学校代码： 11482

学　　号： 201510011013

浙江财经大学

ZHEJIANG UNIVERSITY OF FINANCE AND ECONOMICS

硕士学位论文

MASTER　THESIS

论文题目 高维协变量下的多处理因果效应分析

作者姓名 饶诗诗

专　　业 应用统计

所在学院 数据科学学院

指导教师 赵晓兵

完成日期 2023 年 1 月

硕士学位论文

高维协变量下的多处理因果效应分析

2023 年 1 月

MASTER THESIS

An analysis of multi-treatment causal effects under high-dimensional covariates

January 2023

摘要

新生儿作为生命的起源，其出生数量的下降趋势，一般会对社会经济、人口老龄化等问题造成影响。目前人口老龄化程度加重，国家出台三胎政策来化解此问题，但是面对三胎下的母亲怀孕年龄逐渐增大的趋势，使得对母亲怀孕年龄的探讨又一次成为了研究的热点，因为产妇的高龄一般被认为会对胎儿的发育造成影响，故研究母亲怀孕年龄对婴儿体重的影响是很有价值的，在一定程度上可以避免母亲过早怀孕或是过晚怀孕，使得婴儿尽可能地避免早产、畸形、夭折等情况。然而，样本量小和高维数据势必会对分析生存数据带来挑战，如何对高维数据进行分析，是现有的降维方法所面临的一大难点。

本文基于美国宾夕法尼亚州相关人物的生存数据，选取“母亲怀孕年龄”为处理变量，选取“婴儿出生体重”为响应变量，研究母亲怀孕年龄对婴儿出生体重的影响，此时，处理变量和响应变量均为连续变量，二维的处理变量对二维的响应变量的传统因果效应分析不再适用。基于此，本文在 Imbens 定义的广义倾向得分的基础上定义了连续型处理变量下的广义倾向得分函数，进而为下面的因果分析奠定基础。本文研究内容可分为以下几个步骤：

首先，为了解决高维问题，本文采用切片逆回归降维方法将高维的协变量降维成低维协变量。采用切片逆回归方法较为完整的解决了高维数据下变量选择问题，为后续基于局部似然方法估计倾向得分函数奠定了基础。切片能够将连续型处理变量变成离散型处理变量，从而化解使用积分来估计平均因果效应会造成计算复杂的问题。切片之后的处理变量可能不再是二维处理变量，可能是三维甚至是三维以上的处理变量，需要结合多处理治疗方案模型进行后续分析。

其次，查阅相关文献发现，连续型处理变量下的因果效应大都是假设广义倾向得分函数服从正态分布。然而，在实际情况中正态分布的假设性太强，且对年龄的分析一般是基于指数模型和 Cox 模型进行研究的，故本文基于指数模型和 Cox 模型采用局部似然方法估计广义倾向得分函数，并进行了相关模拟。

最后，基于美国宾夕法尼亚州相关数据对广义倾向得分函数进行逆概率加权来研究母亲的年龄对新生儿体重的因果效应，得出了不同年龄段对新生婴儿体重的因果效应不同：母亲怀孕年龄在 13-22 岁时的因果效应较大，即对婴儿出生体重的影响较大，故不建议女在此年龄段下怀孕；母亲怀孕年龄在 22-26 岁时的因果效应与母亲怀孕年龄在 26-27、27-30、30-35 岁时的因果效应相差较小，即母亲在 22-26 岁和 26-27、27-30、30-35 岁下怀孕对婴儿体重的影响的程度相差不大，且有着较好的身体机能，故在 22-35 岁这个年龄段怀孕为最佳年龄段；35 岁

之后随着年龄的增长母亲怀孕年龄对婴儿体重的影响效果逐渐增大，其原因可能是35岁之后母亲的身体机能开始走下坡路，此时怀孕母体对胎儿的营养供给可能会存在一些不足，故不建议女性在35岁之后怀孕。

综上所述，本文利用充分降维的方法将因果效应应用在多处理变量以及连续型响应变量下，从而拓展了经典因果效应。本文基于非参数方法对广义倾向得分函数进行估计，从而有效的解决因果推断中参数估计的问题，使得因果推断的应用更广泛。实例分析所得的结果是：建议母亲在22-35岁这个年龄段怀孕，相对来说母亲在该年龄段怀孕对婴儿出生体重的影响较小、较稳定。母亲在这个年龄段的身体机能相对稳定，能够为胎儿提供比较充足的营养物质，在一定程度上可以避免母亲过早或过晚怀孕，使婴儿尽可能地避免出现早产、畸形、夭折等情况，从而避免体重过轻或体重过重问题。

关键字：局部似然；切片逆回归；多处理治疗方案；因果推断；逆概率加权

ABSTRACT

As the origin of life, the downward trend in the number of births of newborns generally affects social economy, population aging and other issues. At present, the degree of population aging is increasing, the state has introduced a three-child policy to solve this problem, but in the face of the gradual increase in the mother's age of pregnancy under the third child, the discussion of the mother's pregnancy age has once again become a research hotspot, because the advanced age of the mother is generally considered to have an impact on the development of the fetus, so it is very valuable to study the impact of the mother's gestational age on the weight of the baby, to a certain extent, it can avoid the mother's early pregnancy or too late pregnancy, so that the baby can avoid premature birth, malformation, premature death and other situations as much as possible. However, the small sample size and high-dimensional data are bound to bring challenges to the analysis of survival data, and how to analyze high-dimensional data is a major difficulty faced by existing dimensionality reduction methods.

Based on the survival data of relevant people in Pennsylvania, USA, this paper selects "mother's age of pregnancy" as the treatment variable and "infant birth weight" as the response variable to study the effect of the mother's gestational age on the baby's birth weight. Based on this, this paper defines the generalized propensity score function under continuous processing variables on the basis of the generalized propensity score defined by Imbens, and then lays the foundation for the following causal analysis. The research content of this article can be divided into the following steps:

Firstly, in order to solve the high-dimensional problem, this paper uses the slice inverse regression dimensionality reduction method to reduce the high-dimensional covariate to a low-dimensional covariate. The slice inverse regression method is used to solve the problem of variable selection under high-dimensional data in a complete way, which lays a foundation for the subsequent estimation of the propensity score function based on the local likelihood method. Slicing can turn continuous processing variables into discrete processing variables, thereby resolving the computational complexity of using integrals to estimate mean causal effects. The processing variables after slicing may no longer be two-dimensional processing variables, but may be three-dimensional or even more than three-dimensional processing variables, which need to be combined

with multi-processing treatment plan models for subsequent analysis.

Secondly, reviewing the relevant literature, it is found that the causal effect under continuous treatment variables is mostly assumed that the generalized propensity score function follows a normal distribution. However, the assumption of the normal distribution is too strong in the actual situation, and the analysis of age is generally based on the exponential model and the Cox model, so this paper uses the local likelihood method to estimate the generalized propensity score function based on the exponential model and the Cox model, and carries out relevant simulations.

Finally, based on the relevant data of Pennsylvania, the causal effect of the mother's age on the weight of the newborn was studied by inversely weighting the generalized propensity score function, and it was concluded that the causal effect of the mother's age on the weight of the newborn baby was different at different ages: the causal effect of the mother's pregnancy age at the age of 13-22 years was greater, that is, the impact on the birth weight of the baby was greater, so it was not recommended for women to become pregnant at this age; The causal effect of the mother's pregnancy age at the age of 22-26 is less different from the causal effect of the mother's pregnancy age at the age of 26-27, 27-30, 30-35 years old, that is, the mother's pregnancy at the age of 22-26 and 26-27, 27-30, 30-35 years old has little difference in the degree of effect on the baby's weight, and has better physical function, so pregnancy at the age of 22-35 is the best age group; After the age of 35, with the increase of age, the effect of the mother's age on the weight of the baby gradually increases, the reason may be that the mother's physical function begins to decline after the age of 35, at this time the pregnant mother's nutritional supply to the fetus may have some deficiencies, so it is not recommended for women to become pregnant after the age of 35.

In summary, this paper uses the method of sufficient dimensionality reduction to apply the causal effect to multi-processed variables and continuous response variables, thereby expanding the classical causal effect. This paper estimates the generalized propensity score function based on nonparametric methods, so as to effectively solve the problem of parameter estimation in causal inference, and make the application of causal inference more extensive. The results of the case studies show that mothers are recommended to become pregnant in the age group of 22-35 years, and that maternal pregnancy at this age group has a relatively small and stable impact on the birth weight of the baby. The mother's physical function at this age is relatively stable, can provide the fetus with more sufficient nutrients, to a certain extent can avoid the mother

premature or too late pregnancy, so that the baby as much as possible to avoid premature birth, deformity, premature death, etc., so as to avoid underweight or overweight problems.

Keywords： Local likelihood; Sliced inverse regression; Multi-treatment regimens; Causal inference; Inverse probability weighting

目　录

第一章 绪论

第一节 研究背景

在分析和掌握有关生存统计时，统计学家们总是为统计中的因果关系问题而深感困惑，比如怎样确定同一种药品对特定病症的治疗的有效程度、雇用的劳动记录如何证实雇主具有性别歧视等等。因果关系推断理论广泛应用于流行病学、社会学、计量经济学等研究领域中，以现代医学为例，由于有关学科的进展以及医学模式的改变，人类对疾病成因的了解程度也在日益增加。十九世纪，单病因学说产生，随后更多病因的概念也慢慢产生。虽然有较多书籍，如 Pearl(2000)都提到相关概念与因果并不等价。但是，在流行病学、神经生物学以及其他专业领域中，仍有许多学者将因果和相关混淆。因此，正确理解因果关系尤为重要，这对特定现象的诊断、治疗和预防有重要意义。

因果推断的模型大约在 20 世纪 20 年代开始逐渐被完善。Rubin (1978)的潜在结果模型(Rubin Causal Mode, RCM)和 Pearl (1995)的因果图模型(Causal Diagram, CD)为目前两个代表性模型。其中，因果图模型主要用于表示有向无环图因果关系的点路径关系，但是图模型主要用于小规模因果网络。当数据中含有高维混杂因素的时候，传统因果推断方法的有效性将大大降低，充分降维可以在不损失信息的情况下解决这类问题，再基于潜在结果模型来估计降维数据后的因果效应，使得因果推断的应用更广泛。

因果推断在流行病学、临床医学上常常面临着多处理的问题，例如，在研究不同药物治疗下的患病者康复情况，此时的药物治疗可能不止两种，可能含有三种及以上的多种治疗方案，这时情况将变得复杂起来。一项经典的因果推断研究内容是母亲抽烟与否(处理变量二元变量)对婴儿体重的影响，例如 Almond et al. (2005)对低出生体重的成本进行了分析，其中低出生体重数据来自于美国宾夕法尼亚州相关人物的观察结果，其中包含了 4642 名婴儿的出生信息(Cattaneo, 2010)，数据下载地址及相关信息见表 20。该数据集中共有 23 个变量，响应变量是以克数为计量单位的“婴儿出生体重”，二元处理变量为“母亲的吸烟情况”(1 =吸烟，0=不吸烟)。协变量包括“母亲怀孕的年龄”、“母亲的婚姻状况”(是否结婚)、“母亲怀孕期间饮酒的指标变量”(饮酒与否以及饮酒的次数)、“母亲怀孕之前是否有出生新生儿死亡”、“母亲的教育”、“父亲的教育”、“产前护理的数量”、“母亲的种族”、“父亲的种族”以及“自上个孩子出生到现在的月份间隔数”(月)等其他变量。其次，Liu et al. (2018)基于该数据提出了一种半参数方法来估计“母亲抽烟与否”对“婴儿出生体重”的平均因果效应，他指出如果参数估计倾向得

分模型指定不正确，平均治疗效应的最终估计结果也将不一致。此外，半参数估计不依赖于结果回归模型，在难以获得或计算可靠的结果回归模型时(例如，研究复杂疾病的治疗效果时)具有强烈的吸引力。同时针对高维的协变量，Liu et al. (2018)直接采用几何技术(Geometric Technique)的方法对高维数据进行降维，再使用半参数方法来估计倾向得分函数，最后基于倾向得分函数构建双稳健模型来评价“母亲抽烟与否”对“婴儿出生体重”的平均因果效应。在该数据中，我们注意到，“母亲怀孕年龄”也是一个重要变量(例如，高龄产妇的对新生婴儿的健康有重大影响)。基于此，本文使用原始数据(Cattaneo, 2010)，设置“母亲怀孕年龄”为处理变量，研究处理变量“母亲怀孕年龄”对响应变量“婴儿出生体重”的因果效应分析，其他变量为协变量。此时，协变量共有 21 个，属于高维数据，需要进行降维，以便实现数据可视化降低计算的难度。然而，采用传统的降维方法进行处理会遇到“维数灾难”问题，充分降维刚好能避免此问题。

目前，充分降维方法主要包括：逆回归方法、最小平均方差估计(Minimum Average Variance Estimation, MAVE)、半参数估计。切片逆回归为其中经典的方法，它通过研究多维自变量和因变量之间的关系，在对自变量的降维过程中充分利用了因变量的信息，能有效的降维。MAVE 相对于逆回归方法来说计算是较复杂的，因为 MAVE 需要多次迭代，计算相对复杂；半参数估计是介于参数估计和非参数估计之间的一种方法，它需要平衡参数和非参数，且半参数估计是一种有偏估计。考虑到本文所使用的数据中处理变量为连续型变量，故采用经典切片逆回归(Sliced Inverse Regression, SIR)进行降维处理，而切片后的处理变量可能不再是二元变量，例如，当切片数为 3 或者 3 以上时，处理变量便是一个离散型的多处理变量，此时研究变量“母亲怀孕年龄”对变量“婴儿出生体重”的影响就应该使用多处理治疗方案模型。多处理治疗方案模型是指处理变量不再是 0-1 变量，可能是三种及以上。针对多处理问题，不同的学者对此做了不同的分析和处理：Imbens (2000)基于倾向得分(Propensity Score, PS)引入了多种治疗的观察性研究的广义倾向得分(Generalized Propensity Score, GPS)，提出了一种基于 GPS 的加权估计和回归调整估计方法用于估计观察性研究中涉及多处理的平均因果效应；Feng et al. (2012)基于 Imbens (2000)的 GPS 估计进行了修正，提出了基于 GPS 估计平均因果效应的 PS 加权估计，并采用协方差分析(ANCOVA)对 PS 加权估计和回归调整估计进行了比较，结果显示 PS 加权估计在偏差(Bias)和均方误差(MSE)两方面方面较优；Tu et al. (2013)对广义双稳健估计(Generalized Doubly Robust, GDR)进行了改进，将 GDR 估计与 Imbens 加权估计和 PS 估计进行比较，结果表明，当响应变量为正态分布时，GDR 估计比 Imbens 加权估计和 PS 估计性能好；Tu and Koh (2016)基于 GPS、Imbens 加权估计、PS 估计以及 GDR 估计，考虑了当

响应变量为非正态分布时，进行了多个治疗的因果效应分析，并基于偏差(Bias)和均方误差(MSE)进行性能评估；楼芝兰(2018)采用了在多种治疗方案并存的情况下，通过引入再生核希尔伯特空间(RKHS)来寻找最优分配估计平均因果效应；Austin (2019)描述了定量或连续暴露的广义倾向得分来评估协变量平衡，比较了两种基于广义倾向得分的估计方法：普通最小二乘广义倾向得分和协变量平衡广义倾向得分方法；Garès et al. (2022)提出了连续结果变量下的治疗效果的方差估计，基于广义倾向得分对线性剂量反应函数(Dose-Response Functions, DRFs)进行加权和分层估计的方差估计，最后评价母亲体重指数对婴儿出生体重的影响。

对多处理问题进行归纳总结，可以看出大多数的多处理问题均是基于 GPS 来估计的平均因果效应，GDR 也是建立在 GPS 上的估计，故对多处理的平均因果效应可以基于 GPS 进行估计，考虑到采用参数方法估计 GPS 会有一定的局限性，故采用非参数方法对 GPS 进行估计，再利用估计出来的 GPS 进行逆概率加权来评价母亲怀孕对婴儿出生体重的平均因果效应，依据所得的结果提出相关的建议，能有效防止婴儿出现早产、畸形、夭折等问题。

第二节 文献综述

一、因果推断的现状及发展

在大数据背景下，因果关系推断的发展紧跟时代，人们发现相关关系已经无法满足自己对事物进行深刻理解的需求，数据间的因果关系被大量挖掘出来，能够为各个行业的发展提供基础，甚至引导出新的发展方向。

因果关系最早由哲学家频繁提起，如哲学家 Aristotle 在《Physics》中提出了“因”；在 1690 年英国哲学家 John Locke 定义了“果”。皮尔逊(Pearson, 1911)提出了一种因果分析方法，即列联表法。他指出描述总是被简化为一个列联表，从中可以理解因果关系的关联性。随着时间发展，人们发现关联关系不能满足自己的需求，需进行因果分析，由此推开因果推断的研究大门，不少学者对因果分析进行了研究。Wright (1960)自 1921 年以来一直在修改所提出路径分析模型(Path Analysis)。Neyman (1923)提出了潜在结果(Potential outcomes)概念。Fisher (1936)提出了随机化(Randomization)检验方法。但随机化难以达到，且难以控制偏差，于是 Rubin (1976)提出了控制偏差的匹配抽样理论。Rosebaum and Rubin (1983)提出了倾向得分匹配方法(Propensity Score Matching, PSM)。Pearl (2000)为 RCM 和 CD 提供了等价性证明。这些研究为因果推断提供了理论基础。

随着信息化高速发展，再一次将高维数据下的因果推断助推为热门话题，大量学者进行了研究。为了降低高维因果误差的检测率，Geng et al. (2005, 2006)提

出了一种 D 分离的分解方法，该方法是将图分解为两个子图，再基于子图进行因果推断，最后将子图整合嵌入到整个因果网络中。Tsamardinos et al. (2006)提出了最大——最小爬山法(Max-MinHill Climbing, MMHC)，该方法适应于高维数据的因果推断。Janzing et al. (2012, 2015)提出了信息几何模型的算法 (Information-Geometric Causal Inference, IGCI) ，该方法适合低噪声或者无噪声情况下的因果推断。Cai et al. (2013)提出了 SADA 框架(Scalablec Ausation Discovery Algorithm, SADA)，该方法能够在高维度低样本量的情况下确定因果变量。张浩等(2015)基于互信息提出了高维数据下的因果发现算法(Causal discovery on high dimensional data, CDHD)，该算法通过遍历所有父子节点，构建完整的因果网络图。Hu et al. (2018)对混合数据采用 ANM-MM 算法。Kurthen and Enßlin (2019)在高噪声设置、强离散化数据和非常稀疏的数据情况下提出了一种新的推理方法：贝叶斯因果推理(Bayesian Causal Inference, BCI)。Cheng et al. (2022)利用充分降维来估计平均因果效应，在此基础上提出了一种数据驱动算法来估计平均因果效应，即采用监督内核降维方法学习低维表示从原始协变量空间，然后利用最近邻匹配减少协变量空间估计反事实结果，以避免协变量集过大的问题。

针对目前在高维数据下识别因果关系的问题，现有的方法一般都是将因果推断与机器学习、计算机等方面相结合，最新的论文 Cheng et al. (2022)利用充分降维来估计平均因果效应，将充分降维和因果推断相结合起来发挥两者的优势，故本文将采用充分降维方法对因果效应进行分析。

二、切片逆回归降维方法研究现状

近年来，由于计算机硬件、数据收集方式以及数据保存方式的飞速发展，各行各业都形成了自身的数据库，从而导致了高质量数据的增多。高维数据不同于低维数据，处理低维数据的方法不可直接运用到高维数据处理上，否则会得到意想不到的结果，因此需要先将高维数据进行降维数据处理，对经过处理后的低维数据进行重新分类能更有效地简化运算。

针对高维数据降维问题，各个学者对此进行了研究：Tian (2009)对高维数据分类问题的降维方法进行研究，提出了模拟退火法(SA)和多元适应性随机搜索方法 (MASS) 两种方法，以及用于处理函数型数据的(Functional Adaptive Classification Method, FAC)方法。Clark et al. (2009)介绍了主成分分析、因子分析等降维方法。Rajaraman and Ullman (2011)对特征向量、特征值和奇异值等进行降维。李航(2012)对监督学习方法进行详细的讲述，如感知机、K 近邻、期望最大化(Expectation-Maximum, EM)算法等等。

随着生物医学的发展，人们开始认识到生存的重要性，所以研究对人类生存

有影响的因素是一个热门话题。现有统计方法的最大挑战之一是：在海量的数据中寻找做决策时的有用信息。为了解决此问题，对数据进行降维的方法已成为人们关注的焦点。对数据进行降维处理的优点在于：随着维度的减少，适用于低维数据的统计方法可以成功地应用于降维后的数据；其次，人的视觉受到多维空间的限制，低维数据易于查看。然而，在降维的方法选择上又存在一系列问题：结构复杂多变的高维数据不能用传统的方法，否则可能会存在“维数灾难”问题，而充分降维化解此问题。

SIR 为充分降维的算法中经典的算法，它通过研究多维自变量和单变量之间的关系，在对自变量的降维过程中充分利用了因变量的信息。Li (1991)证明了 SIR 具有良好的渐近性质，且较稳健，因而受了广泛的关注。此后，充分降维的领域得到了极大的发展，一系列基于 SIR 的方法应运而生。

部分专家展开了理论研究：Cook and Weisberg (1991)提出了切片平均方差估计方法(Sliced Average Variance Estimates, SAVE)，相比 SIR，该方法方法所找到的降维空间更加全面。Li (1992)提出了海森主方向(Principal Hessian Directions, PHD)，该方法是基于 Stein 引理的一种新的充分降维方法，且对非线性方向敏感。Hsing and Carroll (1992)认为要使切片方法收敛，则切片数应在$\sqrt{n}$至$n/2$范围内。Li and Wang (2007)提出了方向回归(Directional Regression，DR)，该方法是 SIR 和 SAVE 的组合。Hino et al. (2013) 基于条件熵最小化提出了一种放松或消除线性假设条件降维的方法。Li et al. (2020)用最小的线性组合集来替换原始预测器，以达到降低预测器维数的目的，他们为了解决估计的线性组合通常由所有的变量组成这一困难，提出了稀疏充分降维的方法，在充分降维的框架内进行无模型变量的选择或筛选，并对其进行了进一步的研究。Dong (2021)通过三种优化方法，研究了线性充分降维：首先他通过对一般损失函数的最小化，将经典方法(普通最小二乘和 SIR)以及现代方法(主支持向量机和主分位数回归)放在一个统一的框架下；然后他回顾了通过最大化依赖度量的充分降维方法，其中包括距离协方差、希尔伯特-施密特独立准则、鞅差分散度和期望条件差分；最后提供了充分降维方法的信息准则。

一些学者做了应用上的创新：例如，Chen et al. (1999)将 SIR 应用于缺失值的响应变量中。Li et al. (2003)将 SIR 推广到多维响应变量情形里面。Li et al. (2007)指出当样本量小于维数时，可以应用偏最小二乘与 SIR 相结合的方法进行拓展。Li and Nachtsheim (2006)提出了稀疏切片逆回归方法(Sparse Sliced Inverse Regression, SSIR)，该方法是基于 LASSO 和 SIR 的方法。Park et al. (2009, 2010)定义并应用了时序数据求解中心降维子空间。李向杰等(2018)提出了累积加权切片逆回归法，避免了对切片数的选择，在自变量存在异常值和小样本下比较稳

健。刘金灵(2021)将 SIR 和 SAVE 应用到多响应变量中，丰富了传统的单响应充分降维方法。

此外，一些学者也推广了 SIR 的应用：周文琴等(2001)将 SIR 应用于新型电池的发展中。黄薇等(2004)提出了针对多因素非线性时序数据的方法，将 SIR 和神经网络相结合。李岩岩(2016)将 SIR 方法运用在粮食产量、中国商品房价格中。谢志超(2018)将 SIR 应用于乳腺癌数据中进行实证分析。Cai et al. (2020)提出了一种在线学习的方式来实现 SIR，该方式包括两个步骤：第一步构造核矩阵的在线估计；第二步提出两种在线算法：一种采用摄动法，另一种采用梯度下降优化，进行在线奇异值分解；最后建立了该在线学习的理论性质，并通过模拟和实证分析来证明了该理论性质。

通过上述相关文献不难发现，关于 SIR 的方法研究是比较成熟的，但其应用还是很有限的，因此，本文将 SIR 应用于生存数据，研究其他协变量对母亲怀孕年龄的影响。

由上述文献综述可知，现存大部分文献将因果推断和机器学习等算法相结合来研究因果，少部分人将高维数据和连续型处理变量结合进行因果效应分析。故本文将基于原始数据(Cattaneo, 2010)，采用“母亲怀孕年龄”为处理变量，研究处理变量“母亲怀孕年龄”对响应变量“婴儿出生体重”的因果效应，其他变量为协变量。此时，协变量共有 21 个，属于高维数据，需要对协变量进行降维，将采用 SIR 方法对多个协变量进行降维，SIR 降维后的处理变量可能不是二元处理变量，可能是三元甚至是三元以上的处理变量，这时候变采用多处理模型对因果效应进行分析；然后，将降维后的数据进行非参数方法估计广义倾向得分函数；最后,再基于倾向得分函数进行逆概率加权来估计因果效应，从而比较满意地解决高维数据下的多处理因果效应估计问题。

第三节 文章主要研究内容

本文主要研究多处理下的因果效应分析，考虑的模型是潜在结果模型(Rubin，1978)下含有高维协变量的多处理因果效应分析。在传统的潜在结果模型中，针对的处理变量绝大数是 0-1 两个变量，鲜有文献是针对三个及三个以上的处理变量来做因果效应分析，直至 Imbens (2000)提出了广义倾向得分(Generalized Propensity Score, GPS)，多处理下的因果效应才有了发展，在此基础上，有了对连续型处理变量对二维响应变量的因果效应分析，但是针对连续型处理对连续型响应变量的因果效应分析的文献少之又少，这区别于传统的因果效应分析(针对二维处理变量和二维响应变量)，传统的因果效应分析将不再适用于连续型处理变量以及连续型响应变量，需要重新定义连续型处理变量下的广义倾向得分函数。将

连续型处理变量和高维协变量相结合的文献也较少，本文针对目前研究中的局限性，将主要考虑在连续型处理变量和高维协变量的基础上，使用 SIR 进行切片将之变为多种处理变量和低维协变量下的因果效应。并基于不同分布的协变量进行模拟，进而将此问题推广到其他不同分布下的多处理以及高维协变量的情况中，达到应用广泛的目的。本项研究同时考虑了 SIR 降维算法、非参数方法以及基于 GPS 进行逆概率加权来估计平均因果效应，并将所采用的方法应用在实际例子中，由此来估计母亲怀孕年龄对婴儿体重的影响，基于研究的结果做出相关建议，从而能够达到有效地防止女性怀孕过早或过晚的目的。

为了更好的展示本文结构，将作出具体安排如下：

第一章：本章首先主要介绍了本文所研究的多处理因果效应的研究背景，解释了为何要研究多处理下的因果效应以及如何要研究多处理下的因果效应，再指出本文研究内容的意义。其次介绍了因果模型模型、降维方法以及多处理治疗方案或者连续型处理变量下的因果效应估计的相关文献综述。最后梳理了本文的研究内容、理论意义与使用价值以及介绍了本文的创新与不足。

第二章：介绍因果推断、充分降维以及局部似然方法。本章首先介绍了因果推断模型相关知识；其次介绍了降维算法中 SIR 和结构维数估计问题；最后介绍了非参数方法中的局部似然法，给出了窗宽的一般公式及普通证明。

第三章：主要介绍了高维协变量下的多处理因果效应模型的构建以及模型估计问题。高维协变量下的多处理因果效应模型的构建主要是基于 Imbens (2000)定义的广义倾向得分函数定义了连续型处理变量下的广义倾向得分函数，模型的估计主要是基于指数模型和 Cox 模型，首先是采用 SIR 算法进行充分降维；其次利用非参数方法估计广义倾向得分函数；最后基于逆概率加权方法进行多处理因果效应估计。

第四章：数值模拟部分。针对不同分布、不同模型以及不同的结构维数生成相关数据，首先采用 SIR 算法将高维的协变量降成所需要的结构维数，得出相关的降维空间并评价降维效应；其次利用非参数方法估计广义倾向得分函数；最后基于逆概率加权方法进行多处理因果效应估计。

第五章：实证分析，采用模型估计中所使用的相关方法来分析母亲怀孕年龄对婴儿出生体重的因果效应。在分析不同年龄段的女性怀孕年龄对婴儿出生体重的影响时，协变量的选择主要利用后门准则(Pearl et al., 2020)，年龄段的划分主要利用变量“母亲怀孕年龄”的四分位数，再根据实际情况来随机选择年龄点划分，采用上述的方法来估计“母亲怀孕年龄”对“婴儿出生体重”的因果效应，并据此提出相关建议。

第六章：概述研究内容与不足，以及后续可以研究的方向。

为了更清楚地了解本文的方法流程，本节下面附了一张技术路线图(图 1)，该技术路线图涵盖了本文的主要工作和方法。

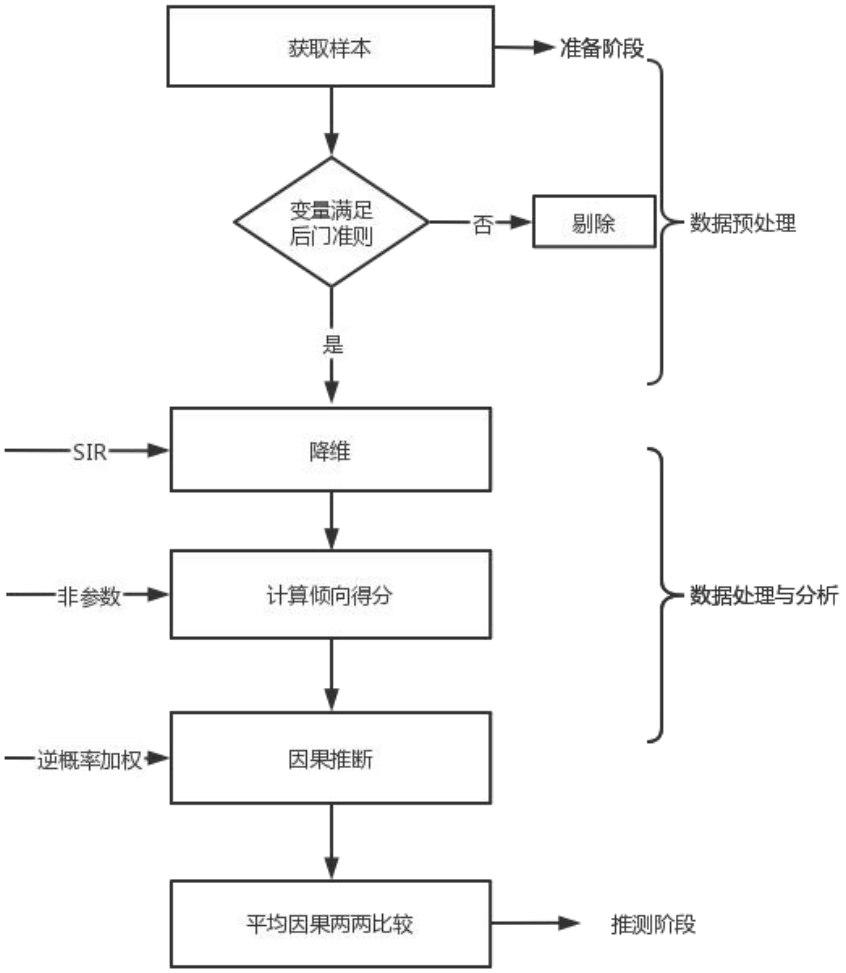

图 1 技术路线图

第四节 理论意义与实用价值

理论意义一：现有文献中，对平均因果效应估计大都基于二分类变量的处理，但是实际生活中更多的多处理平均因果效应，如多种药物治疗下的治疗效果，故本文对多处理情形下的因果效应进行估计。

理论意义二：本文不仅考虑了平均因果效应估计中的高维情况，还考虑了多处理情况，从而较完整的解决了高维数据下多处理因果效应估计问题。

理论意义三：现有文献关于连续型处理变量对连续型响应变量的平均因果效应估计问题研究偏少，本文的研究补充了国内外学术界在该领域的研究。

实用价值：随着计算机等技术的普及，我们所获得的数据越来越多，数据集越来越大。在这些高维数据中，隐藏着重要的信息，如何从高维数据中提取有效

信息成为了如今热点话题。本文对于高维协变量下的多处理因果效应分析的方法，可以应用于其他高维数据分析的领域中，对解决高维问题有很大帮助。

第五节 创新点

本文的创新点：

(1)本文在生成数据模拟部分，考虑了不同分布和不同模型下的处理变量以及协变量，并考虑了不同的结构维数对后续因果效应分析的影响，拓展了因果效应的应用。

(2)本文研究的模型为连续型处理变量对连续型响应变量的平均因果效应模型，在该模型的基础上对连续型处理变量进行切片，将其变成离散型处理变量，从而简化计算。

(3)本文不仅考虑了连续型处理变量的问题，还考虑到了高维协变量的问题，利用协变量和处理变量之间的关系进行充分降维处理，较为完整的解决了高维数据下变量选择问题，为后续的因果效应分析奠定了基础。

(4)本文基于非参数方法对广义倾向得分函数进行估计，将非参数方法应用于平均因果效应分析，从而有效的解决因果推断中参数估计的问题，使得因果推断的应用更广泛。

(5)本文利用充分降维的方法将因果效应应用在多处理变量以及连续型响应变量下，区别于传统的因果效应分析：针对二维处理变量和二维响应变量，从而拓展了经典的因果效应。

第二章 模型介绍

第一节 因果推断

一、因果效应

现阶段的统计问题主要有两个：一，相关关系，量化自变量和相应变量之间的数理关系；二，因果关系，了解对响应变量有影响的因子。100 多年前，基于父母身高和子女身高的关系提出了线性回归模型。回归模型常用于描述变量之间关系，我们常说，相关关系与因果关系不等价，相关学者定义了因果推断的含义：根据某种处理下个体处理结果得出关于因果结论的一个过程。因果推断有以下几方面要素：

(1)个体(Unit)：人、动物或地点，可以在上面进行不同处理的对象。

(2)处理(Treatment)：研究者希望评估进行或不进行某种处理的作用不同；处理必须是可以控制的(注：此处只考虑了处理为两种情况)。

(3)协变量(Confounding Variable)：能够影响个体的处理变量对结果变量产生虚假的相关的变量，如年龄混淆了年收入和患癌概率之间的关系，年龄就是一个协变量。一般来说在此情况下是需要去除影响因果关系的协变量，否则会引起结果变量存在混淆偏倚，产生悖论(注：此处不考虑协变量为对撞的情况)。

(4)潜在结果(Potential Outcome)：每个个体接受或不接受处理时的结果值，该结果通常是反事实的。例如，对于每个特定的个体来说，吃药或不吃药时的结果我们只能观察到一个。

(5)个体因果效应(Individual Causal Effect, ICE)：对于每个个体，接收处理时的潜在结果和不接受处理时的潜在结果的比较。

令 T_i 表示个体 i 接受处理的情况，接受处理为 1，对照组为 0 (此处处理为二值变量，多值可做相应的推广)。Y_i 表示个体 i 的处理的结果，$Y_i^{(T_i)}$ 表示在某种处理 T_i 下的潜在结果。每个个体 i 有一种潜在结果 $\left\{Y_i^{(1)},\ Y_i^{(0)}\right\}$，观察到的结果表示为：$Y_i=T_iY_i^{(1)}+\left(1-T_i\right)Y_i^{(0)}$。ICE 表示个体潜在结果的差或商，例如，

$$Y_i^{(1)}-Y_i^{(0)}\text{或}\frac{Y_i^{(1)}}{Y_i^{(0)}}.$$

然而，对同一个个体观察到的潜在结果只能是一个，例如，我们观测到的数据用表 1 表示，其中 Y=1表示观测到病人的死亡，否则 Y=0；T=1表示个体 i 接受某种处理，否则 T=0。用 $Y^{(0)}$=1 表示个体未接受处理而发生了死亡，$Y^{(1)}$=1 表示个体接受处理时发生了死亡。表 2 为我们想得到理想数据，而实际观测到的数据用表 3 表示，其中有一半数据是缺失的，用“?”表示。

表 1 因果推断的观测数据

i	1	2	3	4	5	6
T	0	0	0	1	1	1
Y	0	1	0	1	1	0

对于某一个个体，如果 $Y^{(0)} \neq Y^{(1)}$，那么在该处理下就存在因果效应。然而，每个个体只能选择接受处理或不处理中的一种，$\left\{Y_i^{(0)}, Y_i^{(1)}\right\}$ 中必然缺失一半，即 ICE 是不可识别的。如果不接受处理，T=0，那么 $Y^{(0)}=Y$，而 $Y^{(1)}$ 是缺失的；同样的，如果接收处理，T=1，那么 $Y^{(1)}=Y$，而 $Y^{(0)}$ 是缺失的。

表 2 因果推断中理想的观测数据(例)

ID	T	Y	$Y^{(0)}$	$Y^{(1)}$
1	0	0	0	0
2	0	1	1	1
3	0	0	0	0
4	1	1	1	1
5	1	1	1	1
6	1	0	0	0

表 3 因果推断中实际的观测数据(例)

ID	T	Y	$Y^{(0)}$	$Y^{(1)}$
1	0	0	0	?
2	0	1	1	?
3	0	0	0	?
4	1	1	?	1
5	1	1	?	1
6	1	0	?	0

在实际观测到的数据中，

$$\Pr\left(Y=1\middle|T=1\right)-\Pr\left(Y=1\middle|T=0\right)=\frac{2}{3}-\frac{1}{3}=\frac{1}{3},$$

而在理想观测的数据中，

$$\Pr\left(Y^{(1)}=1\right)-\Pr\left(Y^{(0)}=1\right)=E\left(Y^{(1)}-Y^{(0)}\right)=0.$$

这会得到两个不同的结论。在观测到的数据中，我们可以说处理 T 跟结果 Y 有相关关系，但实际上两者不存在因果关系，该原因是处理 T 关于结果 Y 还存在其他混淆偏倚(Coufounding Bias)，这就是 Yule-Simpson 悖论。这也说明了相关关系和因果关系不等价，相关性是可以根据数据估计出来的，而因果关系不能。

记(Y_i,T_i,X_i)为我们观测到的数据，其中$i=1,\cdots,n$，假设其独立同分布，其中i表示第i个个体的观测值，Y_i表示观测到的结果，T_i表示接受的处理形式，X_i表示个体在接受处理前的协变量。在一定的假设条件下，因果关系是可估计的。我们关注的重点是利用观测到的数据估计$E\left(Y^{(1)}\right)$和$E\left(Y^{(0)}\right)$，识别总体的平均因果效应。在潜在结果下，为了估计因果效应还需要另一个假设：个体处理稳定假设(Stable Unit Treatment Value Assumption, SUTVA)。即个体之间不会互相被干扰，每个个体只能接受一种处理：

(1)一致性假设：对于个体i，不论处理T的形式如何，不会影响$Y_i^{(T)}$的取值，即：

$$Y_i=T_iY_i^{(1)}+(1-T_i)Y_i^{(0)}.$$

(2)个体间无干扰假设：每个个体的处理结果是独立的，不受其他个体处理的形式所影响。

如果不满足 SUTVA 中的任意一个假设，每个个体的潜在结果不唯一，相对来说是比较复杂的，在这里本文不做考虑。

二、倾向得分(Propensity Score, PS)

统计学家 Rosebaum and Rubin (1983)提出倾向得分概念，该论文指出，平衡得分(Balancing Score)是协变量X_i的一个函数表达式，记为$b(X_i)$。给定$b(X_i)$的条件下，协变量X_i与处理T_i在处理是独立的，即有：

$$X_i \perp T_i \,|\, b(X_i).$$

$b(X_i)$=X$_i$是$b(X_i)$最简单的形式。$b(X_i)$ 是X_i的函数，其形式还存在很多种。Rosebaum and Rubin (1983)将$b(X_i)$的一种形式定义如下：

$$r(X_i)=pr(T_i=1|X_i).$$

称$r(X_i)$为倾向得分，假设

$$pr(T_1,T_2,\cdots,T_n|X_1,X_2,\cdots,X_n)=\prod_{i=1}^{N}r(X_i)^{T_i}\{1-r(X_i)\}^{1-T_i}.$$

故 PS 定义为：给定可观测的协变量下，个体接受处理的条件概率。优点是：将多维协变量X简化为一维的值。在观察研究中，匹配、分层和协方差调整是倾向得分的三种应用。

倾向得分的性质：

(1)平衡处理组与控制组的差异。具有相同或相近倾向得分值的处理组和控制组成员在可观测到的协变量上的差异是平衡的。该性质为研究者带来便利。

(2)给定倾向得分$r(X_i)$条件下，处理T与可观测到的协变量X是独立的，即有：

$$X \perp T \left| r\left(X_i\right). \right.$$

则在控制了倾向得分的情况下，每个个体都具有相同的概率被分到处理组和控制组，能做到类似于随机化要求。

(3)处理分配机制的可忽略性(ignorability of treatment assignment mechanism)(简称可忽略性)假定：对所有的X_i，若有：

$$\left(Y_i^{(1)}, Y_i^{(0)}\right) \perp T_i \left| X_i, \right.$$
$$0 < pr\left(T_i = 1 \left| X_i \right.\right) < 1.$$

成立，则有：

$$\left(Y_i^{(1)}, Y_i^{(0)}\right) \perp T_i \left| b\left(X_i\right), \right.$$
$$0 < pr\left\{T_i = 1 \left| b\left(X_i\right) \right.\right\} < 1.$$

对任意的$b\left(X_i\right)$成立。

(4)在可忽略性假定成立的情况下，有下面的表达式成立：

$$E\left\{Y^{(1)} \left| b(X), T = 1 \right.\right\} - E\left\{Y^{(0)} \left| b(X), T = 0 \right.\right\} = E\left\{Y^{(1)} - Y^{(0)} \left| b(X) \right.\right\}.$$

该性质将反事实框架联系起来，对于倾向得分相同的个体来说，处理组的结果变量Y_1与控制组的结果变量Y_0的均值之差表示为：

$$\begin{aligned} & E\left\{E\left\{Y^{(1)} \left| b(X), T = 1 \right.\right\} - E\left\{Y^{(0)} \left| b(X), T = 0 \right.\right\}\right\} \\ & = E\left\{E\left\{Y^{(1)} - Y^{(0)} \left| b(X) \right.\right\}\right\} \\ & = E\left(Y^{(1)} - Y^{(0)}\right). \end{aligned}$$

三、逆概率加权

Horvitz and Thompson (1952)指出当抽样概率是已知时，抽样人群是从目标人群中抽取的，那么这个概率的倒数被用来加权观测。该方法已经被广泛的应用于统计学的不同的框架中。逆概率加权(Inverse Probability Weighting, IPW)是一种用于解释由于非随机选择观测值或人群信息的非随机缺失而造成的缺失和选择偏差的方法。在因果推断中，IPW 在处理多组间变量混杂偏倚中起到了重要作用。换言之，就是把许多协变量和混杂因素打包成一个概率并进行加权，只需计算它的权重即可，方便了许多，其权重公式用倾向得分的倒数来表示(注：IPW 也满足可忽略性假定)。IPW 可用给定协变量 X 下接受处理 T 的条件概率的倒数来表示，其表达式为：

$$W^T = \frac{1}{r(T \left| X) \right.} = \frac{1}{\Pr\left[T = t \left| X = x \right.\right]}.$$

传统的 IPW 主要用于二维处理变量下的因果推断，则相应的处理$T \in \{0,1\}$，

$T=1$表示接受处理的实验组，$T=0$表示不接受处理的对照组，则实验组和对照组所对应的权重分别为：

$$W^1=\frac{1}{r(T|X)}=\frac{1}{\Pr[T=1|X=x]},$$
$$W^0=\frac{1}{r(T|X)}=\frac{1}{\Pr[T=0|X=x]}.$$

则相应的平均因果效应可以表示为：

$$E\left(Y^{(1)}-Y^{(0)}\right)=E\left(\frac{I(T=1)Y}{r(T=1|X)}\right)-E\left(\frac{I(T=0)Y}{1-r(T=1|X)}\right).$$

Horvitz and Thompson (1952)指出给定可忽略性假定和一定的正则性，逆概率加权估计是平均因果效应的相合估计。

第二节 充分降维

一、充分降维概念

计算机科学的进步导致了数据海量化，而海量数据也使人困扰。怎样才能在尽量没有信息损失的情形下，把高维数据问题转变为常见且好处理的低维数据模型，这一过程就显得特别有意义。而随着对其它专业领域高维数据信息的大量收集，充分降维的研究再一次被统计学界学者所瞩目。

充分降维，是在不假定任意参数模型、不破坏信息的前提下，通过对多维自变量进行线性组合，实现降维的目的。主要思路是：一维响应变量Y关于p维自变量$X=(X_1,\cdots,X_p)^T$的回归问题，响应变量可以是离散或连续的，针对条件分布$Y|X$进行分析，希望找到一个$p\times d$的矩阵$\mathrm{B}(d\le p)$，使得$Y|X$与$Y|\mathrm{B}^TX$相等，与之等价的是去找到满足下式成立的$p\times d$矩阵B：

$$Y\perp X|\mathrm{B}^TX.$$

这里$\perp$表示条件独立。B总是存在的，极端情形就是$d=p,\mathrm{B}=\mathrm{I}_p$，若$d\ll p$，则达到降维的目的了。如果对$\mathrm{B}^T$左乘任意一个非退化的$d\times d$矩阵$\Lambda^T$，仍然满足独立：

$$Y\perp X|\Lambda^T\mathrm{B}^TX.$$

可以看出我们真正想找的是B的列空间Span(B)，将其称为降维空间。当然对任意一个非退化的$p\times p$矩阵A，Span(BA)也是降维空间。Cook (1998)定义了中心降维子空间的概念：所有降维空间的交集仍是一个降维空间，记为$\mathrm{S}_{Y|X}$，并称$\mathrm{S}_{Y|X}$的维数$\mathrm{d}=\dim(\mathrm{S}_{Y|X})$为结构维数。

特别的，如果基于回归均值曲线$\mathrm{E}(Y|X)$所进行的降维，相应的就是找一个

$p\times d$ 的矩阵B，满足：

$$Y \perp \mathrm{E}\left(Y\middle|X\right)\middle|\mathrm{B}^{T}X.$$

称Span(B)为均值降维空间。同样的，中心均值降维子空间也有相同定义：所有均值降维空间的交集仍是一个均值降维空间，记为$\mathrm{S}_{\mathrm{E}(Y|X)}$。

如果，

$$Z=\Lambda X+b,$$

那么，

$$\mathrm{S}_{Y|Z}=\Lambda^{-T}\mathrm{S}_{Y|X}.$$

则称中心降维子空间具有不变性，该性质也适用于中心均值降维子空间。则我们可以将 X 做标准化处理 $Z=\Sigma_{X}^{-1/2}\left(X-\mu\right)$，这里 $\mu=\mathrm{E}\left(X\right)$ 和 $\Sigma_{X}=Cov\left(X\right)$ 分别表示 X 的均值和方差。

二、切片逆回归方法

Li (1991)提出的 SIR (Sliced Inverse Regression)方法被广泛接收和应用，该方法是基于一个线性条件均值假设：

$$\mathrm{E}\left(Z\middle|\mathrm{B}^{T}Z\right)=P_{\mathrm{B}}^{T}Z,$$

其中 P_{B} 是投影阵，$P_{\mathrm{B}}=\mathrm{B}(\mathrm{B}^{T}\mathrm{B})^{-1}\mathrm{B}^{T}$。这一假设在 p 很大时也渐近成立(Hall and Li, 1993)。

基于此，令 $\mathrm{B}_{1}=\Sigma^{1/2}\mathrm{B}$，可以推出：

$$E\left(Z\middle|Y\right)=E\left(\left(Z\middle|Y,\mathrm{B}_{1}^{T}Z\right)\middle|Y\right)=E\left(\left(Z\middle|\mathrm{B}_{1}^{T}Z\right)\middle|Y\right)=P_{\mathrm{B}_{1}}^{T}E\left(Z\middle|Y\right),$$

$$Cov\left[E\left(Z\middle|Y\right)\right]=P_{\mathrm{B}_{1}}^{T}Cov\left[E\left(Z\middle|Y\right)\right]P_{\mathrm{B}_{1}}.$$

故 $\mathrm{Span}\left\{Cov\left[E\left(Z\middle|Y\right)\right]\right\}\subseteq S_{Y|Z}$，即 $Cov\left[E\left(Z\middle|Y\right)\right]$ 的非零特征根所对应的特征向量就是 $S_{Y|Z}$ 的一组基向量。当中心降维子空间的维数为 d 时，表示 $Cov\left[E\left(Z\middle|Y\right)\right]$ 至多有 d 个非零特征根，与之所对应的特征向量包含于中心降维子空间。这样，只需对核矩阵 $Cov\left[E\left(Z\middle|Y\right)\right]$ 进行谱分解，求得其特征值和特征向量即可。参见(Zhu and Zhu, 2007)。

SIR 算法如下：

假设有 n 个独立同分布的样本 $\left(X_{i},Y_{i}\right),\left(i=1,\cdots,n\right)$，

(a)对自变量 X_{i} 标准化，得到 $\hat{Z}_{\mathrm{i}}=\hat{\Sigma}^{-1/2}\left(X_{i}-\bar{X}\right)$，其中 $\hat{\Sigma}$ 和 $\bar{X}$ 分别为样本协方差矩阵和样本均值，并 $Y_{1},\cdots,Y_{n}$ 将中心化 $\hat{Y}_{1},\cdots,\hat{Y}_{n}$。

(b)对数据 $\hat{Y}_{i}$ 进行排序，不妨设 $\hat{Y}_{(1)}\le\hat{Y}_{(2)}\le\cdots\le\hat{Y}_{(n)}$。

(c)尽可能平均的将区间 $\left[\min\left\{\hat{Y}_{1},\cdots,\hat{Y}_{n}\right\},\max\left\{\hat{Y}_{1},\cdots,\hat{Y}_{n}\right\}\right]$ 划分为 H 片，记为

$I_1, I_2, \ldots, I_H$，并计算每一片中 $\hat{Z}$ 的均值，得到 $\hat{\mu}_h = \frac{1}{\hat{p}_h}\sum_{j\in \mathrm{I}_h}\hat{Z}_j$，$\hat{p}_h$ 表示 Y_h 落入片 h 的概率。

(d)建立 SIR 的核矩阵 $\hat{M} = \sum_{h=1}^{H}\frac{\hat{p}_h}{n}\hat{\mu}_h\hat{\mu}_h^T$.

(e)记 $\hat{\eta}_d\left(d=1,2,\ldots,p\right)$表示 d 个特征向量，可以用之来估计 $S_{Y|Z}$.

(f)降维向量：$\hat{\beta}_d = \hat{\Sigma}^{-1/2}\hat{\eta}_d$。而由 $\hat{\beta}_d$ 可以用来估计 $S_{Y|X}$。

三、结构维数的估计

对中心(均值)降维子空间的估计需要确定结构维数 d 之后，再去估计中心(均值)降维子空间。

(一) 序贯检验法

序贯检验法是确定结构维数的方法之一。其思想是：令核矩阵 M 的样本估计 $\hat{M}$ 的特征值为 $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$。那么检验结构维数 H_0: d=m v.s. H_1: d＞m 的检验统计量 T_{m} 有如下形式：

$$T_{\mathrm{m}} = n\sum_{m+1}^{p}\omega\hat{\lambda}_j^k,$$

这里 ω 表示标准化因子，通常与 $\hat{\lambda}_j$ 的渐近方差有关，当 $k=1$ 时，表示核矩阵是半正定的，否则 $k=2$。Li (1991)指出：当 X 服从正态分布时，T_{m} 收敛到标准的卡方分布。序贯检验从 $m=0$ 开始，若接受原假设，则 $m=0$，否则原假设变为 $m=1$，进行再次检验，若拒绝原假设，则 m 继续加 1，直到接受原假设为止，此时的 m 就是结构维数的估计值。易知该方法对维数的估计取决于置信水平，并且在渐进性质中要求结构维数 d 小于切片数减 1，即 $d<H-1$，而实际中结构维数 d 是不知道的。

(二) 准则方法

Zhu et al. (2006)提出了一种 Bayes Information Criterion (BIC)准则方法，该方法是将结构维数选取问题与经典的 BIC 准则相结合。这种方法的优点是保证了所估计的结构维数具有相合性，下面对此做简单的介绍。

令 $\Omega = \mathrm{M} + I_p, \hat{\Omega} = \hat{\mathrm{M}} + I_p$ 为其样本估计，$\left(\lambda_1(\Omega), \cdots, \lambda_p(\Omega)\right)$ 为特征值，$\left(\hat{\lambda}_1(\Omega), \cdots, \hat{\lambda}_p(\Omega)\right)$为 $\hat{\Omega}$ 的特征根，τ 为特征根中大于 1 的个数。定义：

$$logL_d = \frac{n}{2}\sum_{i=1+\min(\tau,d)}^{p}\left(log\left(\hat{\lambda}_i(\Omega)\right)+1-\hat{\lambda}_i(\Omega)\right).$$

然后 BIC 准则$G(d)$定义为：

$$G(d) = \log L_d - C_n d(2p-d+1)/2.$$

其中C_n是一惩罚函数。使得$G(d)$达到最大值的$d=(0,\cdots,p-1)$是结构维数的估计值。该估计值收敛性仅依赖样本特征值的收敛性，相对于序贯检验来说 BIC 准则的应用更广泛，然而目前还未找到最优的C_n。

(三) $R_{d,H}$ 准则

$R_{d,H}$准则主要是一个基于风险函数及其 Bootstrap 估计，该方法可以选择合适的结构维数d和切片数H。该方法首次由 Liquet and Saracco (2012)提出，他们开发了一个交互式的 3D 图形工具，用 R 软件来实现对d和H选择，对应的 R 程序包被命名为“edrGraphicalTools”[1]。

$$R_{d,H} = \mathrm{E}\left[Trace(P_d\hat{P}_{d,H})\right]/d, \quad \forall d = 1,\cdots,p$$

其中，P_d=$\mathrm{B}_d(\mathrm{B}_d^T\Sigma\mathrm{B}_d)^{-1}\mathrm{B}_d^T\Sigma$，$\Sigma$表示$X$的协方差矩阵，Span(B)为降维空间，$\mathrm{B}_d$=$(\beta_1,\cdots,\beta_d)$。$\hat{P}_d$表示$P_d$的估计值，$\hat{\Sigma}$表示$X$的协方差矩阵的估计值。因为$\hat{P}_d$取决于切片数$H$的选择，故用$\hat{P}_{d,H}$表示，其表达式为：$\hat{P}_{d,H}$=$\hat{\mathrm{B}}_{d,H}(\hat{\mathrm{B}}_{d,H}^T\hat{\Sigma}\hat{\mathrm{B}}_{d,H})^{-1}\hat{\mathrm{B}}_{d,H}^T\hat{\Sigma}$，同理$\hat{\mathrm{B}}_{d,H}$也会受到切片数$H$的影响，$\hat{\mathrm{B}}_{d,H}$=$(\hat{\beta}_{1,H},\cdots,\hat{\beta}_{d,H})$。

设Ψ为 Bootstrap 重复的次数，考虑样本$s^{(\varphi)}=\left\{(X_i^{\varphi},Y_i^{\varphi}),i=1,\cdots,n\right\}$，则均方风险函数的朴素 Bootstrap 估计定义为：

$$\hat{R}_{d,H} = \frac{1}{\Psi}\sum_{\varphi=1}^{\Psi}\hat{R}_{d,H}^{(\varphi)}$$

其中$\hat{R}_{d,H}^{(\varphi)}$=$Trace(\hat{P}_{d,H}\hat{P}_{d,H}^{(\varphi)})/d$，$\hat{P}_{d,H}^{(\varphi)}$表示投影在子空间上的第$d$个的特征向量。该式表明$\hat{R}_{d,H}$的值越接近于 1，能够较好的选择$d$和$H$的值，且$d \ll p$，达到降维的目的。本文对于结构维数的估计将采用此方法来估计。

第三节 局部似然

为了介绍局部似然的概念，我们先回顾在参数模型下的最大似然估计方法。假设样本(X_i,Y_i)的对数函数为$l\left\{\eta(X_i),Y_i\right\}$，其中$\eta(\cdot)$是关于$x$的函数。则传统的$n$个样本下的对数似然可以表示为：

[1] 对应的网址为：http://cran.r-project.org.

$$\sum_{i=1}^{n} l\{\eta(X_i), Y_i\}.$$

现在，我们考虑使用非参数估计形式未知的函数$\eta(\cdot)$，假设我们想要估计$\eta(x_0)$，假设函数η在给定点x_0处有$p+1$个连续的导数。数据X_i在给定x_0的领域内，$\eta(X_i)$可以通过p维的泰勒多项式展开近似代替：

$$\eta(X_i) \approx \eta(x_0) + \eta'(x_0)(X_i - x_0) + \cdots + \frac{\eta^{(p)}(x_0)}{p!}(X_i - x_0)^p \equiv \mathbf{X}_i^T \boldsymbol{\beta}^0,$$

其中，$\mathbf{X}_i = \left(1, X_i - x_0, \ldots, (X_i - x_0)^p\right)^{\mathrm{T}}$和$\boldsymbol{\beta}^0 = \left(\beta_0^0, \ldots, \beta_p^0\right)^{\mathrm{T}}$以及$\beta_v^0 = \eta^{(v)}(x_0)/v!$，$v = 0,1,\ldots,p$。对于样本$(X_i, Y_i)$点在$x_0$的领域内所对应的对数似然函数为$l\{\mathbf{X}_i^T\beta, Y_i\}$，对应的权重为$K_h(X_i - x_0)$，其中$K_h(\cdot) = K(\cdot/h)/h$，$\beta = \left(\beta_0, \ldots, \beta_p\right)^{\mathrm{T}}$是该模型的参数。则局部加权对数似然函数为：

$$l(\beta; h, x_0) = \sum_{i=1}^{n} l\left(\mathbf{X}_i^T\beta, Y_i\right) K_h(X_i - x_0). \tag{2.1}$$

最大化式 2.1 中的局部加权对数似然函数，从而获得关于向量β的估计量，即$\hat{\beta} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_p\right)^{\mathrm{T}}$，关于$\eta^{(v)}(x_0)$的估计量可以$\hat{\eta}_v(x_0)$来表示，即：

$$\hat{\eta}_v(x_0) = v!\hat{\beta}_v.$$

其中，$v = 0,1,\ldots,p$。简单的，式 2.1 也被成为局部似然。在式 2.1 中，h表示窗宽，一般的，窗宽的选取可参考 Aerts and Claeskens (1997)。

为了更好的呈现如何选取最优窗宽，我们将考虑以下函数：假设随机变量Y服从的密度函数为$f(\eta(x), y)$，其中表示$\eta(x)$变量x的未知函数。用(X_i, Y_i)表示密度函数所对应的样本，我们将用局部似然去估计$\eta(x_0)$，则对应的对数似然函数为：

$$l = \frac{1}{n}\sum_{i=1}^{n} \log f\left(\eta(X_i), Y_i\right) K_h(X_i - x_0). \tag{2.2}$$

其中x_0为给定的点，估计函数$\eta(\cdot)$的方法主要是对式 2.2 关于$\eta(x_0)$求一阶偏导即可，但是式 2.2 中涉及窗宽的选取问题，窗宽的选取主要是基于式 2.2 的中的渐进偏差和方差，故下面给出一个关于窗宽的选取的思路：对式 2.2 关于$\eta(x_0)$求一阶偏导和二阶偏导，分别表示为：

$$l' = \frac{1}{n}\sum_{i=1}^{n} \left[\log f\left(\eta(X_i), Y_i\right)\right]'_{\eta} K_h(X_i - x_0). \tag{2.3}$$

$$l'' = \frac{1}{n}\sum_{i=1}^{n} \left[\log f\left(\eta(X_i), Y_i\right)\right]''_{\eta} K_h(X_i - x_0). \tag{2.4}$$

将式 2.3 和式 2.4 分别记为：

$$l' = \frac{1}{n}\sum_{i=1}^{n} J\left(\eta\left(X_i\right), Y_i\right) K_h\left(X_i - x_0\right). \tag{2.5}$$

$$l'' = \frac{1}{n}\sum_{i=1}^{n} Q\left(\eta\left(X_i\right), Y_i\right) K_h\left(X_i - x_0\right). \tag{2.6}$$

注，式 2.5 和式 2.6 均是将$\eta\left(x_0\right)$视为一个整体来求导数的。

现在将式 2.5 利用泰勒展式在$\eta\left(x_0\right)$处展开：

$$0 = l'\left(\hat{\eta}\right) = l'\left(\eta\right) + l''\left(\eta\right)\left(\hat{\eta} - \eta\right) + o\left(\left\|\hat{\eta} - \eta\right\|\right), \tag{2.7}$$

利用式 2.7 可以得到：

$$\hat{\eta} - \eta = \left[-l''\left(\eta\right) + o\left(1\right)\right]^{-1} l'\left(\eta\right). \tag{2.8}$$

我们要证明估计的渐进正态性，只需计算以下两个方面即可：一，$l'\left(\eta\right)$的期望和方差，即证明他的渐进正态性；二，$-l''\left(\eta\right)$是一个依概率收敛的函数。

首先求$l'\left(\eta\right)$的期望和方差，证明他的渐进正态性：将$J\left(\eta\left(X_i\right), Y_i\right)$在点$x_0$处进行泰勒展开：

$$\begin{aligned} J\left(\eta\left(X_i\right), Y_i\right) = & J\left(\eta\left(x_0\right), Y_i\right) + J'_x\left(\eta\left(x_0\right), Y_i\right)\left(X_i - x_0\right) \\ & + \frac{1}{2} J''_{xx}\left(\eta\left(x_0\right), Y_i\right)\left(X_i - x_0\right)^2 + \mathrm{O}\left(\left\|X_i - x_0\right\|^2\right), \end{aligned}$$

则相应的$l'\left(\eta\right)$的期望可以表示为：

$$\begin{aligned} E\left[l'\left(\eta\right)\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n} J\left(\eta\left(X_i\right), Y_i\right) K_h\left(X_i - x_0\right)\right] \\ &= E\left[J\left(\eta\left(x_0\right), Y\right) K_h\left(X - x_0\right) + J'_x\left(\eta\left(x_0\right), Y\right)\left(X - x_0\right) K_h\left(X - x_0\right)\right. \\ &\quad \left. + \frac{1}{2} J''_{xx}\left(\eta\left(x_0\right), Y\right)\left(X - x_0\right)^2 K_h\left(X - x_0\right) + \mathrm{O}\left(\left\|X - x_0\right\|^2\right) K_h\left(X - x_0\right)\right] \\ &= E\left[J\left(\eta\left(x_0\right), Y\right) K_h\left(X - x_0\right)\right] + E\left[J'_x\left(\eta\left(x_0\right), Y\right)\left(X - x_0\right) K_h\left(X - x_0\right)\right] \\ &\quad + E\left[\frac{1}{2} J''_{xx}\left(\eta\left(x_0\right), Y\right)\left(X - x_0\right)^2 K_h\left(X - x_0\right)\right] + \mathrm{O}\left(\left\|X - x_0\right\|^2\right). \end{aligned} \tag{2.9}$$

现在只需求式 2.9 的前三部分的期望，即求$E\left(J\left(\eta\left(x_0\right), Y\right) K_h\left(X - x\right)_0\right)$、$E\left(J'_x\left(\eta\left(x_0\right), Y\right)\left(X - x_0\right) K_h\left(X - x_0\right)\right)$和$E\left(2^{-1} J''_{xx}\left(\eta\left(x_0\right), Y\right)\left(X - x_0\right)^2 K_h\left(X - x_0\right)\right)$。

$$\begin{aligned} \mathrm{I} &= E\left[J\left(\eta\left(x_0\right), Y\right) K_h\left(X - x_0\right)\right] \\ &= E\left[J\left(\eta\left(x_0\right), Y\right)\right] \int \frac{1}{h} K\left(\frac{u - x_0}{h}\right) f\left(u\right) du \\ &= E\left[J\left(\eta\left(x_0\right), Y\right)\right] f\left(x_0\right), \end{aligned}$$

$$\begin{aligned}\text{II} &= E\left[J'_x\left(\eta(x_0),Y\right)(X-x_0)K_h(X-x_0)\right]\\ &= E\left[J'_x\left(\eta(x_0),Y\right)\right]\int(u-x_0)\frac{1}{h}K\left(\frac{u-x_0}{h}\right)f(u)du\\ &= E\left[J'_x\left(\eta(x_0),Y\right)\right]\int hyK(y)f(x_0)dy\\ &= E\left[J'_x\left(\eta(x_0),Y\right)\right]hf(x_0)\int yK(y)dy,\end{aligned}$$

$$\begin{aligned}\text{III} &= E\left[\frac{1}{2}J''_{xx}\left(\eta(x_0),Y\right)(X-x_0)^2K_h(X-x_0)\right]\\ &= E\left[J''_{xx}\left(\eta(x_0),Y\right)\right]\int\frac{1}{2}(u-x_0)^2\frac{1}{h}K\left(\frac{u-x_0}{h}\right)f(u)du\\ &= E\left[J''_{xx}\left(\eta(x_0),Y\right)\right]h^2f(x_0)\int y^2K(y)dy,\end{aligned}$$

于是，

$$\begin{aligned}E\left[l'(\eta)\right] &= f(x_0)\left[E\left[J\left(\eta(x_0),Y\right)\right],E\left[J'_x\left(\eta(x_0),Y\right)\right],E\left[J''_{xx}\left(\eta(x_0),Y\right)\right]\right]\\ &\times\left[1,h\int yK(y)dy,h^2\int y^2K(y)dy\right]^T.\end{aligned}$$

相应的$l'(\eta)$的方差可以表示为：

$$\begin{aligned}Var\left[l'(\eta)\right] &= Var\left[\frac{1}{n}\sum_{i=1}^{n}J\left(\eta(X_i),Y_i\right)K_h(X_i-x_0)\right]\\ &= \frac{1}{n}Var\left[J\left(\eta(x_0),Y\right)K_h(X-x_0)\right]\\ &+\frac{1}{n}Var\left[J'_x\left(\eta(x_0),Y\right)(X-x_0)K_h(X-x_0)\right]\\ &+\frac{1}{n}Var\left[\frac{1}{2}J''_{xx}\left(\eta(x_0),Y\right)(X-x_0)^2K_h(X-x_0)\right]\\ &+\frac{1}{n}Var\left[\text{O}\left(\|X-x_0\|^2\right)K_h(X-x_0)\right],\end{aligned}$$

因为第二、三、四项均是高阶的，则只需要求第一项的方差即可：

$$\begin{aligned}Var\left[l'(\eta)\right] &= \frac{1}{n}Var\left[J\left(\eta(x_0),Y\right)K_h(X-x_0)\right]+\text{O}\left(\frac{h^2}{n}\right)\\ &= \frac{1}{n}Var\left[J\left(\eta(x_0),Y\right)\right]\int K_h^2(u-x_0)f(u)du\\ &= \frac{1}{n}Var\left[J\left(\eta(x_0),Y\right)\right]\int\frac{1}{h^2}K^2\left(\frac{u-x_0}{h}\right)f(u)du\\ &= \frac{1}{nh}Var\left[J\left(\eta(x_0),Y\right)\right]f(x_0)\int K^2(y)dy.\end{aligned}$$

则有：

$$\sqrt{nh}\left(l'(\eta)-E\left[l'(\eta)\right]\right)\xrightarrow{d}N\left(0,(nh)^{-1}Var\left[l'(\eta)\right]\right). \tag{2.10}$$

计算$-l''(\eta)$的依概率收敛函数：

$$\begin{aligned}-l''(\eta)&=-\frac{1}{n}\sum_{i=1}^{n}Q\left(\eta(X_i),Y_i\right)K_h(X_i-x_0)\\&\xrightarrow{p}-E\left[Q\left(\eta(X),Y\right)K_h(X-x_0)\right]\\&=-\int\frac{1}{h}K\left(\frac{u-x_0}{h}\right)f(u)E\left[Q\left(\eta(X),Y\right)|X=x_0\right]du\\&=-E\left[Q\left(\eta(X),Y\right)|X=x_0\right]\int\frac{1}{h}K\left(\frac{u-x_0}{h}\right)f(u)du\\&=-E\left[Q\left(\eta(X),Y\right)|X=x_0\right]\int K(y)f(hy+x_0)dy.\end{aligned} \tag{2.11}$$

将式 2.11 中的$f(hy+x_0)$在x_0处用泰勒展式展开：

$$\begin{aligned}f(hy+x_0)&=f(x_0)+f'(x_0)(hy)+\frac{1}{2}f''(x_0)(hy)^2+o\left((hy)^2\right)\\&=f(x_0)+o(hy).\end{aligned}$$

因为$K(y)$是对称核，所以有$\int yK(y)dy=0$，于是可以得到：

$$\begin{aligned}-l''(\eta)\xrightarrow{p}&=-E\left[Q\left(\eta(X),Y\right)|X=x_0\right]\int K(y)f(x_0)dy\\&=-E\left[Q\left(\eta(X),Y\right)|X=x_0\right]f(x_0).\end{aligned} \tag{2.12}$$

由式 2.10 和 2.12 可以得到：

$$\begin{aligned}\sqrt{nh}(\hat{\eta}-\eta)&=\sqrt{nh}\left[-l''(\eta)+o(1)\right]^{-1}l'(\eta)\\&\xrightarrow{d}N\Big(\left[-l''(\eta)+o(1)\right]^{-1}E\left[l'(\eta)\right],\\&\qquad\left[-l''(\eta)+o(1)\right]^{-2}(nh)^{-1}Var\left[l'(\eta)\right]\Big).\end{aligned} \tag{2.13}$$

所以，对应的$Bias(x_0)$和$Var(x_0)$分别为：

$$\begin{aligned}Bias(x_0)&=\left[-l''(\eta)+o(1)\right]^{-1}E\left[l'(\eta)\right]\\&=\left(-E\left[Q\left(\eta(X),Y\right)|X=x_0\right]f(x_0)\right)^{-1}E\left[l'(\eta)\right],\end{aligned}$$

$$\begin{aligned}Var(x_0)&=\left[-l''(\eta)+o(1)\right]^{-2}(nh)^{-1}Var\left[l'(\eta)\right]\\&=\frac{1}{nh}Var\left[J\left(\eta(x),Y\right)\right]f(x_0)\int K^2(y)dy.\end{aligned}$$

基于$Bias(x_0)$和$Var(x_0)$，可以得到最优窗宽的公式：

$$h_{opt}=\arg\min_h\left(\int Bias^2(x_0)dx_0+\int Var(x_0)dx_0\right). \tag{2.14}$$

对应的不同分布下的最优窗宽可根据式 2.14 来求得，具体的计算过程请参考 Aerts and Claeskens (1997)，下面仅给出 Aerts and Claeskens (1997)中的具体的窗宽表达式：

$$h_{opt} = C\left(K, f_X\right) C\left(\eta\left(x_0\right)\right) n^{-1/(2p+3)}. \tag{2.15}$$

其中 p 表示 $\eta\left(x_0\right)$ 泰勒展式展开的阶数，$C\left(K, f_X\right)$ 和 $C\left(\eta\left(x_0\right)\right)$ 分别表示：

$$C\left(K, f_X\right) = \left(\frac{(p+1)!\, p! \int_{R_{x_0}} K_p^2(z)\, dz}{f_X\left(x_0\right)\left(\int_{R_{x_0}} z^{p+1} K_p(z)\, dz\right)^2} \right)^{1/(2p+3)},$$

$$C\left(\eta\left(x_0\right)\right) = \left(\frac{\mathbf{I}^{-1}\left(\eta\left(x_0\right)\right)}{2\left[\eta^{(p+1)}\left(x_0\right)\right]^2} \right)^{1/(2p+3)}.$$

其中，R_{x_0} 表示 x_0 的定义域，$\mathbf{I}^{-1}(\cdot)$ 表示信息矩阵的逆矩阵，即：

$$\mathbf{I}^{-1}\left(\eta\left(x_0\right)\right) = \left[E\left(-\frac{\partial^2 l\left(\eta\left(x_0\right)\right)}{\partial^2 \eta\left(x_0\right)} \right) \right]^{-1}.$$

理论上是采用式 2.15 来求最优窗宽，但是实际上却是相当复杂的，Aerts and Claeskens (1997)指出一般用交叉验证的这个方法替代：

$$\hat{h}_{CV} = \arg\max_{h>0} \sum_{i=1}^{n} \log f\left(Y_i, \hat{\eta}^{[i]}\left(x_i\right)\right), \tag{2.16}$$

其中 $\hat{\eta}^{[i]}\left(x_i\right)$ 表示去掉第 i 个观测值 $\left(x_i, Y_i\right)$ 的估计量。本文的两个模型所对应的窗宽均是采用的式 2.16 的方法来选取的。

第三章 模型构建与估计

第一节 多处理因果效应模型构建

Feng et al. (2012)基于 Imbens (2000)所提出的多维处理变量进行了展开研究，估计不同的处理水平下的平均因果效应，为后续的学者研究多处理下的因果效应奠定了基础。多处理治疗方案模型是指处理变量不再是 0-1 变量，可能是三种及以上。Feng et al. (2012)定义的多处理的广义倾向得分如下：

我们有 m 个处理，$\zeta=\{1,2,\cdots,m\}$，让 $Y_i(t)$ 表示第 i 个个体接受处理 t 的潜在结果。设 Y_{it} 表示第 i 个个体所观察到的结果，设 $I(t)$ 为接受处理 t 的指标：

$$I(t)=\begin{cases}1, & T=t,\\ 0, & \text{其他}.\end{cases}$$

这里的 T 是表示个体所接受处理的随机变量。因此，每个个体都有 m 个潜在的结果。然而，在实际情况中，我们只能观察到指定处理下的一个结果。让 T_i 表示实际受到的处理。广义倾向得分 $r(t,X)$ 定义为下式：

$$r(t,x)=\Pr\left(T=t\middle|X=x\right),\quad (0<r(t,x)<1).$$

基于 Feng et al. (2012)离散化的多维处理变量下的倾向得分，本文基于连续型的倾向得分构建了连续型处理变量下的因果效应模型，如下所示。

生存分析下的数据一般是恒为正数的，本文所研究的处理变量“母亲怀孕年龄”对响应变量“婴儿出生体重”的影响，此处的处理变量为“母亲怀孕年龄”，恒为正数，故假设我们有 m 个处理，$m\in R^{+}$，让 $Y_i^{(t)}$ 表示第 i 个个体接受处理 t 的潜在结果。设 Y_{it} 表示第 i 个个体所观察到的结果，设 $I(t)$ 为接受处理 t 的指标：

$$I(t)=\begin{cases}1, & T\le t,\\ 0, & \text{其他}.\end{cases}$$

同理，T 是表示个体所接受处理的随机变量，t 表示所接受的处理 T 的值。因此，每个个体都有 m 个潜在结果。然而，在实际情况中，我们只能观察到指定处理下的一个结果。对广义倾向得分 $r(t,X)$ 进行定义：给定协变量 X 的接受处理 t 的条件概率，其表达式为：

$$r(t,x)=\Pr\left(T\le t\middle|X=x\right),\quad (0<r(t,x)<1).$$

其中，T 表示所接受处理，t 表示所接受处理 T 的值，而 $I(t)$ 是所接受处理 t 的指标。

给定观察到的协变量 X，$I(t)$ 是与 $Y^{(t)}$ 是独立的：

$$I(t)\perp Y^{(t)}\middle|X.$$

Imbens (2000)表明，如果给定观察到的协变量，处理T是与$Y^{(t)}$是独立的，那么给定广义倾向得分$r(t,x)$，处理T是与$Y^{(t)}$是独立的：

$$T \perp Y^{(t)} \left| r(t,x). \right.$$

设$\beta(t,r)$表示在广义倾向得分$r(t,x)=r$的治疗下个体因果效应。给定的协变量$X=x$，对所有的$t \in T$，有以下结果：

$$\beta(t,r)=E\left\{Y^{(t)} \left| r(t,x)=r \right.\right\}=E\left\{Y \left| T=t, r(t,x)=r \right.\right\}.$$

在处理t下的受试者的潜在结果的平均因果效应为：

$$E\left\{Y^{(t)}\right\}=E\left\{\beta\left(t, \left| r(t,x) \right.\right)\right\}.$$

在响应变量为连续型变量时，在处理t下的受试者的潜在结果的平均因果效应为：

$$\begin{aligned} E\left\{Y^{(t)}\right\} &= E\left\{\beta\left(t, \left| r(t,x) \right.\right)\right\} \\ &= \int y dF_{Y^{(t)}}(y). \end{aligned}$$

基于第二章中所示的离散的逆概率加权的权重，重新定义了连续型的逆概率加权权重，其表达式为：

$$W^{T}=\frac{1}{r(t|x)}=\frac{1}{\Pr\left[T \le t \left| X=x \right.\right]}.$$

此时，逆概率加权下的在处理t下的受试者的潜在结果的因果效应可以写成如下形式：

$$E\left\{Y^{(t)}\right\}=E\left\{\frac{Y^{(t)} I(T \le t)}{r(T \le t, X)}\right\}.$$

本文以“母亲怀孕年龄”为处理变量，“年龄”这个变量可以看成是一个连续的变量，可以用一个连续的分布函数来表示，再针对这个连续的分布函数进行分析。连续下的倾向得分就是一个分布函数，而本文选取的是“母亲怀孕年龄”为处理变量，所以倾向得分函数实际上就是一个关于年龄的分布函数。一般的，在生存分析中大多数学者对年龄的分布函数进行分析是基于指数模型还有 Cox 模型来分析年龄的分布的，如周天枢和陈崇帼(1992)指出指数分布常用作动物寿命的近似分布。当然，对倾向得分函数进行估计也可以基于其他分布来进行分析，不失一般性，本文对倾向得分函数进行估计时考虑指数模型和 Cox 模型这两个模型来进行估计，进而对估计出来的倾向得分函数进行加权计算出平均因果效应。

第二节 基于指数模型的多处理因果效应估计

在对多处理平均因果效应进行估计之前要对较多维的协变量进行降维处理，

因为较多的协变量会对平均因果效应的估计造成影响，协变量中的混杂会对影响正确的结果，如：将相关关系错误地认为是因果关系。当存在许多协变量时，非参数方法无法正确估计倾向得分函数，需要对协变量进行降维，因为本文针对的处理变量为连续型处理变量，故本文采用 SIR 算法进行充分降维，计算降维空间 $\hat{B}$。倾向得分函数具体降维的算法如下：

输入关于协变量和处理变量的样本$\left(X_i,T_i\right),\left(i=1,\cdots,n\right)$，

(a)对变量 X 进行标准化处理得到 $\tilde{X}_i=\hat{\Sigma}_X^{-1/2}\left(X_i-\overline{X}\right)$，其中 $\hat{\Sigma}_X$ 和 $\overline{X}$ 分别表示协变量 X 的样本协方差矩阵和样本均值。

(b)对数据 T 进行排序，不妨设 $T_{(1)}\le T_{(2)}\le\cdots\le T_{(n)}$。

(c)尽可能平均的将 T 划分为 H 个片，$I_1,I_2,\ldots,I_H$，$\hat{p}_h$ 表示 T_i 落入片 h 的概率，即：

$$\hat{p}_h=\frac{1}{n}\sum_{i=1}^{n}\delta_h\left(T_i\right),\delta_h\left(T_i\right)=\begin{cases}1,T_i\in I_h\\0,T_i\notin I_h\end{cases}.$$

(d)计算每个切片的样本均值：

$$\hat{m}_h=\frac{1}{n\hat{p}_h}\sum\nolimits_{i\in\mathbf{I}_h}\tilde{X}_i=\frac{1}{n\hat{p}_h}\sum\tilde{X}_i\,\delta_h\left(T_i\right).$$

(e)建立加权协方差阵 $\hat{M}=\sum_{h=1}^{H}\hat{p}_h\hat{m}_h\hat{m}_h^T$，求出其特征值和所对应的特征向量。

(f)记 $\hat{\eta}_d\left(d=1,2,\ldots,p\right)$ 表示 d 个特征向量，降维向量则为 $\hat{\beta}_d=\hat{\Sigma}_X^{-1/2}\hat{\eta}_d$。

在上述倾向得分函数降维的算法中，第(c)步和第(d)步对逆回归曲线进行标准化，得到$\mathrm{E}\left(\tilde{X}\middle|T\right)$，第(e)步处理了不同切片中样本量不同的情况，第(f)步把降维方向进行旋转，将其转回原始方向。因此 $\hat{\beta}_d$ 表示用来估计降维方向，而由 $\hat{\beta}_d$ 张开的空间 $\hat{B}$ 表示估计降维空间 B。

评价估计 $\hat{\beta}_d\left(d=1,2,\ldots,p\right)$ 的效果，将采用偏差(Bias)和均方误差(MSE)来评价。设 N 为重复的次数，则 Bias 和 MSE 的定义分别为：

$$\begin{aligned}Bias_l&=\frac{1}{N}\sum_{q=1}^{N}\hat{\beta}_{d,l,q}-\beta_{d,l},\\MSE_l&=\frac{1}{N}\sum_{q=1}^{N}\left(\hat{\beta}_{d,l,q}-\beta_{d,l}\right)^2.\end{aligned}\tag{3.1}$$

其中 $\hat{\beta}_{d,l,q}$ 表示结构维数为 $d\left(d\le p\right)$ 下第 $l\left(l=1,\cdots,p\right)$ 个分量重复 $q\left(q=1,\cdots,N\right)$ 次的估计值，$\beta_{d,l}$ 表示结构维数为 $d\left(d\le p\right)$ 下第 $l\left(l=1,\cdots,p\right)$ 个分量的真实值。$Bias_l$ 和 MSE_l 分别表示个分量的偏差和均方误差。

假设给定协变量 X 后，处理变量 T 的概率分布函数为指数分布：

$$r\left(t,x\right)=\Pr\left(T\le t\middle|X=x\right)=F\left(t\middle|x\right)=1-\exp\left[-\eta\left(x^TB\right)\cdot t\right],\tag{3.2}$$

因为 SIR 降维已经估计出来了降维空间 $\hat{B}$，所以对应的密度函数为：

$$f\left(t|x\right)=\eta\left(x^{T}\hat{B}\right)\cdot\exp\left[-\eta\left(x^{T}\hat{B}\right)\cdot t\right].$$

相应的似然函数为：

$$l=\prod_{i=1}^{n}\eta\left(X_{i}^{T}\hat{B}\right)\cdot\exp\left[-\eta\left(X_{i}^{T}\hat{B}\right)\cdot T_{i}\right].$$

相应的对数似然函数为：

$$\begin{aligned}\log l&=\sum_{i=1}^{n}\log\left\{\eta\left(X_{i}^{T}\hat{B}\right)\cdot\exp\left[-\eta\left(X_{i}^{T}\hat{B}\right)\cdot T_{i}\right]\right\}\\&=\sum_{i=1}^{n}\left\{\log\left[\eta\left(X_{i}^{T}\hat{B}\right)\right]-\eta\left(X_{i}^{T}\hat{B}\right)\cdot T_{i}\right\}\\&=\sum_{i=1}^{n}\log\left[\eta\left(X_{i}^{T}\hat{B}\right)\right]-\sum_{i=1}^{n}\eta\left(X_{i}^{T}\hat{B}\right)\cdot T_{i}.\end{aligned}$$

将 $\eta\left(X_{i}^{T}\hat{B}\right)$ 利用泰勒展式在给定 $x^{T}\hat{B}$ 处展开：

$$\begin{aligned}\eta\left(X_{i}^{T}\hat{B}\right)&=\eta\left(x^{T}\hat{B}\right)+\eta'\left(x^{T}\hat{B}\right)\left(X_{i}^{T}\hat{B}-x^{T}\hat{B}\right)+o\left(X_{i}^{T}\hat{B}-x^{T}\hat{B}\right)\\&\approx\alpha_{0}+\alpha_{1}\left(X_{i}^{T}\hat{B}-x^{T}\hat{B}\right).\end{aligned}$$

则相应的局部对数似然函数为：

$$\begin{aligned}L\approx&\frac{1}{n}\sum_{i=1}^{n}\log\left[\alpha_{0}+\alpha_{1}\left(X_{i}^{T}\hat{B}-x^{T}\hat{B}\right)\right]K_{h}\left(X_{i}^{T}\hat{B}-x^{T}\hat{B}\right)\\&-\frac{1}{n}\sum_{i=1}^{n}\left[\alpha_{0}+\alpha_{1}\left(X_{i}^{T}\hat{B}-x^{T}\hat{B}\right)\right]\cdot T_{i}\cdot K_{h}\left(X_{i}^{T}\hat{B}-x^{T}\hat{B}\right).\end{aligned}\tag{3.3}$$

本文中只对 α_{0} 感兴趣，所以对式 3.3 关于 α_{0} 求一阶偏导数，并令其为零，可以得到最大值，记为 $\hat{\alpha}_{0}$，即 $\hat{\eta}\left(x^{T}\hat{B}\right)=\hat{\alpha}_{0}$；同样的，如果一些学者对 α_{1} 感兴趣，也可采用相同的方法来估计 α_{1}，这里不再展开。在窗宽的选取上，可以利用式 2.15 来求得，但是本文利用式 2.16 来选取最优窗宽，利用相同的方法可以计算出指数模型下的窗宽，这里不再赘述。

将 $\hat{\eta}\left(x^{T}\hat{B}\right)=\hat{\alpha}_{0}$ 代入指数模型的分布函数中即可得到倾向函数的估计值，其估计值得表达式为：

$$\hat{r}\left(t,x\right)=\Pr\left(T\le t\middle|X=x\right)=F\left(t|x\right)=1-\exp\left[-\hat{\eta}\left(x^{T}\hat{B}\right)\cdot t\right],\tag{3.4}$$

下面将对平均因果效应估计，首先介绍个体因果效应。对于个体 i，观察结果 Y_{i} 可以表示为：

$$Y_{i}=\sum_{t=1}^{m}Y_{i}^{(t)}I\left(T_{i}\le t\right).$$

其中 T_{i} 表示个体 i 实际受到的处理，对个体 i 的处理 j 与处理 $k\,(j\neq k)$ 的处理

效果定义为：

$$TE = Y_i^{(j)} - Y_i^{(k)}.$$

若 $Y_i^{(j)}$ 或 $Y_i^{(k)}$ 都可以被观察到，但不能同时观察到(同上面的二分类下的处理结果)，所以个体的处理效果不能被识别。换句话说，不能评估处理 j 和处理 k 对个体 i 的影响。我们考虑种群中处理 j 与处理 k ($j \neq k$)的平均因果效应，其定义为：

$$ATE_{jk} = E\left(Y_i^{(j)}\right) - E\left(Y_i^{(k)}\right).$$

其中 X_i 为每个个体 i 的所含协变量。如果我们能观察到所有的潜在结果，那么上述利用 GPS 进行逆概率加权估计的平均因果效应就可以通过下面式子来表示：

$$\begin{aligned} A\hat{T}E_{jk} &= \frac{1}{n}\sum_{i=1}^{n} Y_i^{(j)} - \frac{1}{n}\sum_{i=1}^{n} Y_i^{(k)} \\ &= \left[\sum_{i=1}^{n} \frac{Y_i \cdot I\left(T_i \le j\right)}{\hat{r}\left(j, X_i\right)}\right]\left[\sum_{i=1}^{n} \frac{I\left(T_i \le j\right)}{\hat{r}\left(j, X_i\right)}\right]^{-1} \\ &\quad - \left[\sum_{i=1}^{n} \frac{Y_i \cdot I\left(T_i \le k\right)}{\hat{r}\left(k, X_i\right)}\right]\left[\sum_{i=1}^{n} \frac{I\left(T_i \le k\right)}{\hat{r}\left(k, X_i\right)}\right]^{-1}. \end{aligned} \tag{3.5}$$

为了使模型估计内容更紧凑，使得多处理平均因果效应估计程序结构清晰、可读性好、更直观，基于指数模型的多处理平均因果效应估计的伪代码如下：

算法：基于指数模型的多处理平均因果效应估计

输入：样本 $\left(X_1, T_1, Y_1\right), \left(X_2, T_2, Y_2\right), \cdots, \left(X_n, T_n, Y_n\right)$；

输出：处理 j 处理 k ($j \neq k$)的平均因果效应 $A\hat{T}E_{jk}$；

1 利用 SIR 算法估计降维空间 $\hat{B}$；

2 基于指数模型中式 3.3 估计函数 $\hat{\eta}\left(x^T \hat{B}\right)$；

3 将 $\hat{\eta}(\cdot)$ 代入式 3.2 中，得到式 3.4 中指数模型下的广义倾向得分函数 $\hat{r}(t, x)$；

4 将 $\hat{r}(t, x)$ 代入式 3.5 中，计算处理 j 处理 k ($j \neq k$)的平均因果效应 $A\hat{T}E_{jk}$。

第三节 基于 Cox 模型的多处理因果效应估计

同样的，基于 Cox 模型对多处理因果效应的估计也是要先进行 SIR 降维的，从而估计相应的降维空间 $\hat{B}$。下面，我们考虑一个基于 Cox 模型的半参数模型，假设处理变量 T 的强度函数为：

$$T \sim \lambda_0(t) \cdot \exp\left(\eta\left(x^T B\right)\right).$$

对应的概率分布是：

$$r(t,x)=\Pr\left(T\le t\middle|X=x\right)=F\left(t\middle|x\right)=1-\exp\left(-\Lambda_0(t)\cdot\exp\left(\eta\left(x^TB\right)\right)\right). \quad (3.6)$$

相应的记 $F_0(t)=1-\exp\left(-\Lambda_0(t)\right)$，则式 3.6 的分布函数变为：

$$\begin{aligned} r(t,x)&=\Pr\left(T\le t\middle|X=x\right)\\ &=F\left(t\middle|x\right)=1-\exp\left(-\Lambda_0(t)\exp\left(\eta\left(x^TB\right)\right)\right)\\ &=1-\left[1-F_0(t)\right]^{\exp\left(\eta\left(x^TB\right)\right)}. \end{aligned} \quad (3.7)$$

对式 3.7 进行估计，主要利用 Fan et al. (1997)中的方法，该方法分两步来进行：一，对 $\eta\left(x^T\hat{B}\right)$ 的函数形式进估计，二，对 $\Lambda_0(t)$ 进行估计。

定义累积基准危险函数：

$$\Lambda_0(t;\theta)=\sum_{j=1}^{N}\theta_jI\left\{T_j\le t\right\}.$$

其中，θ 表示跳跃度，T_j 表示第 j 个个体在时刻 t 死亡，$R_j=\left\{i:Z_i\ge T_j\right\}$，$Z_i=\min\left[T_i,C_i\right]$，$C_i$ 表示删失，则：

$$\Lambda_0(Z_i;\theta)=\sum_{j=1}^{N}\theta_jI\left\{i\in R_j\right\}.$$

因为 SIR 降维已经估计出来了降维空间 $\hat{B}$，所以对应的对数似然函数为：

$$\log l=\sum_{j=1}^{N}\left(\log\theta_j+\eta\left(X_{(j)}^T\hat{B}\right)\right)-\sum_{i=1}^{n}\left(\sum_{j=1}^{N}\theta_jI\left\{i\in R_j\right\}\exp\left(\eta\left(X_i^T\hat{B}\right)\right)\right). \quad (3.8)$$

最大化式 3.8，可以得到 $\theta_j\left(j=1,\ldots,N\right)$ 的估计值：

$$\hat{\theta}_j=\left[\sum_{i\in R_j}\exp\left\{\eta\left(X_i^T\hat{B}\right)\right\}\right]^{-1}. \quad (3.9)$$

将式 3.9 代入式 3.8 中，则：

$$\max_{\lambda_0}\log l=\sum_{j=1}^{N}\left(\eta\left(X_{(j)}^T\hat{B}\right)-\log\left[\sum_{i\in R_j}\exp\left\{\eta\left(X_i^T\hat{B}\right)\right\}\right]\right)-N.$$

Fan et al. (1997)同时指出：当 $\eta\left(X^T\hat{B}\right)$ 和 $\Lambda_0(t)$ 没有被明确的定义时，应该采用局部部分似然方法来估计 $\eta\left(x^T\hat{B}\right)$ 和 $\Lambda_0(t)$。

将 $\eta\left(X_i^T\hat{B}\right)$ 利用泰勒展式在给定 $x^T\hat{B}$ 处展开：

$$\begin{aligned} \eta\left(X_i^T\hat{B}\right)&=\eta\left(x^T\hat{B}\right)+\eta'\left(x^T\hat{B}\right)\left(X_i^T\hat{B}-x^T\hat{B}\right)+o\left(X_i^T\hat{B}-x^T\hat{B}\right)\\ &\approx\delta_0+\delta_1\left(X_i^T\hat{B}-x^T\hat{B}\right). \end{aligned}$$

则相应的对应的局部部分似然函数为：

$$L \approx \sum_{j=1}^{N} K_h\left(X_{(j)}^T\hat{B}-x^T\hat{B}\right)\left\{\delta_0+\delta_1\left(X_{(j)}^T\hat{B}-x^T\hat{B}\right)\right.$$
$$\left.-\log\left(\sum_{i\in R_j}\exp\left(\delta_0+\delta_1\left(X_i^T\hat{B}-x^T\hat{B}\right)\right)K_h\left(X_i^T\hat{B}-x^T\hat{B}\right)\right)\right\}. \tag{3.10}$$

对式 3.10 关于 δ_0 求一阶偏导，即可获得 $\hat{\eta}\left(x^T\hat{B}\right)=\hat{\delta}_0$，同理，一些学者要是对 δ_1 感兴趣，也可以进行相同的估计。

相应的：

$$\hat{\Lambda}_0(t)=\sum_{j=1}^{N}\left[\sum_{i\in R_j}\exp\left\{\hat{\eta}\left(X_i^T\hat{B}\right)\right\}\right]^{-1} I\left\{T_j\le t\right\}.$$

若对 $\lambda_0(t)$ 感兴趣，则可以利用下式来估计 $\lambda_0(t)$：

$$\hat{\lambda}_0(t)=\int W_g(t-x)\,d\hat{\Lambda}_0(x).$$

其中，W_g 表示给定窗宽 g 下的核函数。

最优窗宽的计算方式同指数模型下的方法一样，均是采用式 2.16 的方法来选取，这里也不再赘述。

将 $\hat{\eta}\left(x^T\hat{B}\right)=\hat{\delta}_0$ 代入 Cox 模型的分布函数中即可得到倾向函数的估计值，其估计值得表达式为：

$$\hat{r}(t,x)=\Pr\left(T\le t\middle|X=x\right)=F\left(t\middle|x\right)=1-\exp\left(-\Lambda_0(t)\cdot\exp\left(\hat{\eta}\left(x^T\hat{B}\right)\right)\right). \tag{3.11}$$

将上式 3.11 估计出来的 $\hat{r}(t,x)$ 代入式 3.5 中，可以得到 IPW 估计的平均因果效应。

同样的，基于 Cox 模型下的多处理平均因果效应估计的伪代码如下：

算法：基于 Cox 模型的多处理平均因果效应估计

输入：样本 $\left(X_1,T_1,Y_1\right),\left(X_2,T_2,Y_2\right),\cdots,\left(X_n,T_n,Y_n\right)$；

输出：处理 j 处理 $k\,(j\neq k)$ 的平均因果效应 $A\hat{T}E_{jk}$；

1 利用 SIR 算法估计降维空间 $\hat{B}$；

2 基于 Cox 模型中式 3.10 估计函数 $\hat{\eta}\left(x^T\hat{B}\right)$；

3 将 $\hat{\eta}(\cdot)$ 代入式 3.6 中，得到式 3.11 中指数模型下的广义倾向得分函数 $\hat{r}(t,x)$；

4 将 $\hat{r}(t,x)$ 代入式 3.5 中，计算处理 j 处理 $k\,(j\neq k)$ 的平均因果效应 $A\hat{T}E_{jk}$。

第四章 模拟

针对本文提出的多处理因果模型，本文采用 SIR 方法对高维协变量进行降维，对降维后的数据将采用非参数数方法估计 GPS，最后再基于 GPS 进行 IPW 来估计平均因果效应。

针对上述的模型，我们将进行模拟研究对广义倾向得分函数估计的效果，再进行因果推断。模拟按照以下步骤进行：首先，生成四维协变量 $X=\left(X_1,X_2,X_3,X_4\right)^T, p=4$，其中每个 X_i 都来自不同的分布：$X_1\sim N\left(0,1\right)$，$X_2\sim N\left(4,2\right)$，$X_3\sim B\left(\mathrm{n},0.5\right)$，$X_4\sim U\left(0,1\right)$。由于在医学和流行病学研究中经常看到非正态的结果变量，当结果变量来自非正态分布时，在估计多处理变量下的平均因果效应是至关重要的。

第一节 基于指数模型的多处理因果效应模拟

一、结构维数 $d=1$.

考虑以下函数形式：当 $\eta\left(X^T\beta\right)=\sin\left(X^T\beta\right)$ 时，其中 $d=1$，以及 $\beta=(1,0,0,0)^T$。

为了便于比较，假设式 3.2 和式 3.2 中的 $\eta(\cdot)$ 的函数形式是已知的，结构维数 d 和降维空间 B 均是已知的。尽管这是不现实的，但它提供了一个基准，因为它是人们期望获得的最佳估计。我们重复每个实验 $\Phi=500$ 次，样本量分别为 $n=500$ 和 $n=1000$。因为 Hsing and Carroll (1992)指出切片数在 $\sqrt{n}$ 至 $n/2$ 范围内，切片方法是收敛的，故选择切片数 $H=\sqrt{n}$ 到 $H=n/2$ 之间，等间距划分。设置 $\hat{R}_{d,H}$ 中的 Bootstrap 重复的次数为 Ψ=50。基于式 3.2 生成关于 $T_i\left(i=1,\cdots,n\right)$ 的数据；再将样本数据 $\left\{X_i,T_i\right\}\left(i=1,\cdots,n\right)$，利用下式生成响应变量 $Y_i\left(i=1,\cdots,n\right)$：

$$Y=\gamma_0 T+X^T\boldsymbol{\gamma}+\varepsilon.$$

其中 γ_0 表示随机的常数，$\boldsymbol{\gamma}$ 表示一个随机的向量，ε 表示随机误差项，即 $\varepsilon\sim N\left(0,1\right)$。这里设定 $\gamma_0=0.8$，$\boldsymbol{\gamma}=\left(0.5,0.2,-0.1,0.1\right)^T$，$\varepsilon\sim N\left(0,1\right)$。

基于指数模型按照相应的步骤生成样本 $\left\{X_i,T_i,Y_i\right\}\left(i=1,\cdots,n\right)$，首先利用样本数据 $\left\{X_i,T_i\right\}\left(i=1,\cdots,n\right)$ 进行 SIR 降维处理，估计降维空间 B。需要确定切片数的范围：当 $n=500$ 时，因为切片数在 $\sqrt{n}$ 至 $n/2$ 范围内，切片方法是收敛的，所以选择的切片数为 25、40、70、100、130、160、190、220、250；同理，当 $n=1000$ 时，所选择的切片数为 40、90、140、190、240、290、340、390、440、490。结构维数的取值范围为 $\left(1,p-1\right)$，其中 $p=4$。利用 R 软件进行数据分析，得到表

4。

从表 4 可以看出：当$n=500$时，$\hat{R}_{d,H}$的值接近于 1 时的结构维数和切片数分别为$d=1$和$H=25$，即当$n=500$时，所选择最优的结构维数和切片数分别为 1 和 25。当$n=1000$时，$\hat{R}_{d,H}$的值接近于 1 时的结构维数和切片数分别为$d=1$和$H=40$，即当$n=1000$时，所选择最优的结构维数和切片数分别为 1 和 40。

表 4 不同的切片数和不同的结构维数下的$\hat{R}_{d,H}$值

样本量	H \ d	1	2	3
n=1000	40	0.9993	0.8070	0.8110
	90	0.9990	0.8240	0.8330
	140	0.9980	0.7730	0.8500
	190	0.9980	0.6630	0.8530
	240	0.9980	0.7360	0.8560
	290	0.9970	0.7730	0.7870
	340	0.9970	0.7660	0.8030
	390	0.9960	0.7330	0.7890
	440	0.9940	0.7450	0.8120
	490	0.9940	0.7250	0.7510
n=500	25	0.9980	0.7860	0.8200
	40	0.9970	0.7520	0.8200
	70	0.9970	0.7490	0.8300
	100	0.9960	0.6870	0.8060
	130	0.9950	0.7340	0.8300
	160	0.9950	0.6940	0.8130
	190	0.9940	0.6760	0.8360
	220	0.9930	0.7020	0.8200
	250	0.9900	0.6970	0.8340

表 5 为表 4 重复 500 次后，均选取$d=1$和$H=40$下所进行的 SIR 降维的估计值。表 5 为降维空间各分量的估计效果，不管是$n=1000$还是$n=500$，Bias 和 MSE 均较小，且跟真实值相差较小，说明估计较准确，说明各分量的估计效果较好。

表 5 不同样本量下降维空间各分量的估计效果

		β_{11}	β_{12}	β_{13}	β_{14}
样本量	True	1	0	0	0
n=1000	Median	0.9993	-0.0028	-0.0152	0.0345
	Bias	-0.0007	-0.0028	-0.0152	0.0345
	MSE	0.0000	0.0000	0.0002	0.0012
n=500	Median	0.9818	0.0147	-0.0224	-0.1615
	Bias	-0.0171	0.0147	-0.0206	-0.1549
	MSE	0.0003	0.0002	0.0004	0.0242

关于T的四分位数点如表6，其四分位点一目了然，根据四分位点分析因果效应，如表7所示的不同处理段的平均因果效应的两两比较：当$n=1000$时，在处理段(-0.3004,0.0028)下的因果效应与在(0.0028,0.0254)下的因果效应相差0.2591，与在(0.0254,0.3291)下的因果效应相差0.3951；在(0.0028,0.0254)下的因果效应与在(0.0254,0.3291)下的因果效应相差0.1360。当$n=500$时，在处理段(-0.3603,-0.0027)下的因果效应与在(-0.0027,-0.0428)下的因果效应相差0.3868，与在(-0.0428,0.2720)下的因果效应相差0.5425；在(-0.0027,-0.0428)下的因果效应与在(-0.0428,0.2720)下的因果效应相差0.2412。总的来说，不管的在样本量为1000还是500时，不同处理段下的因果效应相差不大。

表6 关于T的描述性统计

样本量	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
n=1000	-3.7514	-0.3004	0.0028	0.0254	0.3291	5.2092
n=500	-5.53226	-0.3603	-0.0027	-0.0428	0.2720	5.2127

表7 不同处理段的因果效应两两比较

样本量	不同处理段	(-0.3004,0.0028)	(0.0028,0.0254)	(0.0254,0.3291)
n=1000	(-0.3004,0.0028)	0	0.2591	0.3951
	(0.0028,0.0254)	-	0	0.1360
	(0.0254,0.3291)	-	-	0
样本量	不同处理段	(-0.3603,-0.0027)	(-0.0027,-0.0428)	(-0.0428,0.2720)
n=500	(-0.3603,-0.0027)	0	0.3868	0.5425
	(-0.0027,-0.0428)	-	0	0.2412
	(-0.0428,0.2720)	-	-	0

二、结构维数$d=2$.

考虑以下函数形式：当$\eta\left(X^T\beta\right)=\left(X^T\beta_1\right)^2+\left(X^T\beta_2\right)$时，其中$d=2$，$\beta_1=(1,0,0,0)^T$，以及$\beta_2=(0,0,0,1)^T$。其中每个样本分量的设置均同$d=1$的情况，这里不再赘述。

表8 不同的切片数和不同的结构维数下的$\hat{R}_{d,H}$值(n=1000)

H \ d	1	2	3
40	0.4710	0.8280	0.8050
90	0.5060	0.7920	0.7980
140	0.5040	0.7140	0.8080
190	0.4760	0.6930	0.8000
240	0.5570	0.6250	0.7740
290	0.3260	0.6550	0.7810
340	0.4630	0.6010	0.8030
390	0.3170	0.4470	0.7560

440	0.3780	0.5620	0.7930
490	0.4060	0.6590	0.7790

从表 8 可以看出：当$n=1000$时，$\hat{R}_{d,H}$的值接近于 1 时的结构维数和切片数分别为$d=2$和$H=40$，即当$n=1000$时，所选择最优的结构维数和切片数分别为 2 和 40。从表 9 可以看出：降维空间各个分量的估计效果较好，因为 Bias 和 MSE 相差均不大，说明估计的效果较好。

表 9 降维空间各分量的估计效果(n=1000)

	β_{11}	β_{12}	β_{13}	β_{14}	β_{21}	β_{22}	β_{23}	β_{24}
True	1	0	0	0	0	0	0	1
Median	0.9562	0.1497	-0.0957	0.1025	-0.1092	0.0435	-0.0228	0.9844
Bias	-0.0421	0.1571	-0.0957	0.1025	-0.1092	0.0435	-0.0228	-0.0158
MSE	0.0015	0.0255	0.0013	0.0230	0.0223	0.0023	0.0010	0.0003

关于T的四分位数点如表 10，其四分位点一目了然，最小值为-9.2391，最大值为 19.0791，然后根据四分位点来分析因果效应，如表 11 所示的不同处理段的平均因果效应，不同处理段下的因果效应相差不大。在处理段(0.0426,0.4664)下的因果效应与在(0.4664,1.2882)下的因果效应相差 0.4281，与在(1.2882,1.6924)下的因果效应相差 0.5117；在(0.4664,1.2882)下的因果效应与在(1.2882,1.6924)下的因果效应相差 0.0836。

表 10 关于 T 的描述性统计

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-9.2391	0.0426	0.4664	1.2882	1.6924	19.0791

表 11 不同处理段的因果效应两两比较

	(0.0426,0.4664)	(0.4664,1.2882)	(1.2882,1.6924)
(0.0426,0.4664)	0	0.4281	0.5117
(0.4664,1.2882)	-	0	0.0836
(1.2882,1.6924)	-	-	0

第二节 基于 Cox 模型的多处理因果效应模拟

一、结构维数$d=1$.

考虑以下函数形式：当$\eta\left(X^T\beta\right)=\sin\left(X^T\beta\right)$时，其中$d=1$，以及$\beta=(1,0,0,0)^T$。

基于 Cox 模型的多处理因果效应模拟生成的各种参数均和指数模型下$d=1$的情形相同，设定$\lambda_0=0.8$；这里不再赘述。

从表 12 可以看出：当$n=500$时，$\hat{R}_{d,H}$的值接近于 1 时的结构维数和切片数

分别为 $d=1$ 和 $H=25$，即当 $n=500$ 时，所选择最优的结构维数和切片数分别为 1 和 25。此时，结构维数小于 $X_{n\times p}=\left(X_1,X_2,X_3,X_4\right)^T$ 的维数，达到了降维的目的。当 $n=1000$ 时，$\hat{R}_{d,H}$ 的值接近于 1 时的结构维数和切片数分别为 $d=1$ 和 $H=40$，即当 $n=1000$ 时，所选择最优的结构维数和切片数分别为 1 和 40。此时，结构维数小于 $p=4$，也达到了降维的目的。总的来看，当 $n=1000$ 时，不同的切片数和不同的结构维数下的 $\hat{R}_{d,H}$ 值相比于 $n=500$ 时的 $\hat{R}_{d,H}$ 值较大，说明样本量越大，$\hat{R}_{d,H}$ 值越接近于 1，能较好的选择结构维数和切片数。

表 12 不同的切片数和不同的结构维数下的 $\hat{R}_{d,H}$ 值

样本量	H \ d	1	2	3
n=1000	40	0.9900	0.7390	0.8050
	90	0.9840	0.8080	0.8710
	140	0.9770	0.8110	0.8190
	190	0.9720	0.7960	0.8920
	240	0.9590	0.8060	0.8660
	290	0.9430	0.7800	0.8810
	340	0.9360	0.8460	0.8370
	390	0.9220	0.8200	0.8550
	440	0.8820	0.7480	0.8760
	490	0.8890	0.7490	0.8360
n=500	25	0.9820	0.7900	0.8340
	40	0.9570	0.7550	0.8140
	70	0.9620	0.7700	0.8390
	100	0.9540	0.8090	0.8540
	130	0.9430	0.7940	0.8150
	160	0.9350	0.7190	0.8330
	190	0.9280	0.6650	0.7920
	220	0.9000	0.7770	0.7760
	250	0.8680	0.7240	0.7970

表 13 不同样本量下降维空间各分量的估计效果

		β_{11}	β_{12}	β_{13}	β_{14}
样本量	True	1	0	0	0
n=1000	Median	0.9818	-0.0462	0.0319	-0.1813
	Bias	-0.0238	-0.0462	0.0319	-0.1813
	MSE	0.0039	0.0002	0.0010	0.0329
n=500	Median	0.9083	-0.0699	-0.0270	0.4115
	Bias	-0.0917	-0.0699	-0.0270	0.4115
	MSE	0.0084	0.0049	0.0007	0.0732

针对不同的样本量进行 SIR 降维，采用估计的中位数来表示降维空间 $\hat{B}$ 的估

计值，表 13 是基于表 12 重复 500 次的实验而得到的，重复 500 次发现：不管 $n=500$ 还是 $n=1000$ 每个的 $\hat{R}_{d,H}$ 值均是在 $d=2$ 时取得最大值。使用 Bias 和 MSE 来评价降维空间估计的效果，重复次数为 500 次，得到表 13。从表 13 可以看出降维空间各分量的估计效果不错，所对应的 MSE 和 Bias 均较小，且各分量的中位数跟真实值相差不大。

表 14 为不同样本量下的关于 T 的描述性统计表，根据四分位点分析因果效应，如表 15 所示的不同处理段的平均因果效应：当 $n=1000$ 时，在(0.2938,0.7971)处理段下的因果效应与在(0.7971,1.3662)处理段下的因果效应相差 0.0042，与在(1.3662,1.7821)处理段下的因果效应相差 0.0504；在(0.7971,1.3662)处理段下的因果效应与在(1.3662,1.7821)处理段下的因果效应相差 0.0462。当 $n=500$ 时，在(0.2818,0.8184)处理段下的因果效应与在(0.8184,1.6411)处理段下的因果效应相差为 0.0302，与在(1.6411,2.0704)处理段下的因果效应相差 0.0814；在(0.8184,1.6411)处理段下的因果效应与在(1.6411,2.0704)处理段下的因果效应相差 0.0512。总的来说，不管的在样本量为 1000 还是 500 时，不同处理段下的因果效应相差较小。

表 14 关于 T 的描述性统计

样本量	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
n=1000	0.0001	0.2938	0.7971	1.3662	1.7821	13.0353
n=500	0.0008	0.2818	0.8184	1.6411	2.0704	18.8579

表 15 不同处理段的因果效应

样本量	不同处理段	(0.2938,0.7971)	(0.7971,1.3662)	(1.3662,1.7821)
n=1000	(0.2938,0.7971)	0	0.0042	0.0504
	(0.7971,1.3662)	-	0	0.0462
	(1.3662,1.7821)	-	-	0
样本量	不同处理段	(0.2818,0.8184)	(0.8184,1.6411)	(1.6411,2.0704)
n=500	(0.2818,0.8184)	0	0.0302	0.0814
	(0.8184,1.6411)	-	0	0.0512
	(1.6411,2.0704)	-	-	0

二、结构维数 $d=2$.

考虑以下函数形式：当 $\eta\left(X^T\beta\right)=\left(X^T\beta_1\right)^2+\left(X^T\beta_2\right)$ 时，其中 $d=2$，$\beta_1=(1,0,0,0)^T$，以及 $\beta_2=(0,0,0,1)^T$。其中每个样本分量的设置均同 $d=1$ 的情况，这里不再赘述。

表 16 为不同的切片数和不同的结构维数下的 $\hat{R}_{d,H}$ 值，虽然在生成数据时，我们设定的结构维数为 2，且当 $n=1000$ 时，可以明显的看出当 $\hat{R}_{d,H}$ 值最大时，

$d=2$ 和 $H=40$，即选择结构维数为 2 和切片数为 40 时进行降维空间估计，此时 $d=2<p=4$，也达到了降维的目的。同样的，当 $n=500$ 时，在 $d=2$ 和 $H=25$ 时，$\hat{R}_{d,H}$ 值最大，故选择结构维数为 2 和切片数为 25 时进行降维空间估计。总的来看，当 $n=1000$ 时，不同的切片数和不同的结构维数下的 $\hat{R}_{d,H}$ 值相比于 $n=500$ 时的 $\hat{R}_{d,H}$ 值较大，说明样本量越大，$\hat{R}_{d,H}$ 值越接近于 1，能较好的选择结构维数和切片数。

表 16 不同的切片数和不同的结构维数下的 $\hat{R}_{d,H}$ 值

样本量	H \ d	1	2	3
n=1000	40	0.8220	0.9800	0.8670
	90	0.7300	0.9620	0.8760
	140	0.6440	0.9440	0.8860
	190	0.5930	0.9290	0.8610
	240	0.5290	0.8920	0.8620
	290	0.4040	0.8900	0.8720
	340	0.4070	0.8770	0.8160
	390	0.4720	0.8680	0.8420
	440	0.4210	0.8370	0.7777
	490	0.4100	0.8050	0.8360
n=500	25	0.9390	0.9470	0.8160
	40	0.9430	0.9240	0.8030
	70	0.9000	0.9000	0.8190
	100	0.8020	0.8770	0.8510
	130	0.7900	0.7910	0.8540
	160	0.7890	0.8110	0.8150
	190	0.7350	0.8050	0.8390
	220	0.7650	0.6810	0.8520
	250	0.6320	0.6690	0.8680

表 17 不同样本量下降维空间各分量的估计效果

		β_{11}	β_{12}	β_{13}	β_{14}	β_{21}	β_{22}	β_{23}	β_{24}
样本量	True	1	0	0	0	0	0	0	1
n=500	Median	0.9947	-0.0372	0.0156	-0.0057	-0.1661	-0.3257	0.1922	0.9107
	Bias	-0.0053	-0.0372	0.0156	-0.0057	-0.1661	-0.3257	0.1922	-0.0893
	MSE	0.0421	0.0014	0.0002	0.0003	0.0276	0.1061	0.0369	0.0043
n=1000	Median	0.9594	0.0803	0.0108	0.2701	0.0072	0.0045	0.0280	0.9996
	Bias	-0.0406	0.0803	0.0108	0.2701	0.0072	0.0045	0.0280	-0.0004
	MSE	0.0016	0.0065	0.0001	0.0730	0.000	0.0000	0.0008	0.0000

表 17 是基于重复 500 次表 16 的实验而得到的，重复 500 次发现不管 $n=500$ 还是 $n=1000$ 每个的 $\hat{R}_{d,H}$ 值均是在 $d=2$ 时取得最大值，在下面的指数模型中也是

同样如此。在表 17 中，当$n=500$时和当$n=1000$时，可以看出降维空间$\hat{B}$的各分量的估计效果不错，所对应的 MSE 和 Bias 均较小，各分量的中位数跟真实值相差不大。

表 18 为关于T的描述性统计，根据四分位点分析因果效应，如表 19 所示的不同处理段的平均因果效应，当$n=1000$时，在(0.1258,0.3336)的处理段下的因果效应与在(0.3336,0.5958)的处理段下的因果效应相差 0.0146，与在(0.5958,0.7802)的处理段下的因果效应相差 0.0131；在(0.3336,0.5958)的处理段下的因果效应与在(0.5958,0.7802)的处理段下的因果效应相差 0.0129。同样的，当$n=500$时，在(0.1205,0.3421)的处理段下的因果效应与在(0.3421,0.7095)的处理段下的因果效应相差 0.1288，与在(0.7095,0.9059)的处理段下的因果效应相差 0.0131；在(0.3421,0.7095)的处理段下的因果效应与在(0.7095,0.9059)的处理段下的因果效应相差 0.0021。总的来说，不管的在样本量为 1000 还是 500 时，不同处理段下的因果效应相差不大，每个处理段下的因果效应均在一个固定值波动。

表 18 关于 T 的描述性统计

样本量	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
n=1000	0.0000	0.1258	0.3336	0.5958	0.7802	5.8048
n=500	0.0004	0.1205	0.3421	0.7095	0.9059	10.8216

表 19 不同处理段的因果效应

样本量	不同处理段	(0.1258,0.3336)	(0.3336,0.5958)	(0.5958,0.7802)
	(0.1258,0.3336)	0	0.0146	0.0017
n=1000	(0.3336,0.5958)	-	0	0.0129
	(0.5958,0.7802)	-	-	0
样本量	不同处理段	(0.1205,0.3421)	(0.3421,0.7095)	(0.7095,0.9059)
	(0.1205,0.3421)	0	0.1288	0.0131
n=500	(0.3421,0.7095)	-	0	0.0021
	(0.7095,0.9059)	-	-	0

第五章 实例分析

随着生物医学的发展，人们开始认识到生存的重要性，所以研究对人类生存有影响的因素是一个重要的问题。生存分析广泛应用于医疗数据研究中，常常用追踪的方式来研究事物的发展变化，如研究某一药物对个体的治疗效果，因而常常和因果推断相结合。由于生存数据具有样本小、维数高和非线性的特点，传统的降维方法将面临“小样本”和“维数灾难”的问题。现有统计方法的最大挑战之一是从海量数据中提取有效信息，并将其转化为做决策时所需的信息。为了解决此问题，数据的降维方法已成为人们关注的焦点。对高维数据进行降维具有以下优点：随着维度的减少，用于低维数据的统计方法可以应用于降维后的数据。其次，人的视觉受到多维空间的限制，低维数据易于查看。如果直接将传统的降维方法应用于结构复杂多变的高维数据中可能会存在“维数灾难”等问题，充分降维能化解此类问题。SIR 为充分降维重的经典算法，该方法研究多维自变量和单变量之间的关系，在对自变量的降维过程中充分利用了因变量的信息，能够有效的进行降维，从而化解了“维数灾难”等问题。

健康和疾病的起源说表明，基因本身和所处的环境会对人的健康产生影响。如果胎儿在出生后早期的关键阶段遇到不利条件，可能会导致一些人存在潜在的慢性疾病。新生儿作为生命的起源，其出生数量的下降趋势，在一定程度上会对社会经济、人口老龄化等问题造成影响。目前人口老龄化程度加重，国家出台三胎政策来化解问题，但是面对三胎下的母亲怀孕年龄逐渐增大的趋势，使得对母亲怀孕年龄的探讨又一次成为了研究的热点，因为产妇的高龄一般认为会对胎儿的发育造成影响，衡量指标为婴儿出生体重，因为出生体重已成为婴儿健康的主要指标，也是婴儿健康政策的中心焦点。低出生体重儿会经历发育困难等问题，这不仅会给家庭带来经济问题，还会给社会带来巨大的成本。母亲体内的营养和代谢情况是影响婴儿出生体重的因素之一，孕期的母亲承担着自身及胎儿营养的双重供给，是胎儿发育的特殊时期，且孕妇体内营养的供给与母亲怀孕年龄一般情况下是有一定的相关性的，其原因是如果母亲过早的怀孕，母亲自身发育的成熟度不够，胎儿与发育中的母亲可能会存在争夺营养的情况，这不仅会影响母亲的健康，还会影响胎儿的发育；如果母亲怀孕太晚，可能会存在母亲的身体机能减退的情况，一般来说高龄产妇容易引发一些如妊娠期高血压疾病或糖尿病等并发症和合并症，一旦有了合并症，胎儿的畸形率就会偏高一些。故研究母亲怀孕年龄对婴儿体重的影响是很有价值的，在一定程度上可以避免过早怀孕或是过晚怀孕，使得婴儿尽可能地避免面临早产、畸形、夭折等情况。

表 20 数据集描述性统计

Variable	Value	Interpretation
bweight	连续变量	婴儿体重
mmarried	married.已结婚；notmarried.未结婚	母亲是否结婚
mhisp	0 否；1 是	母亲是否非西班牙裔黑人
fhisp	0 否；1 是	父亲是否非西班牙裔黑人
foreign	0 否；1 是	是否外国人
alcohol	0 否；1 是	怀孕间是否饮酒
deadkids	0 否；1 是	之前是否有新生儿死亡
mage	连续变量	母亲年龄
medu	有序分类变量	母亲受教育程度
fage	连续变量	父亲年龄
fedu	有序分类变量	父亲受教育程度
nprenatal	连续变量	产前检查次数
monthslb	离散变量	自上个孩子出生到现在的月数间隔
order	有序分类变量	该婴儿在多胞胎中的排序
msmoke	分类变量	母亲每天抽烟的根数
mbsmoke	0 否；1 是	母亲是否抽烟
mrace	0-1 变量	母亲的种族
frace	0-1 变量	父亲的种族
prenatal	有序分类变量	从怀孕到首次产检的月份间隔
birthmonth	离散变量	婴儿出生月份
lbweight	0 否；1 是	是否是低体重婴儿
fbaby	0 否；1 是	是否是第一个孩子
prenatal1	0-1 变量	是否首次怀孕

注：数据下载地址：http://www.stata-press.com/data/r13/cattaneo2.dta.

本文基于原始数据(Cattaneo, 2010)，其中包含了 4642 名婴儿的出生信息，共有 23 个变量，本文认为母亲的年龄也是一个重要变量(例如，高龄产妇的对新生婴儿的健康有重大影响)，故采用“母亲怀孕年龄”为处理变量，“婴儿出生体重”为响应变量，此时其余的 21 个变量均为协变量，研究处理变量“母亲怀孕年龄”对响应变量“婴儿出生体重”的因果效应。同时 Liu et al. (2018)论文数据也是基于此原始数据进行的分析，该论文里考虑的“母亲抽烟与否对婴儿体重的影响”，针对的是 0-1 变量，面对高维协变量的时候，无法使用 SIR 进行降维，他采用的是几何技术进行降维，该方法计算较复杂。值得注意的是，母亲的年龄

是一个连续变量，因此可以使用 SIR 来考虑高维协变量的降维，从而简化计算，再采用非参数方法估计降维后的倾向得分函数，切片之后的处理变量为多处理变量，进而基于逆概率加权方法使用多治疗方案模型来分析母亲怀孕年龄对新生婴儿体重的因果效应，并依据此提出相关的建议，能有效防止婴儿出现早产、畸形、夭折等问题。

实证数据集样本量为 4642，此处要研究母亲怀孕年龄对婴儿体重的因果效应，首先了解相关变量的含义及变量类型，如表 20。表中显示：处理变量和响应变量均是连续型的，剩下的为协变量，协变量的变量类型是多样的：连续、离散、有序等。

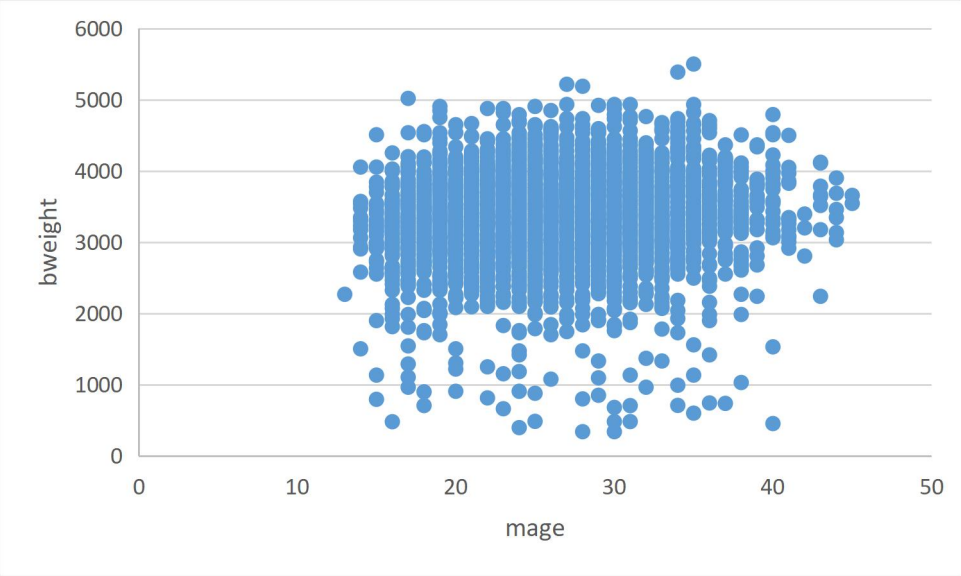

图 2 婴儿出生体重关于母亲怀孕年龄的散点图

图 2 为响应变量关于处理变量的散点图，可以明显得看出绝大多数的数据点均聚集在 2000(g)到 5000(g)之间，还有少部分散落在 0(g)到 2000(g)之间，仅有少部分的数据在大于 5000(g)之外。查验相关资料发现，婴儿出生体重的正常范围为 2500(g)-4000(g)之间；如果婴儿出生体重在 1000(g)以下，则为超低出生体重儿，身体发育情况可能会非常差，甚至出现死亡现象；如果婴儿出生体重在 1000(g)-1500(g)之间，为极地出生体重儿，比较容易患有各种疾病，身体健康状况欠佳；如果婴儿出生体重在 1500(g)-2500(g)之间，则为低出生体重儿，可能出现营养不良或者其他疾病等；如果婴儿出生体重在 4000(g)以上，则为巨大儿，比较肥胖，容易发生缺血缺氧性脑部疾病。从图 4 可以看出有部分婴儿出生体重的数据落在 2500(g)-4000(g)之外的，故研究母亲怀孕年龄对婴儿出生体重的影响是有意义的，可以在婴儿出生体重出现异常时，及时与医生沟通并进行相关治疗，当然研究此项目的目的主要是为了预防母亲过早或过晚怀孕。

实例分析将基于 Cox 模型进行因果效应分析，首先要对数据进行预处理，即

剔除一些无关变量，且倾向得分函数只有在协变量 X 是处理 T 的后门路径的时候才能使用，即 X 对 T 有影响(Pearl et al., 2020)；在实例分析中，协变量 X 必须选择对处理变量 T 有影响的变量(后门准则)。因为我们不知道在这 21 个变量中到底哪一个变量或者哪几个变量会影响处理变量，故将这 21 个变量全部纳入协变量中，和处理变量一起进行 SIR 处理，但是变量“msmoke”和变量“mbsmoke”这两者之间存在高度的共线性，即母亲是否抽烟(mbsmoke)与母亲每天抽烟的根数(msmoke)存在共线性，因为如果母亲抽烟一定的大于等于 1 根的，会造成在 SIR 所求得的特征向量矩阵不存在，所以我们将剔除变量“msmoke”。剔除变量“msmoke”之后还剩 20 个协变量。

只有当结构维数 d 小于协变量的个数 p 时，才能达到降维的目的，故 d=l,⋯,19 。表 21 为不同切片数和不同的结构维数下的 $\hat{R}_{d,H}$ 值，可以明显的看出：相同切片数下，随着结构维数的不断增大，$\hat{R}_{d,H}$ 值在逐渐减小，其中结构维数在 8-19 之间的 $\hat{R}_{d,H}$ 值均小于结构维数在 1-7 之间的 $\hat{R}_{d,H}$ 值，故忽略不计。纵向来看，随着切片数 H 的不断增加，不同结构维数下的 $\hat{R}_{d,H}$ 值变化各有不同，相同结构维数下的 $\hat{R}_{d,H}$ 值随着切片数的增加呈现减小的趋势，其中当 $d=2$ 和 $H=70$ 时，$\hat{R}_{d,H}$ 值达到最大，故选择 $d=2$ 和 $H=70$ 进行后续分析。

表 21 不同的切片数和不同的结构维数下的 $\hat{R}_{d,H}$ 值

H \ d	1	2	3	4	5	6	7	8-19
70	0.5170	0.8890	0.7680	0.7840	0.7940	0.7220	0.7300	...
270	0.6200	0.7130	0.7200	0.7130	0.6320	0.6040	0.6150	...
470	0.3950	0.6940	0.6600	0.5500	0.5040	0.4930	0.5130	...
670	0.5280	0.6270	0.6250	0.5190	0.4810	0.4800	0.4910	...
870	0.5070	0.5890	0.5310	0.4490	0.4220	0.4240	0.4430	...
1070	0.4520	0.4830	0.4670	0.4060	0.3860	0.4070	0.4330	...
1270	0.4410	0.4020	0.3250	0.3090	0.3240	0.3550	0.4010	...
1470	0.4650	0.3930	0.3210	0.3130	0.3370	0.3670	0.4030	...
1670	0.4390	0.3580	0.3060	0.3020	0.3270	0.3610	0.4000	...
1870	0.3990	0.3190	0.2810	0.2960	0.3220	0.3520	0.3980	...
2070	0.3960	0.2640	0.2560	0.2820	0.3070	0.3540	0.4050	...
2270	0.3850	0.2500	0.2390	0.2670	0.3000	0.3520	0.3980	...

计算出来的 $\hat{B}$ 为：

$$\hat{B}=\left(\hat{\beta}_1,\hat{\beta}_2\right)$$

$$\hat{\beta}_1=\begin{bmatrix}-0.6108\\0.3255\\0.1554\\-0.3045\\-0.2941\\-0.2502\\-0.1418\\-0.0411\\0.0339\\-0.0049\\-0.0074\\-0.2326\\-0.0065\\-0.2686\\0.2333\\0.0093\\-0.0009\\0.0082\\-0.1084\\-0.1809\end{bmatrix},\hat{\beta}_2=\begin{bmatrix}-0.4480\\0.1925\\0.0731\\-0.0850\\0.0551\\-0.0245\\-0.0122\\0.0144\\-0.0002\\-0.0005\\0.0090\\0.1873\\-0.2483\\0.1597\\-0.2174\\0.0515\\-0.0032\\0.0970\\0.7506\\-0.0228\end{bmatrix}$$

再基于 $\hat{B}$ 的估计值进行局部似然，再将局部似然得到的估计值代入 Cox 模型中，得到估计的倾向得分函数 $\hat{r}(t,x)$。再将 GPS 进行 IPW 来估计因果效应，表 22 为母亲怀孕年龄描述性分析表，以母亲怀孕年龄的四分位数进行划分，样本中母亲怀孕年龄的最小值为 13，最大值为 45，平均数为 26.5。因为年龄没有小数，所以选择 27 岁为一个年龄点，一般来说，18 岁以上母亲为成年人，35 岁以上的母亲怀孕为高龄的孕妇，在随机选取了年龄为 40 岁的点，所以选取了 13、18、22、27、30、35、40 以及 45 岁这八个年龄点来进行年龄段的划分。

表 22 母亲怀孕年龄描述性分析

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
13	22	26	26.5	30	45

表 23 是不同年龄段下的因果效应的两两比较，即不同年龄段下的因果效应相减，其结果呈现出一个对角线上为零的上三角矩阵。从表 23 横向的来看，母亲怀孕年龄在 13-18 岁时与母亲怀孕年龄在 18-22 岁时的因果效应差为 837.5031，说明在 13-18 岁时怀孕，对婴儿出生的体重的影响较大，可能是因为 13-18 岁的女性的身体各方面的机能还未发育完全，且在 13-18 岁这个年龄段下怀孕的因果效应比在 18-22 岁这个年龄段下怀孕的因果效应大 837.5031；同理，母

亲怀孕年龄在 13-18 岁时与母亲怀孕年龄在 22-26 岁时的因果效应差为 872.4980；母亲怀孕年龄在 13-18 岁时与母亲怀孕年龄在 26-27 岁时的因果效应差为 891.2130；母亲怀孕年龄在 13-18 岁时与母亲怀孕年龄在 27-30 岁时的因果效应差为 896.2041；母亲怀孕年龄在 13-18 岁时与母亲怀孕年龄在 30-35 岁时的因果效应差为 896.3441；母亲怀孕年龄在 13-18 岁时与母亲怀孕年龄在 35-40 岁时的因果效应差为 933.6485；母亲怀孕年龄在 13-18 岁时与母亲怀孕年龄在 40-45 岁时的因果效应差为 4158.6860；总的来说，随着年龄的增加，母亲在 13-18 岁这个年龄段怀孕下的因果效应跟其他年龄段下的因果效应的差值在逐渐增大，即横向来看：13-18 岁这个年龄段的因果效应和 18-22、22-26、26-27、27-30、30-35、35-40、40-45 各个年龄段下的因果效应的差异在逐渐上升。同理，其他各个年龄段下的因果效应差值跟年龄在 13-18 岁下的因果效应差的计算一样，这里不再赘述，显然的，其他各个年龄段下的因果效应的差异跟年龄在 13-18 岁下的因果效应差异所呈现的趋势也一样，均有增大的趋势。

表 23 各个年龄段对婴儿体重的因果效应两两比较

	13-18	18-22	22-26	26-27	27-30	30-35	35-40	40-45
13-18	0	837.5031	872.4980	891.2130	896.2041	896.3441	933.6485	4158.6860
18-22	-	0	34.9949	53.7099	58.7009	58.8409	96.1454	3321.1830
22-26	-	-	0	18.7150	23.7060	23.8460	61.1505	3286.1880
26-27	-	-	-	0	4.9910	5.1310	42.4355	3267.4730
27-30	-	-	-	-	0	0.1400	37.4445	3262.4820
30-35	-	-	-	-	-	0	37.3045	3262.3420
35-40	-	-	-	-	-	-	0	3225.0380
40-45	-	--	-	-	-	-	-	0

注：此表数值均是用前一个(小的)年龄段下的因果效应减去后一个(大的)年龄段下的因果效应所得。

特别地，母亲怀孕年龄在 18-22 岁下与在 22-26、26-27、27-30、30-35、35-40、40-45 岁下的因果效应差值分别为 34.9949、53.7099、58.7009、58.8409、96.1454、3321.1830，这说明母亲怀孕年龄在 18-22 岁下的因果效应与其他各个年龄段下的因果效应相差较大，且随着年龄的增加，其差异在不断增加，母亲怀孕年龄在 18-22 岁这个年龄段下，母亲的身体机能可能发育完全，但可能是 18-22 岁下母亲的心理未成熟，而且本文中的因果效应是基于 SUTVA 假设的，而 18-22 这个年龄段下的母亲初为人母不了解相关知识，使得自己被他人所影响，造成 SUTVA 假设不成立，所以造成 18-22 岁这个年龄段下的因果效应较大。母亲怀孕年龄在 22-26 岁下的因果效应与母亲怀孕年龄在 26-27、27-30、30-35、35-40、40-45 岁下的因果效应差值分别为 18.7150、23.7060、23.8460、61.1505、

3286.1880，这说明，不管女性是 22-26 岁时怀孕还是在 26-27、27-30、30-35 岁下怀孕，对婴儿出生体重的影响是差不多的，即建议母亲怀孕年龄在 22-35 岁。从表 23 可以看出：35-40、40-45 岁下的因果效应两两比较值均是在上升的，均是在 35 岁这个年龄点下上升的，这也是符合常理的，在 35 岁时母亲身体机能开始走下坡路，容易导致一系列的并发症和合并症，一旦有合并症了，就会对胎儿产生一定的影响。一般来说，35 岁之后怀孕的产妇被定义为高龄产妇，因为在 35 岁之后一直到 45 岁时，母亲怀孕年龄对婴儿出生体重的因果效应又开始逐渐上升，因为母亲的身体机能随年龄的上升而下降，建议此时母亲不应再怀孕，否则婴儿在体内的营养得不到充分供给，容易引发各种疾病危害婴儿的健康。

综上所述：女生在 22-35 岁时怀孕，对婴儿体重的影响较小。故在 22-35 岁这个年龄段中怀孕对婴儿出生体重的影响较小，有利于婴儿出生体重保持在正常值范围，在一定程度上可以避免过早怀孕或是过晚怀孕，使得婴儿尽可能地避免面临早产、畸形、夭折等情况。

第六章 总结

本文中，我们主要研究高维协变量下的连续型处理变量对连续型响应变量的因果效应分析。区别于传统的因果效应分析，它针对二维处理变量和二维响应变量，传统的因果效应分析将不再适用于连续型处理变量以及连续型响应变量，Imbens (2000)提出了 GPS 的概念，为后续基于广义倾向得分进行因果效应分析奠定了基础，他指出倾向得分$r(t,X)$中的t值可能不再是 0 或 1 两个值，可能是三个或者三个以上的值，二维处理变量下的倾向得分不再适用，他指出如果t有 K+1 个值$t\in\{0,1,\cdots,K\}$，为了保证处理变量和多维协变量的条件独立性，我们需要对整个集合中 K +1 个值进行条件处理，并基于二维处理变量下所定义的倾向得分对多维处理变量下的倾向得分进行了定义，此时的倾向得分被称为广义倾向得分，针对广义倾向得分进行因果效应估计的模型就是离散型多处理因果模型。然而，当处理变量为连续型变量时，离散型多处理模型不再适用，故本文在 Imbens (2000)定义的广义倾向得分的基础上定义了连续型处理变量下的广义倾向得分函数。查阅相关文献发现，连续型处理变量下的因果效应大都是假设广义倾向得分函数服从正态分布等，本文将使用非参数的方法估计广义倾向得分函数。

随着计算机的普遍适用，获取的数据也将变得方便起来，但是在海量的数据中找到与研究对象相关的变量也变复杂起来，如何将高维数据变为低维的有效数据面临着极大的挑战，传统的降维方法可能不再适用，(Zhu and Zhu, 2007)指出充分降维可以有效的解决相关“维数灾难”问题，故本文采用充分降维的方法将高维的协变量降维成低维协变量，为后续基于非参数方法估计倾向得分函数奠定了基础。最后再基于倾向得分函数进行逆概率加权来估计平均因果效应。本文构建的连续型处理变量对连续型响应变量的因果模型具有以下创新点：(1)本文在生成数据模拟部分，考虑了不同分布和不同模型下的处理变量以及协变量，以及不同的结构维数对后续因果效应分析的影响，从而拓展了因果效应的应用。(2)本文研究的模型为连续型处理变量对连续型响应变量下的平均因果效应模型，在该模型的基础上对连续型处理变量进行切片，将之变成离散型处理变量，从而简化计算。(3)本文不仅考虑到连续型处理变量的问题，还考虑到了高维协变量的问题，利用协变量和处理变量之间的关系进行降维处理，较为完整的解决了高维数据下变量选择问题，为后续的因果效应分析奠定了基础。(4)本文基于非参数方法对广义倾向得分函数进行估计，将非参数方法应用于平均因果效应分析，从而有效的解决因果推断中参数估计的问题，使得因果效应的应用更广泛。(5)本文利用 SIR 算法将因果效应应用在多处理变量以及连续型响应变量下，区别于传统的因果效应分析针对二维处理变量和二维响应变量，从而拓展了经典因果效应。

最后，本文采用(Cattaneo, 2010)数据，研究母亲的年龄对新生儿体重的因果效应，得出了不同年龄段对新生婴儿体重的因果效应不同，母亲怀孕年龄在 13-22 岁时，对婴儿体重的影响较大；母亲怀孕年龄在 22-26 岁时的因果效应与母亲怀孕年龄在 26-27、27-30、30-35 岁时的因果效应相差较小，即母亲在 22-26 岁和 26-27、27-30、30-35 岁下怀孕对婴儿体重的影响的程度相差不大，且有着较好的身体机能，故在 22-35 岁这个年龄段怀孕为最佳年龄段；35 岁随着年龄的增长母亲怀孕年龄对婴儿体重的影响效果逐渐增大。故建议母亲在 22-35 岁这个年龄段怀孕，相对来说 22-35 岁这个年龄段对婴儿出生体重的影响较小较稳定，在一定程度上可以避免女性过早怀孕或是过晚怀孕，使得婴儿尽可能地避免面临早产、畸形、夭折等情况。

另外，本文的不足之处在于：在使用方法上，仅考虑了一种充分降维算法和一种非参数方法来估计平均因果效应。接下来，我们可以考虑在不同方法充分降维算法和不同的非参数方法下的因果效应，这是日后我们继续研究的方向。

参考文献

[1] 耿直. 因果推断与 Simpson 悖论[J]. 统计与信息论坛, 2000, 15(3): 9-12.

[2] 黄薇, 王惠文, 张志慧. 分段逆回归与神经网络组合建模方法[J]. 系统工程, 2004, 22(4):104-107.

[3] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[4] 李向杰, 吴燕燕, 张景肖. 基于切片逆回归的稳健降维方法[J]. 统计研究, 2018, 7.

[5] 李岩岩. SIR 降维方法与半参数可加回归的应用研究[D]. 重庆工商大学, 2016.

[6] 刘金灵. 多响应充分降维方法的改进[D].云南财经大学, 2021.

[7] 楼芝兰. 个性化医疗中多种治疗方案下的最优分配准则估计[D]. 华东师范大学, 2018.

[8] 肖招娣. 高维数据集上的降维算法及其应用[D]. 华南理工大学, 2013.

[9] 谢志超. 基于 SIR 的数据降维算法研究及其应用[D]. 南京邮电大学, 2018.

[10] 张浩, 郝志峰, 蔡瑞初, 等. 基于互信息的适用于高维数据的因果推断算法[J]. 计算机应用研究, 2015, 32(2): 382-385.

[11] 周天枢, 陈崇帼. 指数分布及其在寿命表中的应用[J]. 中国卫生统计, 1992(4): 62-63.

[12] 周文琴, 冯鸣鸣, 王惠文. SIR 方法在小型二次电池市场分析上的应用[J]. 数理统计与管理, 2001, 20(6): 5-9.

[13] Aerts M, Claeskens C. Local Polynomial Estimation in Multiparameter Likelihood Models[J]. Journal of the American Association, 1997, 440(92): 1536-1545.

[14] Almond D, Chay K Y, Lee D S. The costs of low birth weight[J]. The Quarterly Journal of Economics, 2005, 120(3): 1031-1083.

[15] Austin P C. Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures[J]. Statistical methods in medical research, 2019, 28(5): 1365-1377.

[16] Cai R, Zhang Z, Hao Z. Sada: A general framework to support robust causation discovery[C]. International conference on machine learning. PMLR, 2013: 208-216.

[17] Cai Z R, Li R Z, Zhu L P. Online Sufficient Dimension Reduction Through Sliced Inverse Regression[J]. Journal of Machine Learning Research, 2020, 21:1-25.

[18] Chen C H, Li K C, Wang J L. Dimension reduction for censored regression data[J]. The Annals of Statistics, 1999, 27(1): 1-23.

[19] Cheng D B, Li J Y, Liu L, et al. Sufficient dimension reduction for average causal effect Estimation[J]. Data Mining and Knowledge Discovery, 2022, 36:1174–1196.

[20] Clarke B, Fokoue E, Zhang H H. Principles and Theory for Data Mining and Machine Learning[M]. Springer, 2009.

[21] Cook R D, Weisberg S. Sliced inverse regression for dimension reduction: Comment[J]. Journal of the American Statistical Association, 1991, 86(414): 328-332.

[22] Cook R D. Regression Graphics: Ideas for Studying Regressions Through Graphics. New York, NY: Wiley, 1998.

[23] Dong Y X. A brief review of linear sufficient dimension reduction through optimization[J]. Journal of Statistical Planning and Inference, 2021, 211: 154–161.

[24] Feng P, Zhou X H, Zou Q M, et al. Generalized propensity score for estimating the average treatment effect of multiple treatments[J]. Statistics in medicine, 2012, 31(7): 681-697.

[25] Fisher R A. Design of experiments[J]. British Medical Journal, 1936, 1(3923): 554.

[26] Fan J Q, Gijbels I. and King M. Local likelihood and local partial likelihood in hazard regression[J]. Ann. Statist., 1997, 25, 1661-1690.

[27] Geng Z, Wang C, Zhao Q. Decomposition of search for v-structures in DAGs[J]. Journal of Multivariate Analysis, 2005, 96(2): 282-294.

[28] Garès V, Chauvet G, Hajage D. Variance estimators for weighted and stratified linear dose–response function estimators using generalized propensity score[J]. Biometrical Journal, 2022, 64(1): 33-56.

[29] Hall P, Li K C. On almost linearity of low dimensional projections from high dimensional data[J]. The annals of Statistics, 1993: 867-889.

[30] Hino H, Wakayama K, Murata N. Entropy-based sliced inverse regression[J]. Computational statistics & data analysis, 2013, 67: 105-114.

[31] Horvitz D. G., Thompson D. J. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1952, 47 (260): 663–685.

[32] Hsing T, Carroll R J. An asymptotic theory for sliced inverse regression[J]. The Annals of Statistics, 1992: 1040-1061.

[33] Hu S, Chen Z, Partovi Nia V, et al. Causal inference and mechanism clustering of a mixture of additive noise models[J]. Advances in Neural Information Processing Systems, 2018, 31.

[34] Imbens G W. The role of the propensity score in estimating dose-response functions[J]. Biometrika, 2000, 87(3): 706-710.

[35] Janzing D, Steudel B, Shajarisales N, et al. Justifying information-geometric causal inference[M]. Measures of complexity. Springer, Cham, 2015: 253-265.

[36] Janzing D, Mooij J, Zhang K, et al. Information-geometric approach to inferring causal directions[J]. Artificial Intelligence, 2012, 182: 1-31.

[37] Kurthen M, Enßlin T A. A Bayesian Model for Bivariate Causal Inference[J]. Entropy, 2019, 22(1): 46.

[38] Li B, Wang S. On directional regression for dimension reduction[J]. Journal of the American Statistical Association, 2007, 102(479): 997-1008.

[39] Li K C. Sliced inverse regression for dimension reduction[J]. Journal of the American Statistical Association, 1991, 86(414): 316-327.

[40] Li K C. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma[J]. Journal of the American Statistical Association, 1992, 87(420): 1025-1039.

[41] Li K C, Aragon Y, Shedden K, et al. Dimension reduction for multivariate response data[J]. Journal of the American Statistical Association, 2003, 98(461): 99-109.

[42] Li L, Cook R D, Tsai C L. Partial inverse regression[J]. Biometrika, 2007, 94(3): 615-625.

[43] Li L, Nachtsheim C J. Sparse sliced inverse regression[J]. Technometrics, 2006, 48(4): 503-510.

[44] Li L, Wen X M, Yu Z. A selective overview of sparse sufficient dimension reduction[J]. statistical theory and related fields, 2020, 4(2): 121–133.

[45] Liquet, B. and Saracco, J. A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. Computational Statistics, 2012, 27(1), 103-125.

[46] Liu J, Ma Y, Wang L. An alternative robust estimator of average treatment effect in causal inference[J]. Biometrics, 2018, 74(3): 910-923.

[47] Neyman J S. On the application of probability theory to agricultural experiments. essay on principles. section 9[J]. Annals of Agricultural Sciences, 1923, 10: 1-51.

[48] Park J H, Sriram T N, Yin X. Central mean subspace in time series[J]. Journal of Computational and Graphical Statistics, 2009, 18(3): 717-730.

[49] Park J H, Sriram T N, Yin X. Dimension reduction in time series[J]. Statistica Sinica, 2010: 747-770.

[50] Pearson K. On the probability that two independent distributions of frequency are really samples from the same population[J]. Biometrika, 1911, 8(1-2): 250-254.

[51] Pearl J. Causal diagrams for empirical research[J]. Biometrika, 1995, 82(4): 669-688.

[52] Pearl J. Models, reasoning and inference[J]. Cambridge, UK: CambridgeUniversityPress, 2000, 19(2).

[53] Pearl J, Glymour M, Jewell N P. Causal Inference in Statistics:A Primer[M], 杨娇云,等,译. 北京: 高等教育出版社, 2020.

[54] Rajaraman A, Ullman J D. Mining of massive datasets[M]. Cambridge University Press, 2011.

[55] Rosenbaum P R, Rubin D B. The central role of the propensity score in observational studies for causal effects[J]. Biometrika, 1983, 70(1): 41-55.

[56] Rubin D B. Causal inference using potential outcomes: Design, modeling, decisions[J]. Journal

of the American Statistical Association, 2005, 100(469): 322-331.

[57] Rubin D B. Inference and missing data[J]. Biometrika, 1976, 63(3): 581-592.

[58] Rubin D B. Comment: Which ifs have causal answers[J]. Journal of the American statistical association, 1986, 81(396): 961-962.

[59] Rubin D B. Multivariate matching methods that are equal percent bias reducing, I: Some examples[J]. Biometrics, 1976: 109-120.

[60] Tian T S. Dimensionality reduction for classification with high-dimensional data[M]. University of Southern California, 2009.

[61] Tsamardinos I, Brown L E, Aliferis C F. The max-min hill-climbing Bayesian network structure learning algorithm[J]. Machine learning, 2006, 65(1): 31-78.

[62] Tu C, Koh W Y, Jiao S. Using generalized doubly robust estimator to estimate average treatment effects of multiple treatments in observational studies[J]. Journal of Statistical Computation and Simulation, 2013, 83(8): 1518-1526.

[63] Tu C, Koh W Y. Causal inference for average treatment effects of multiple treatments with non-normally distributed outcome variables[J]. Journal of Statistical Computation and Simulation, 2016, 86(5): 855-861.

[64] Wright S. The treatment of reciprocal interaction, with or without lag, in path analysis[J]. Biometrics, 1960, 16(3): 423-445.

[65] Xie X, Geng Z, Zhao Q. Decomposition of structural learning about directed acyclic graphs[J]. Artificial Intelligence, 2006, 170(4-5): 422-439.

[66] Yu K, Jones M C. Likelihood-based local linear estimation of the conditional variance function[J]. Journal of the American Statistical Association, 2004, 99(465): 139-144.

[67] Zhu L P, Zhu L X. A data adaptive hybrid method for dimension reduction[J]. Journal of Nonparametric Statistics, 2007.

[68] Zhu L X, Miao B Q, Peng H. Sliced Inverse Regression with Large Dimensional Covariates[J]. Journal of the American Statistical Association, 2006, 101: 630-643.

Zhejiang University of Finance and Economics

浙 江 财 经 大 学

地 址：中国杭州市下沙高教园区学源街 18 号

邮 编：310018

电 话：0571-86754515

网 址：https://gs.zufe.edu.cn