2/(2h^2)\}$. 记 $(X_1, Y_1), \dots, (X_n, Y_n)$ 为数据中的观测值. 对任意矩阵 $\beta \in \mathbb{R}^{p \times d}$, 利用局部线性回归构造对 $E(Y | \beta^T X)$ 的非参数估计, 从而以

$$\sum_{i=1}^n \left[\min_{a_i \in \mathbb{R}, b_i \in \mathbb{R}^d} \sum_{j=1}^n \{Y_j - a_i - b_i^T \beta^T (X_j - X_i)\}^2 K_h(\|X_j - X_i\|_2) \right] \quad (3.1)$$

的最小值点 $\widehat{\beta} \in \mathbb{R}^{p \times d}$ 生成对 $\mathcal{S}_{E(Y|X)}$ 的估计. 这一过程无需对 X 的分布作任何参数假设, 因此相较于逆回归方法适用于更一般的数据.

为了最小化 (3.1), 文献 [85] 提出了交替迭代最小二乘算法. 给定 β 的初始值后, 这一算法具有很快的收敛速度. 然而, 由于 (3.1) 是 β 的非凸且奇异的函数, 所以它是否收敛到全局最小值点 $\widehat{\beta}$ 取决于 β 初始值的相合性. 为此, 文献 [85] 提出了主梯度方法. 该方法先利用核均值估计方法近似 $\partial E(Y | X) / \partial X$, 得到 $M_{\text{OPG}} \equiv E[\{\partial E(Y | X) / \partial X\}^{\otimes 2}]$ 的相合估计; 再取其前 d 个特征向量作为 β 的初值. 由于 $\partial E(Y | X) / \partial X = \beta_0 \{\partial E(Y | \beta_0^T X) / \partial (\beta_0^T X)\}^T$, 所以 M_{OPG} 的列空间等于 $\mathcal{S}_{E(Y|X)}$. 因此, 主梯度方法具有相合性. 同时, 由于仅涉及核均值估计和矩阵奇异值分解, 所以主梯度方法计算简便. 因此, 这一方法为上述交替迭代最小二乘算法提供了合适的初始值. 基于最小平均方差估计 $\mathcal{S}(\widehat{\beta})$, 文献 [85] 进一步将 (3.1) 中的权重项 $K_h(\|X_j - X_i\|_2)$ 替换为定义在低维实空间 $\mathcal{S}(\widehat{\beta})$ 上的 $K_h(\|\widehat{\beta}^T X_j - \widehat{\beta}^T X_i\|_2)$, 从而更好地抓取每一个 X_i 的有效邻域. 相应的方法被称为精炼最小平均方差估计方法 [85]. 通常, 研究者在中序贯地采用主梯度方法、最小平均方差估计方法、精炼最小平均

方差估计方法. 这一序贯估计过程也被简称为最小平均方差估计方法, 其中对窗宽 h 的两次调参均基于回归分析中经典的以最小均方误差为准则的交叉验证方法. 由于 $\mathcal{S}_{E(Y|X)}$ 的维数 d 决定了 (3.1) 中的参数数量, $\hat{E}(Y | \beta^T X)$ 与 d 的关系符合常见的偏差 - 方差权衡现象, 即 $\hat{E}(Y | \beta^T X)$ 在 d 被低估时会欠拟合, 而在 d 被高估时会过拟合. 因此, 交叉验证方法也适用于对 $\mathcal{S}_{E(Y|X)}$ 的定阶.

基于最小平均方差估计方法, 文献 [84] 将 Y 替换为对条件密度函数 $f(Y | X = x)$ 在一系列 x 上的取值, 再集成所得的对 $\mathcal{S}_{E\{f(Y|X=x)|X\}}$ 的估计, 从而将对 $\mathcal{S}_{E(Y|X)}$ 的估计改造为对 $\mathcal{S}_{Y|X}$ 的估计. 相应的方法被称为密度最小平均方差估计方法. 基于同一思路, 文献 [53] 将 Y 替换为一族 Fourier 变换 $\exp(itY)$, 从而将其构造的对 $\mathcal{S}_{E(Y|X)}$ 的估计改造为对 $\mathcal{S}_{Y|X}$ 的估计. 文献 [42] 将这一思路推广为一般化的集成充分降维方法论: 假设 $\Omega(Y)$ 是 $L_2(Y)$ 的一个稠密线性子空间的一组基, 则 $\mathcal{S}_{Y|X} = \mathcal{S}(\bigcup_{\omega(Y) \in \Omega(Y)} \mathcal{S}_{E\{\omega(Y)|X\}})$. 从而, 可先估计每一个 $\mathcal{S}_{E\{\omega(Y)|X\}}$, 再将其集成为对 $\mathcal{S}_{Y|X}$ 的估计. 由于 $\mathcal{S}_{Y|X}$ 是有限维线性空间, 在实际应用中研究者只需要采用 $\Omega(Y)$ 的一个有限子集即可. 作为特例, 若 $\Omega(Y)$ 是基于 Y 的 Borel 域的所有示性函数 $I(Y \in A)$ 的集合, 则可采用 Y 的 H 个互补的切片函数 $I(Y_S = 1), \dots, I(Y_S = H)$. 从这个角度上, 集成充分降维方法也解释了第 2 节中提到的关于估计 $\mathcal{S}_{E(Y|X)}$ 和估计 $\mathcal{S}_{Y|X}$ 的相关逆回归方法之间的联系, 如最小二乘方法与切片逆回归方法的联系, 以及主 Hessian 方向方法与切片平均方差估计方法之间的联系等.

因为最小平均方差估计方法与逆回归方法的出发点不同, 且具有应用范围上的优势和计算方便性上的劣势, 在前者发表以后的十年间, 研究者普遍认为两类方法互不相干. 它们之间的联系在 Luo 等 [64] 以及 Ma 和 Zhu [71-74] 等学者自 2012 年以来的一系列文献中才得以揭示, 并且被统一在更一般化的估计方程框架下. 对 $\mathcal{S}_{Y|X}$, 文献 [71] 给出了一类正则化估计方法:

$$\Lambda^T E_n[g(Y, \beta^T X) - \hat{E}\{g(Y, \beta^T X) | \beta^T X\}][\alpha(X) - \hat{E}\{\alpha(x) | \beta^T X\}] = 0, \quad (3.2)$$

其中 Λ 为使得 (3.2) 有解的适当矩阵, $E_n(\cdot)$ 表示样本均值, $g(\cdot, \cdot)$ 和 $\alpha(\cdot)$ 是任意的平方可积函数, 而 $\hat{E}\{g(Y, \beta^T X) | \beta^T X\}$ 和 $\hat{E}\{\alpha(x) | \beta^T X\}$ 的其中之一 (不必两者同时) 须在 $\mathcal{S}(\beta) = \mathcal{S}_{Y|X}$ 时是其真值的相合估计, 后者也被称为双重稳健性质 [71]. 在渐近意义下, 估计方程 (3.2) 包含了前述所有的逆回归方法. 例如, 取 $g(Y, \beta^T X) = \{I(Y_S = 1), \dots, I(Y_S = H)\}^T$ 或 $\{E(X | Y_S = 1), \dots, E(X | Y_S = H)\}^T$, 取 $\alpha(X) = X^T$, 并在线性条件 (2.2) 下利用线性回归得到 $E\{\alpha(X) | \beta^T X\}$ 的相合估计, 再基于上述双重稳健性质令 $\hat{E}\{g(Y, \beta^T X) | \beta^T X\} \equiv 0$, 则 (3.2) 是切片逆回归方法在渐近意义下的估计方程形式 (参见文献 [71, 73]). 当对 $E\{\alpha(X) | \beta^T X\}$ 假设一般化的参数模型或者半参数模型并基于双重稳健性质令 $\hat{E}\{g(Y, \beta^T X) | \beta^T X\} \equiv 0$ 时, (3.2) 则给出了前述文献 [21, 44] 对逆回归方法推广的渐近等价形式.

当充分降维的目标参数为 $\mathcal{S}_{E(Y|X)}$ 时, 通过将 (3.2) 中的 $g(Y, \beta^T X)$ 限制为 Y , 文献 [71] 同样给出了一类基于估计方程的半参数充分降维方法:

$$\Lambda^T E_n\{Y - \hat{E}(Y | \beta^T X)\}[\alpha(X) - \hat{E}\{\alpha(x) | \beta^T X\}] = 0, \quad (3.3)$$

其中 Λ 为使得 (3.3) 有解的适当矩阵. 当 $\alpha(X) = X$ 且 $\hat{E}\{\alpha(x) | \beta^T X\}$ 来自于局部线性估计时, 该方法在渐近意义下等同于最小平均方差估计方法 (收录于本文作者一篇未公开发表的论文); 当 $\alpha(X) = X$ 且线性条件 (2.2) 成立而 $\hat{E}\{\alpha(x) | \beta^T X\}$ 来自于线性回归时, 该方法等同于最小二乘估计; 当 $\alpha(X) = X^{\otimes 2}$ 且线性条件 (2.2) 和常数方差条件 (2.3) 同时成立时, 该方法等同于主 Hessian 方向方法. 对于前述一般的泛函 $T(Y | X)$, 文献 [64] 也构建了一般的估计方程框架以估计 $\mathcal{S}_{T(Y|X)}$.

除了揭示最小平均方差估计方法和逆回归方法的联系, 以及构建统一的充分降维方法框架外, 文献 [64, 72, 74] 等还为这类方法建立了相应的渐近理论, 从而证明了在一定的正则化条件下, 充分降维方

法可在 $Y|X$ 和 X 均为模型自由的前提下实现对 $\mathcal{S}_{Y|X}$ 或 $\mathcal{S}_{E(Y|X)}$ (或简单理解为 β_0) 的 $\sqrt{n}$ 相合估计. 相比较于逆回归方法因为限制 X 且形式简单而显然具有 $\sqrt{n}$ 相合性, 以及最小平均方差方法因为形式复杂而容易被认为不具 $\sqrt{n}$ 相合性, 这些渐近理论回答了关于充分降维子空间的估计精度的根本问题, 证明了低维结构假设 (1.1) 足以保证完全模型自由的精确统计分析. 在此基础上, 文献 [64, 72, 74] 分别对 $\mathcal{S}_{Y|X}$ 和 $\mathcal{S}_{E(Y|X)}$ 以及一般的 $\mathcal{S}_{T(Y|X)}$ 构造了渐近有效估计方法. 这些方法达到了半参数渐近有效界, 因而是理论上的最优方法. 类似于最小平均方差估计方法, 基于 (3.2) 的对 $\mathcal{S}_{Y|X}$ 的定阶问题以及基于 (3.3) 的对 $\mathcal{S}_{E(Y|X)}$ 的定阶问题也可利用交叉验证方法.

尽管具有理论上的诸多优势, 但是 (3.2) 和 (3.3) 的求解过程一般需要涉及非凸优化算法. 因此, 基于估计方程的充分降维方法在高维数据场景下会遭遇计算成本上的瓶颈. 这也是目前该类方法悬而未决的主要问题. 利用深度神经网络等机器学习工具来拟合 $E\{g(Y, \beta^T X) | \beta^T X\}$ 和 $E\{\alpha(X) | \beta^T X\}$, 也许会是一个有效的途径, 相关工作可参见文献 [37].

4 高维场景下的充分降维

如第 2 节所言, 在高维数据场景下, 逆回归方法是现实可行的充分降维方法. 简单的切片逆回归方法已在过去十几年得到了丰富的发展. 在渐近统计理论中, 高维数据是指 X 的维数 p 随着样本量 n 增加而趋于无穷的数据. 如果存在固定的 $k > 1$, 使得 p 趋于无穷的速度至少为 $\exp(n/k)$, 则这样的数据被称为超高维数据 [25].

在基因组学和生物学等许多应用场景中, X 对 Y 的作用普遍被认为或被验证具有稀疏性, 即存在 X 的一部分变量 X_A , 称为激活子集, 使得

$$Y \perp\!\!\!\perp X | X_A. \quad (4.1)$$

尽管 X_A 未知, 但在线性回归等统计模型中, 研究者已经证明了对激活子集 X_A 的变量选择过程不会对后续统计分析带来渐近意义下的显著影响 (oracle 性质). 因此, 在稀疏性假设 (4.1) 下, 可粗略认为只需要对 (Y, X_A) 这一低维数据建模, 从而提高数据分析精度. 结合 (1.1) 和 (4.1), 稀疏充分降维假设存在稀疏且低维的 $\mathcal{S}_{Y|X}$, 其任意基矩阵 β_0 中只有一部分行 (对应于 X_A) 非零. 稀疏充分降维方法的目的在于构造 $\mathcal{S}_{Y|X}$ 的稀疏且相合的估计.

在高维稀疏数据场景下, 基于惩罚最小二乘估计方法的线性回归分析得到了长足的发展 (参见文献 [24, 25, 82, 98, 108]), 并衍生出了稀疏主成分分析 [108] 等相关统计方法. 由于后者可直接推广为对一般矩阵型参数的稀疏估计, 文献 [51] 将其改造为稀疏逆回归方法, 将 $\widehat{M}$ 的特征分解改写为最小二乘形式, 再加入惩罚项. 文献 [5] 提出, 由于充分降维的目标参数是中心降维子空间, 而非其任一基矩阵 β_0 , 该惩罚项必须是以矩阵 $\beta \in \mathbb{R}^{p \times d}$ 的每一行为组别的分组惩罚函数, 从而保证所得的稀疏矩阵估计在任意可逆列变换下具有不变性. 例如, 可取惩罚函数为分组 LASSO (least absolute shrinkage and selection operator) 函数

$$\rho_\lambda(\beta) = \lambda \sum_{i=1}^p \|\beta_i\|_2, \quad (4.2)$$

其中 λ 为该函数的权重, β_i 为 β 的第 i 行, $\|\cdot\|_2$ 表示向量的 Euclid 范数. 由于逆回归方法的候选矩阵 M 普遍涉及 X 的协方差矩阵的逆矩阵 Σ^{-1} , 此类直接基于 $\widehat{M}$ 的稀疏逆回归方法无法应用于 $p > n$ 的情形.

鉴于上述高维惩罚最小二乘方法的蓬勃发展, 稀疏逆回归方法的另一思路是将 M (而非其特征分解) 重写为最小二乘的形式. 例如, 由 (2.1) 可得, 切片逆回归方法可被重写为

$$\Sigma B_0 = E(X \vec{Y}_S), \quad (4.3)$$

其中 $\vec{Y}_S = \{I(Y_S = 1), \dots, I(Y_S = H)\}$. 显然, $B_0 \in \mathbb{R}^{p \times H}$ 具有与 M_{SIR} 相同的列空间, 因此在线性条件 (2.2) 以及其他正则化条件下同样生成了 $\mathcal{S}_{Y|X}$. 通过将 (4.3) 进一步改写为

$$\min_{B \in \mathbb{R}^{p \times H}} E(\|\vec{Y}_S - B^T X\|_2^2), \quad (4.4)$$

切片逆回归方法也具有最小二乘的形式. 由于 H 的取值必须大于 d , 为减少参数数量, 文献 [75] 将 (4.4) 中的参数 B 限定为 βA , 其中 $\beta \in \mathbb{R}^{p \times d}$, $A \in \mathbb{R}^{d \times H}$, 再引入惩罚项 (4.2). 相应的惩罚最小二乘估计在超高维数据情形下依然具有相合性. 然而, 由于目标函数是 (β, A) 的非凸函数, 该方法在计算上存在劣势. 文献 [81] 部分解决了这一问题, 其思路是以 β_0 的投影矩阵作为目标参数, 以所有 p 维半正定矩阵的集合作为参数空间, 从而实现凸优化. 由于该方案中优化问题的参数数量从 $O(p)$ 提升到 $O(p^2)$, 因此在计算性能上仍有不足. 最近, 文献 [97] 直接利用 (4.4) 构造了以 $B \in \mathbb{R}^{p \times H}$ 为参数的惩罚凸优化估计方法, 实现了超高维场景下对 $\mathcal{S}_{Y|X}$ 的简便易算的相合估计. 此外, 在 $p = o(n^2)$ 的高维场景下, 文献 [57] 也基于类似思路提出了 LASSO 切片逆回归方法. 关于高阶逆回归方法的稀疏高维调整可参见文献 [69, 75, 83], 此处不再赘述.

早期的超高维稀疏数据分析并非基于惩罚最小二乘估计方法, 而是采用特征筛选方法^[25, 104]. 在实际应用中, 针对 p 远大于 n 的情形 (如 $p \approx 10^6$, $n \approx 10^3$), 特征筛选方法因为形式简单, 目前依然是主流的数据降维方法. 如果在 (1.1) 中限制 β_0 为 p 维单位矩阵的列的子集, 则充分降维假设 (1.1) 等价于数据稀疏性假设 (4.1). 相应地, 前述充分降维方法可被改造为模型自由的变量选择方法或者特征筛选方法. 相关工作包括基于切片逆回归方法和 Dantzig 选择器的迹追踪方法^[96], 基于文献 [96] 并利用对 Σ^{-1} 的 l_1 最小化估计, 从而改进了 M_{SIR} 估计的边际切片逆回归方法^[95], 以及基于 Σ 的稀疏性假设的 DT-切片逆回归方法^[56]. 在这些模型自由方法的加持下, 可对所得的低维数据进一步应用常规的充分降维方法, 从而分两步实现超高维稀疏数据的充分降维.

关于上述稀疏充分降维方法的发展, 在综述文献 [54] 中也有详尽的总结与讨论, 供读者参考.

在经济学等一些高维数据分析场景中, 稀疏性假设 (4.1) 被普遍认为不再成立. 例如, 如果用某地过去一段时间的日均房价 (X) 来预测其明天的房价 (Y), 则 X 中各变量对 Y 的影响理应呈现按时间倒序而总体缓慢衰减但全部非零的特点. 针对这些高维非稀疏数据, 文献 [12] 讨论了以切片逆回归方法为代表的一类充分降维方法的渐近性质; 文献 [56] 给出了切片逆回归方法的适用范围, 即 $p = o(n)$; 对一类弱奇异性高维非稀疏数据, 文献 [26] 提出了先应用主成分分析再应用切片逆回归方法的两步充分降维方法; 基于文献 [26, 66] 放宽了对数据的弱奇异性限制, 并将第二步中的切片逆回归方法推广为一般逆回归方法.

本文作者认为, 由于高维数据分析的精髓在于对数据结构作合理假设以大幅降低统计建模所需要的参数数量 (暂时忽略机器学习中的良性过拟合现象), 充分降维方法在不同的高维场景下仍有很大的发展空间.

5 充分降维在因果推断中的应用

如第 1 节所述, 由于观察性数据在实际应用中普遍存在, 因果推断是统计学研究的热门问题. 假

设关注某二分干预变量 (原因) 对某数值型响应变量 (结果) 的因果作用, 例如, 是否服药对特定患者的康复时间的影响, 以及是否施行新政对未来经济走势的影响等, 对观察性数据直接进行统计分析往往会导致错误甚至相反的因果结论. 因此, 因果推断也以其复杂性闻名. 记二分干预变量为 T , 其中 $T = 0$ 表示相应观测点未接受干预 (对照组), $T = 1$ 表示相应观测点接受了干预 (实验组, 尽管这不是随机实验). 在潜在结果变量框架这一因果推断研究的主流范式下, 将响应变量 Y 理解为两个潜在的结果变量 Y_0 和 Y_1 , 其中 Y_0 表示观测点未接受干预 ($T = 0$) 时对应的结果, 而 Y_1 表示观测点接受了干预 ($T = 1$) 时对应的结果. 干预变量 T 对响应变量 Y 的因果效应可表示为 $Y_1 - Y_0$. 为了排除随机因素的干扰, 通常关注平均因果效应 $E\{Y_1 - Y_0\}$ 或者条件因果效应 $E\{Y_1 - Y_0 | X\}$. 后者中的 X 表示观测点的个体特征, 因此后者也常常是个性化因果推断的目标参数.

由于数据中的每个观测点只能取值 $T = 0$ 或 $T = 1$, 两个潜在结果变量 Y_0 和 Y_1 均部分缺失. 为了正则化它们的缺失机制, 研究者通常采用可忽略性假设 (参见文献 [77]), 即对于 $t = 0, 1$, 有

$$Y_t \perp\!\!\!\perp T | X. \quad (5.1)$$

在此假设下, 每个观测点取值 $T = 0$ 或者 $T = 1$ 与该观测点对 Y_0 和 Y_1 的取值均无关. 此外, 研究者还需要假设公共支撑条件: 令 $\pi(X) = P(T = 1 | X)$ 为给定 X 后 $T = 1$ 的概率, 则

$$P(0 < \pi(X) < 1) = 1. \quad (5.2)$$

通常, $\pi(X)$ 也被称为倾向得分函数. 基于 (5.1) 和 (5.2), 条件因果效应可被重写为

$$E(Y_1 | X, T = 1) - E(Y_0 | X, T = 0) \quad \text{或} \quad E\left\{\frac{TY_1}{\pi(X)} \middle| X\right\} - E\left\{\frac{(1-T)Y_0}{1-\pi(X)} \middle| X\right\}, \quad (5.3)$$

从而可由观测到的数据得到相合估计. (5.3) 中的前者被称为结果变量回归分析^[76], 后者被称为逆概率加权分析; 两者也可结合, 实现对条件因果效应的双重稳健估计^[2] 以及渐近有效估计^[78]. 对 (5.3) 取均值后, 这些方法也可用于估计平均因果效应.

与一般的统计学问题相同, 当 X 的维数较大时, 基于 (5.3) 的因果推断方法会遭受“维数灾难”. 因为充分降维方法具有模型自由的特点, 所以恰当地应用充分降维可改进因果推断的精确度和可信度. 在上述可忽略性假设和公共支撑条件下, 相关研究始于文献 [29], 其提出对 $T | X$ 应用充分降维, 从而获得对倾向得分函数 $\pi(X)$ 的低维非参数估计, 再利用 (5.3) 中的逆概率加权分析估计平均因果效应. 此后, 文献 [68] 对 (5.3) 中的结果变量回归分析应用充分降维, 即在相应的低维结构假设下, 分别估计 $E(Y_1 | X, T = 1)$ 和 $E(Y_0 | X, T = 0)$ 的中心均值降维子空间. 这一工作探讨了在渐近意义下充分降维对因果效应估计的改进作用^[102], 并引发了在高维数据场景下, 利用充分降维实现精确可信的因果推断的研究热潮 (参见文献 [36, 65, 67, 69, 69]). 文献 [36, 69] 对 $T | X$, $Y_0 | X$ 和 $Y_1 | X$ 均应用了充分降维, 再合并所得低维自变量; 为了提高降维效率, 文献 [65] 构造了新的结果变量 $TY_1/\pi(X) - (1-T)Y_0/\{1-\pi(X)\}$, 再对此应用充分降维; 文献 [69] 构造了稀疏主 Hessian 方向估计方法, 并以此实现了超高维数据情形下因果效应的相合估计.

当研究者只关注平均因果效应时, 类似 (5.1), 降维所得的低维自变量只需要满足

$$Y_t \perp\!\!\!\perp T | \beta_t^T X, \quad t = 0, 1. \quad (5.4)$$

从而, 可效仿 (5.3) 将平均因果效应写为 $E\{E(Y_1 | \beta_1^T X, T = 1) - E(Y_0 | \beta_0^T X, T = 0)\}$. 在可忽略性假设 (5.1) 下, $\mathcal{S}_{Y_t|X}$ 的任意基矩阵以及 $\mathcal{S}_{T|X}$ 的任意基矩阵均满足 (5.4). 因此, 上一段提到的诸多降维

方法都可看作是以 $Y_t \perp\!\!\!\perp X \mid \beta^T X$ 或 $T \perp\!\!\!\perp X \mid \beta^T X$ 为途径实现 (5.4) 这一最终目标. 然而, 这些方法需要在 (5.4) 之外假设 $Y_t \mid X$ 或 $T \mid X$ 的低维结构, 而且所得低维自变量无法保证达到 (5.4) 所允许的最低维数, 从而无法最大化降维所带来的收益 (详见文献 [61]). 基于这些局限性, 文献 [61] 将 (5.4) 视为全新的降维问题, 并初步构建了相应的理论框架和估计方法. 由于满足 (5.4) 的最低维线性空间通常不唯一, 文献 [61] 引入了局部有效降维子空间的集合作为目标参数族. 在类似线性条件 (2.2) 和常数方差条件 (2.3) 的假设下, 文献 [61] 指出这一目标参数族可等价表示为方程

$$M_{Y_t}^T \Sigma P(\Sigma, \beta) M_T = M_Y^T \Sigma M_T \quad (5.5)$$

的解空间, 其中 M_{Y_t} 和 M_T 分别是对 $Y_t \mid X$ 和 $T \mid X$ 应用逆回归方法所得的候选矩阵. 基于 (5.5) 的降维方法被称为双重逆回归方法. 该类方法具有对 $(Y_0, Y_1, T) \mid X$ 模型自由、不假设 $(Y_0, Y_1) \mid X$ 或者 $T \mid X$ 的低维结构, 以及所得低维自变量达到 (5.4) 所允许的最低维数等优点. 文献 [61] 同时讨论了局部有效降维子空间的唯一性的充分必要条件. 基于双重逆回归方法, 对局部有效降维子空间的定阶问题可应用文献 [62, 63] 等方法. 由于 (5.4) 比只涉及单个响应变量的一般充分降维 (1.1) 更加复杂, 相关工作还处于初始阶段.

由于可将潜在结果变量 Y_t 看成普通的存在部分缺失的响应变量, 以及直观上缺失数据 $Y_t I(T = t)$ 与删失数据 $\min\{Y, C\}$ 具有相似性, 上述工作与缺失数据分析场景下的充分降维^[20, 22, 31]以及生存分析场景下的充分降维^[19]存在天然联系. 近年来, 因果推断方法的发展着重于融入更广泛的现实场景, 包括可忽略性假设 (5.1) 不成立或者干预变量 T 为连续型变量等复杂情形. 对于前者, 常见的思路包括引入工具变量^[1]或代理变量^[16]等. 充分降维需作适当调整, 以契合相应的数据结构假设, 从而使这些方法在保留原有优势的基础上, 适用于高维数据场景. 本文作者认为这也是充分降维在因果推断场景下新的发展方向.

6 非线性充分降维

如本文开头所述, 充分降维的另一个发展方向是非线性充分降维^[28, 39, 46], 即假设存在低维函数 $R(X)$, 使得

$$Y \perp\!\!\!\perp X \mid R(X). \quad (6.1)$$

这一假设推广了线性充分降维假设 (1.1), 并在同方差回归分析等情形自动成立 ($R(X) = E(Y \mid X)$), 因此具有广泛的应用场景. 类似 (1.1), (6.1) 对 $R(X)$ 的任意可逆变换均成立. 为保证目标参数的可识别性, 文献 [39] 将 (6.1) 视为对 $R(X)$ 生成的 σ 域的刻画, 并在广泛的正则化条件下证明了满足 (6.1) 的最小 σ 域的唯一性. 这一 σ 域被称为中心降维 σ 域. 它是非线性充分降维的目标参数, 也是对线性充分降维中的目标参数 $S_{Y \mid X}$ 的自然推广.

类似线性充分降维, 为方便估计中心降维 σ 域, 非线性充分降维方法通常引入中间参数. 由于 $R(X)$ 可理解为连接 X 的函数空间与 Y 的函数空间的桥梁, 这一中间参数通常是两个函数空间之间的算子. 该算子须满足其值域内所有函数所生成的 σ 域等于中心降维 σ 域. 例如, 在线性充分降维的特殊情形下, 切片逆回归方法对应的算子将 $I(Y_S = i)$ 映射为 X 的线性函数 $X^T \Sigma^{-1} E\{X I(Y_S = i)\}$. 因此, 这一思路也是线性充分降维中逆回归方法的自然推广.

目前, 主流的非线性充分降维方法将 X 的函数空间取为再生核 Hilbert 空间, 从而构造相应的算子 (参见文献 [28, 34, 39, 41, 42, 88]). 由于再生核 Hilbert 空间特殊的内积运算, 上述线性算子在 Riesz

表示定理下具有简单的 (可数维) 矩阵表示形式. 这一表示形式也常被称为核技巧 (the kernel trick). 例如, 记 $\mathcal{K}_X$ 为 X 的某内积函数空间, 记 $\mathcal{K}_Y$ 为 Y 的某函数空间, 则广义 (非线性) 切片逆回归方法^[39,41] 对应的算子为 $\Sigma_{XX}^{-1}\Sigma_{XY}A_Y : \mathcal{K}_Y \rightarrow \mathcal{K}_X$, 其中算子 $\Sigma_{XX} : \mathcal{K}_X \rightarrow \mathcal{K}_X$ 和 $\Sigma_{XY} : \mathcal{K}_Y \rightarrow \mathcal{K}_X$ 的定义分别为, 对任意的 $f(X), g(X) \in \mathcal{K}_X$ 和任意的 $\ell(Y) \in \mathcal{K}_Y$, 都有

$$\langle f(X), \Sigma_{XX}g(X) \rangle_X = E\{f(X)g(X)\}, \quad \langle f(X), \Sigma_{XY}\ell(Y) \rangle_X = E\{f(X)\ell(Y)\},$$

而 A_Y 是任意定义的 $\mathcal{K}_Y \rightarrow \mathcal{K}_Y$ 的可逆算子. 如果 $\mathcal{K}_X$ 是由核函数 $k_X(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ 生成的 X 的再生核 Hilbert 空间, 且 $\mathcal{K}_Y$ 是由核函数 $k_Y(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ 生成的 Y 的再生核 Hilbert 空间, 且另取 $A_Y \equiv \Sigma_{YY}^{-1/2}$ 使得 $\langle f(Y), A_Y^{-1}A_Y^{-1}g(Y) \rangle_Y = E\{f(Y)g(Y)\}$ 对于任意 $f(Y), g(Y) \in \mathcal{K}_Y$ 均成立, 则算子 $\Sigma_{XX}^{-1}\Sigma_{XY}A_Y$ 在坐标系 $\{k_X(\cdot, X_1), \dots, k_X(\cdot, X_n)\}, \{k_Y(\cdot, Y_1), \dots, k_Y(\cdot, Y_n)\}$ 下具有矩阵表示形式: $(\mathbb{G}_X + \lambda_n \mathbb{I}_n)^{-2} \mathbb{G}_X (\mathbb{G}_X + \lambda_n \mathbb{I}_n)^{-1} \mathbb{G}_X \mathbb{G}_Y (\mathbb{G}_Y + \lambda_n \mathbb{I}_n)^{-1} \mathbb{G}_Y^{1/2}$, 其中 λ_n 是趋于零的常数序列; $\mathbb{I}_n$ 表示 n 维单位矩阵; $\mathbb{G}_X = \mathbb{Q}\mathbb{K}_X\mathbb{Q}$, $\mathbb{G}_Y = \mathbb{Q}\mathbb{K}_Y\mathbb{Q}$, 这里 $\mathbb{K}_X, \mathbb{K}_Y$ 和 $\mathbb{Q}$ 均为 $n \times n$ 维矩阵, $\mathbb{K}_X$ 的第 (i, j) 位置元素为 $k_X(X_i, X_j)$, $\mathbb{K}_Y$ 的第 (i, j) 位置元素为 $k_Y(Y_i, Y_j)$, $\mathbb{Q} = \mathbb{I}_n - n^{-1}(1, \dots, 1)^T(1, \dots, 1)$. 广义切片平均方差估计方法等其他非线性充分降维方法的算子形式和矩阵表示形式也可类似得出 (参见文献 [39,60]).

从核技巧的角度出发, 非线性充分降维理论与方法也可粗略地理解为先对 X 升维, 再进行线性充分降维: 假设 (6.1) 中的 $R(X)$ 属于 X 的某线性函数空间 (该空间通常包含 X 的非线性函数, 因此维数远大于 p), 则先将自变量从 X 升维到该函数空间的一组基函数, 如上述 $k_X(\cdot, X_1), \dots, k_X(\cdot, X_n)$, 再对升维后的数据应用逆回归方法等线性充分降维方法.

不同于线性充分降维理论, 非线性充分降维理论目前仍有许多基础问题仍待探讨. 例如, 在以中心降维 σ 域为目标参数的框架下, 如何定义 (6.1) 中 $R(X)$ 的维数? 当中心降维 σ 域无法用低维光滑函数生成时, 如何平衡 $R(X)$ 与给定函数空间 $\mathcal{K}_X$ 的近似误差以及 $\mathcal{K}_X$ 自身复杂性所引起的估计误差 (也称偏差 - 复杂度权衡)? 此外, 非线性充分降维方法与机器学习中的表征学习^[3] 和流形学习^[58] 密切相关. 机器学习算法的最新进展, 如深度神经网络等工具, 也可替代上述再生核 Hilbert 空间, 从而引出计算性能更好的非线性充分降维方法^[35].

7 讨论与展望

因篇幅所限, 本文未能回顾所有充分降维研究的重要工作, 包括但不限于如下工作.

- 在逆回归方法的最小二乘等价形式中, L_2 损失函数也可换成其他损失函数, 如 L_1 损失函数、hinge 损失函数和 Kullback-Leibler 散度, 从而得到稳健切片逆回归方法^[23]、主支持向量机^[42] 和加权主支持向量机^[80] 等对 X 中的异常值稳健的充分降维方法, 以及基于似然函数估计的充分降维方法^[11,91,92] 及更一般化的方法类^[101]; 另一种稳健逆回归方法的思路则是在保持其相关系数矩阵不变的前提下, 对 X 作适当的压缩变换, 详见文献 [23,59].

- 当数据的观测点具有复杂的网络结构而非相互独立时, 文献 [103] 建立了逆回归方法的渐近理论.

- 当 X 中的部分变量不参与降维过程, 也就是 $X = (W, Z)$ 而降维所得低维自变量为 $(\beta^T W, Z)$ 时, 充分降维可调整为部分充分降维^[6,27,43]. 这一概念可被推广至低维变量为 $(\beta^T W, \gamma^T Z)$ 的分组充分降维^[52].

• 当 X 包含随机观测误差, 使得我们只能观测到 $X + \varepsilon$ 时, 文献 [48] 提出了 $\mathcal{S}_{Y|X}$ 的不变性理论, 并据此提出了 $\mathcal{S}_{Y|X}$ 可估的充分条件; 这一不变性理论在文献 [26, 66] 中被推广;

• 当 X 中部分变量存在缺失值时, 基于类似 (5.1) 的可忽略性假设, 文献 [53, 86] 利用逆概率加权调整了逆回归方法和最小平均方差估计方法.

• 当 X 具有矩阵或者张量形式时, 充分降维所得自变量可调整为 $\alpha^T X \beta$ 等, 详见文献 [18, 41, 45].

• 当 X 为函数型随机变量时, 文献 [40, 46] 等提出了相应的非线性充分降维理论与方法.

• 当 Y 为某非 Euclid 度量空间中的随机个体, 如概率分布时, 文献 [94, 99, 100] 等提出了相应的充分降维理论与方法.

• 除因果推断外, 充分降维也被诸多学者用于构造模型自由的假设检验方法, 如关于 X 中各变量对 Y 作用的显著性的假设检验方法^[9] 和在高维情形下的拓展 [30], 以及关于单指标参数模型的假设检验方法^[32] 等.

在此, 本文作者也希望读者能海涵本文的所有纰漏.

在高维复杂数据日益普及的背景下, 充分降维依然是实现模型自由精确统计分析的重要途径. 在机器学习研究的推动下, 非线性充分降维理论与方法似乎将成为充分降维发展的主流之一. 同时, 如何与因果推断等诸多统计学研究问题有效结合, 也将丰富充分降维的理论与方法, 并提高充分降维的应用价值. 如第 5 节所述, 后者尚处于起步阶段.

参考文献

- 1 Angrist J D, Imbens G W, Rubin D B. Identification of causal effects using instrumental variables. *J Amer Statist Assoc*, 1996, 91: 444–455
- 2 Bang H, Robins J M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 2005, 61: 962–973
- 3 Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE Trans Pattern Anal Mach Intell*, 2013, 35: 1798–1828
- 4 Bura E, Yang J. Dimension estimation in sufficient dimension reduction: A unifying approach. *J Multivariate Anal*, 2011, 102: 130–142
- 5 Chen X, Zou C, Cook R D. Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann Statist*, 2010, 6: 3696–3723
- 6 Chiaromonte F, Cook R D, Li B. Sufficient dimension reduction in regressions with categorical predictors. *Ann Statist*, 2002, 30: 475–497
- 7 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*. Anaheim: Amer Statist Assoc, 1994, 18–25
- 8 Cook R D. *Regression Graphics*. New York: John Wiley & Sons, 1998
- 9 Cook R D. Testing predictor contributions in sufficient dimension reduction. *Ann Statist*, 2004, 32: 1062–1092
- 10 Cook R D, Forzani L. Principal fitted components for dimension reduction in regression. *Statist Sci*, 2008, 23: 485–501
- 11 Cook R D, Forzani L. Likelihood-based sufficient dimension reduction. *J Amer Statist Assoc*, 2009, 104: 197–208
- 12 Cook R D, Forzani L, Rothman A J. Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *Ann Statist*, 2012, 40: 353–384
- 13 Cook R D, Li B. Dimension reduction for conditional mean in regression. *Ann Statist*, 2002, 30: 455–474
- 14 Cook R D, Ni L. Sufficient dimension reduction via inverse regression. *J Amer Statist Assoc*, 2005, 100: 410–428
- 15 Cook R D, Weisberg S. Sliced inverse regression for dimension reduction (with discussion). *J Amer Statist Assoc*, 1991, 86: 316–342
- 16 Cui Y, Pu H, Shi X, et al. Semiparametric proximal causal inference. *J Amer Statist Assoc*, 2024, 119: 1348–1359
- 17 Diaconis P, Freedman D. Asymptotics of graphical projection pursuit. *Ann Statist*, 1984, 12: 793–815
- 18 Ding S, Cook R D. Tensor sliced inverse regression. *J Multivariate Anal*, 2015, 133: 216–231
- 19 Ding S, Qian W, Wang L. Double-slicing assisted sufficient dimension reduction for high-dimensional censored data. *Ann Statist*, 2020, 48: 2132–2154
- 20 Ding X, Wang Q. Fusion-refinement procedure for dimension reduction with missing response at random. *J Amer Statist Assoc*, 2011, 106: 1193–1207

- 21 Dong Y, Li B. Dimension reduction for non-elliptically distributed predictors: Second-order methods. *Biometrika*, 2010, 97: 279–294
- 22 Dong Y, Xia Q, Tang C Y, et al. On sufficient dimension reduction with missing responses through estimating equations. *Comput Statist Data Anal*, 2018, 126: 67–77
- 23 Dong Y, Yu Z, Zhu L. Robust inverse regression for dimension reduction. *J Multivariate Anal*, 2015, 134: 71–81
- 24 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 25 Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B Stat Methodol*, 2008, 70: 849–911
- 26 Fan J, Xue L, Yao J. Sufficient forecasting using factor models. *J Econometrics*, 2017, 201: 292–306
- 27 Feng Z, Wen X M, Yu Z, et al. On partial sufficient dimension reduction with applications to partially linear multi-index models. *J Amer Statist Assoc*, 2013, 108: 237–246
- 28 Fukumizu K, Bach F R, Jordan M I. Kernel dimension reduction in regression. *Ann Statist*, 2009, 36: 1871–1905
- 29 Ghosh D. Propensity score modelling in observational studies using dimension reduction methods. *Statist Probab Lett*, 2011, 81: 813–820
- 30 Guo X, Li R, Zhang Z, et al. Model-free statistical inference on high-dimensional data. *J Amer Statist Assoc*, 2024, 119: 1–12
- 31 Guo X, Wang T, Xu W, et al. Dimension reduction with missing response at random. *Comput Statist Data Anal*, 2014, 69: 228–242
- 32 Guo X, Wang T, Zhu L. Model checking for parametric single-index models: A dimension reduction model-adaptive approach. *J R Stat Soc Ser B Stat Methodol*, 2016, 78: 1013–1035
- 33 Hall P, Li K C. On almost linearity of low dimensional projections from high dimensional data. *Ann Statist*, 1993, 47: 867–889
- 34 Hsing T, Ren H. An RKHS formulation of the inverse regression dimension-reduction problem. *Ann Statist*, 2009, 37: 726–755
- 35 Huang J, Jiao Y, Liao X, et al. Deep dimension reduction for supervised representation learning. *IEEE Trans Inform Theory*, 2024, 70: 3583–3598
- 36 Huang M Y, Chan K C G. Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika*, 2017, 104: 583–596
- 37 Kapla D F, Fertl L, Bura E. Fusing sufficient dimension reduction with neural networks. *Comput Statist Data Anal*, 2022, 168: 107390
- 38 Kong E, Xia Y. An adaptive composite quantile approach to dimension reduction. *Ann Statist*, 2014, 42: 1657–1688
- 39 Lee K Y, Li B, Chiaromonte F. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Ann Statist*, 2013, 41: 221–249
- 40 Lee K Y, Li L. Functional sufficient dimension reduction through average Fréchet derivatives. *Ann Statist*, 2022, 50: 904–929
- 41 Li B. *Sufficient Dimension Reduction: Methods and Applications with R*. Boca Raton: CRC Press, 2018
- 42 Li B, Artemiou A, Li L. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann Statist*, 2011, 39: 3182–3210
- 43 Li B, Cook R D, Chiaromonte F. Dimension reduction for the conditional mean in regressions with categorical predictors. *Ann Statist*, 2003, 31: 1636–1668
- 44 Li B, Dong Y. Dimension reduction for nonelliptically distributed predictors. *Ann Statist*, 2009, 37: 1272–1298
- 45 Li B, Kim M K, Altman N. On dimension folding of matrix- or array-valued statistical objects. *Ann Statist*, 2010, 38: 1094–1121
- 46 Li B, Song J. Dimension reduction for functional data based on weak conditional moments. *Ann Statist*, 2022, 50: 107–128
- 47 Li B, Wang S. On directional regression for dimension reduction. *J Amer Statist Assoc*, 2007, 102: 997–1008
- 48 Li B, Yin X. On surrogate dimension reduction for measurement error regression: An invariance law. *Ann Statist*, 2007, 35: 2143–2172
- 49 Li K C. Sliced inverse regression for dimension reduction. *J Amer Statist Assoc*, 1991, 86: 316–327
- 50 Li K C, Duan N. Regression analysis under link violation. *Ann Statist*, 1989, 17: 1009–1052
- 51 Li L. Sparse sufficient dimension reduction. *Biometrika*, 2007, 94: 603–613
- 52 Li L, Li B, Zhu L X. Groupwise dimension reduction. *J Amer Statist Assoc*, 2010, 105: 1188–1201
- 53 Li L, Lu W. Sufficient dimension reduction with missing predictors. *J Amer Statist Assoc*, 2008, 103: 822–831
- 54 Li L, Wen X M, Yu Z. A selective overview of sparse sufficient dimension reduction. *Statist Theory Related Fields*, 2020, 4: 121–133
- 55 Li Y, Zhu L X. Asymptotics for sliced average variance estimation. *Ann Statist*, 2007, 35: 41–69

- 56 Lin Q, Zhao Z, Liu J S. On consistency and sparsity for sliced inverse regression in high dimensions. *Ann Statist*, 2018, 46: 580–610
- 57 Lin Q, Zhao Z, Liu J S. Sparse sliced inverse regression via lasso. *J Amer Statist Assoc*, 2019, 114: 1726–1739
- 58 Lin T, Zha H B. Riemannian manifold learning. *IEEE Trans Pattern Anal Mach Intell*, 2008, 30: 796–809
- 59 Luo R, Wang H, Tsai C L. Contour projected dimension reduction. *Ann Statist*, 2009, 37: 3743–3778
- 60 Luo W. On the second-order inverse regression methods for a general type of elliptical predictors. *Statist Sinica*, 2018, 28: 1415–1436
- 61 Luo W. On efficient dimension reduction with respect to the interaction between two response variables. *J R Stat Soc Ser B Stat Methodol*, 2022, 84: 269–294
- 62 Luo W, Li B. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 2016, 103: 875–887
- 63 Luo W, Li B. On order determination by predictor augmentation. *Biometrika*, 2021, 108: 557–574
- 64 Luo W, Li B, Yin X. On efficient dimension reduction with respect to a statistical functional of interest. *Ann Statist*, 2014, 42: 382–412
- 65 Luo W, Wu W, Zhu Y. Learning heterogeneity in causal inference using sufficient dimension reduction. *J Causal Inference*, 2019, 7: 20180015
- 66 Luo W, Xue L, Yao J, et al. Inverse moment methods for sufficient forecasting using high-dimensional predictors. *Biometrika*, 2022, 109: 473–487
- 67 Luo W, Zhu Y. Matching using sufficient dimension reduction for causal inference. *J Bus Econom Statist*, 2020, 38: 888–900
- 68 Luo W, Zhu Y, Ghosh D. On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, 2017, 104: 51–65
- 69 Ma S, Zhu L P, Zhang Z, et al. A robust and efficient approach to causal inference based on sparse sufficient dimension reduction. *Ann Statist*, 2019, 47: 1505–1535
- 70 Ma Y, Zhang X. A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika*, 2015, 102: 409–420
- 71 Ma Y, Zhu L P. A semiparametric approach to dimension reduction. *J Amer Statist Assoc*, 2012, 107: 168–179
- 72 Ma Y, Zhu L P. Efficient estimation in sufficient dimension reduction. *Ann Statist*, 2013, 41: 250–268
- 73 Ma Y, Zhu L P. Efficiency loss and the linearity condition in dimension reduction. *Biometrika*, 2013, 100: 371–383
- 74 Ma Y, Zhu L P. On estimation efficiency of the central mean subspace. *J R Stat Soc Ser B Stat Methodol*, 2014, 76: 885–901
- 75 Qian W, Ding S, Cook R D. Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *J Amer Statist Assoc*, 2019, 114: 1277–1290
- 76 Robins J. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Math Model*, 1986, 7: 1393–1512
- 77 Rosenbaum P R, Rubin D B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983, 70: 41–55
- 78 Rotnitzky A, Lei Q, Sued M, et al. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 2012, 99: 439–456
- 79 Setodji C M, Cook R D. *K*-means inverse regression. *Technometrics*, 2004, 46: 421–429
- 80 Shin S J, Wu Y, Zhang H H, et al. Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*, 2017, 104: 67–81
- 81 Tan K, Wang Z, Zhang T, et al. A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika*, 2018, 105: 769–782
- 82 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol*, 1996, 58: 267–288
- 83 Wang C J B, Zhu L P. Penalized interaction estimation for ultrahigh dimensional quadratic regression. *Statist Sinica*, 2021, 31: 1549–1570
- 84 Xia Y. A constructive approach to the estimation of dimension reduction directions. *Ann Statist*, 2007, 35: 2654–2690
- 85 Xia Y, Tong H, Li W K, et al. An adaptive estimation of dimension reduction space. *J R Stat Soc Ser B Stat Methodol*, 2002, 64: 363–410
- 86 Yang X, Wang Q. Sufficient dimension reduction under dimension-reduction-based imputation with predictors missing at random. *Statist Sinica*, 2019, 29: 1751–1777
- 87 Ye Z, Weiss R E. Using the bootstrap to select one of a new class of dimension reduction methods. *J Amer Statist Assoc*, 2003, 98: 968–979
- 88 Yeh Y-R, Huang S-Y, Lee Y-J. Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Trans Knowl Data Eng*, 2009, 21: 1590–1603
- 89 Yin X. Estimating central subspaces via inverse third moments. *Biometrika*, 2003, 90: 113–125

- 90 Yin X, Cook R D. Dimension reduction for the conditional k th moment in regression. *J R Stat Soc Ser B Stat Methodol*, 2002, 64: 159–175
- 91 Yin X, Cook R D. Direction estimation in single-index regressions. *Biometrika*, 2005, 92: 371–384
- 92 Yin X, Li B. Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann Statist*, 2011, 106: 3392–3416
- 93 Yin X, Li B, Cook R D. Successive direction extraction for estimating the central subspace in a multiple-index regression. *J Multivariate Anal*, 2008, 99: 1733–1757
- 94 Ying C, Yu Z. Fréchet sufficient dimension reduction for random objects. *Biometrika*, 2022, 109: 975–992
- 95 Yu Z, Dong Y, Shao J. On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *Ann Statist*, 2016, 44: 2594–2623
- 96 Yu Z, Dong Y, Zhu L X. Trace pursuit: A general framework for model-free variable selection. *J Amer Statist Assoc*, 2016, 111: 813–821
- 97 Zeng J, Mai Q, Zhang X. Subspace estimation with automatic dimension and variable selection in sufficient dimension reduction. *J Amer Statist Assoc*, 2024, 119: 343–355
- 98 Zhang C H. Nearly unbiased variable selection under minimax concave penalty. *Ann Statist*, 2010, 38: 894–942
- 99 Zhang Q, Li B, Xue L. Nonlinear sufficient dimension reduction for distribution-on-distribution regression. *J Multivariate Anal*, 2024, 202: 105302
- 100 Zhang Q, Xue L, Li B. Dimension reduction for Fréchet regression. *J Amer Statist Assoc*, 2023, 118: 1–15
- 101 Zhang X, Mai Q, Zou H. Maximum separation subspace in sufficient dimension reduction with categorical response. *J Mach Learn Res*, 2020, 21: 1–36
- 102 Zhang Y, Shao J, Yu M, et al. Impact of sufficient dimension reduction in nonparametric estimation of causal effect. *Statist Theory Related Fields*, 2018, 2: 89–95
- 103 Zhao J, Liu X, Wang H, et al. Dimension reduction for covariates in network data. *Biometrika*, 2021, 109: 85–102
- 104 Zhu L P, Li L, Li R, et al. Model-free feature screening for ultrahigh-dimensional data. *J Amer Statist Assoc*, 2011, 106: 1464–1475
- 105 Zhu L P, Miao B, Peng H. On sliced inverse regression with high-dimensional covariates. *J Amer Statist Assoc*, 2006, 101: 630–643
- 106 Zhu L P, Wang T, Zhu L, et al. Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, 2010, 97: 295–304
- 107 Zhu L P, Zhu L X. Dimension reduction for conditional variance in regressions. *Statist Sinica*, 2009, 19: 869–883
- 108 Zou H. The adaptive lasso and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429

Some recent advances in sufficient dimension reduction

Wei Luo

Abstract Sufficient dimension reduction (SDR) is a main stream of the dimension reduction technique in modern statistical research. It aims at replacing the original predictor X in the data with some reduced predictor $\beta^T X$ without loss of information for modeling the response Y . In this paper, we briefly review the literature of SDR, including its basic theory and major types of methods such as the inverse regression methods, the minimum average variance estimation, and the semi-parametric methods based on estimating equations. We also review some recent advances including the sparse and non-sparse SDR methods under high-dimensional settings, the application of SDR in causal inference, and nonlinear SDR.

Keywords sufficient dimension reduction, causal inference, supervised learning

MSC(2020) 62B05

doi: 10.1360/SSM-2024-0034