因果约束的稳定学习理论方法研究

（申请清华大学工学博士学位论文）

培 养 单 位：计 算 机 科 学 与 技 术 系
学　　科：计 算 机 科 学 与 技 术
研　究　生：况　琨
指 导 教 师：杨　士　强　教　授

二〇一九年六月

Causally Regularized Stable Learning: Theorem and Method

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Computer Science and Technology

by

Kun Kuang

Dissertation Supervisor : Professor Shiqiang Yang

June, 2019

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

随着大数据时代的兴起，海量数据在各行各业都被充分利用起来。通过大数据赋能机器学习，现有预测模型的预测力得到了显著提升，并成功运用于各个领域来解决工业界实际问题。然而，现有大部分预测模型是数据关联驱动的算法，且需要假设测试数据集和模型训练的数据集是独立同分布的。其中关联驱动学习使得现有预测模型会基于虚假关联进行预测，导致预测缺乏可解释性；而独立同分布假设会使得预测模型对未知分布的测试数据集预测能力失去保障，导致预测缺乏稳定性。预测模型缺乏可解释性和稳定性，会使得用户对其结果产生不信任，进而导致其很难应用于很多关键领域，尤其是那些需要决策制定的领域，例如金融，医疗，军事和政治等。因此如何提升预测模型的可解释性和稳定性对学术研究和实际应用都十分重要。

为了提升预测模型的可解释性和稳定性，本文提出了因果约束的稳定学习框架，通过大数据因果推理方法挖掘数据之间不变的因果关系，并利用因果关系约束预测模型的训练学习，实现对未知分布的测试数据集的稳定预测。本文主要研究内容和创新成果如下：

- **面向高维数据，提出了混淆变量自动分离的因果推理模型**：在传统因果推理框架中，所有的观测变量都被视为会影响因果推理的混淆变量，导致其很难运用于大数据高维变量情况。针对该问题，本文提出了面向高维数据的因果推理框架，将所有的观测变量分为三部分：混淆变量，调整变量和无关变量。基于提出的因果推理框架，本文通过构造混淆变量和调整变量之间的正交性约束项，提出了数据驱动的变量自动分离和因果推理算法。其中分离的混淆变量可以帮助无偏地评估因果效应，而调整变量通过回归可以帮助降低因果效应评估的方差。在理论和实验上，我们都验证了该模型在面向高维数据时因果推理的准确性和鲁棒性。
- **面向变量差异性，提出了混淆变量区分性平衡的因果推理模型**：本文从理论上论证了，在观测性学习中不同的混淆变量给因果推理带来的误差影响是不一样的。在此基础上，本文通过同时学习混淆变量权重和样本权重，提出了混淆变量区分性平衡的因果推理算法。其中混淆变量权重用于选择混淆变量和区分混淆变量的影响，而样本权重则用于平衡混淆变量分布来消除其对因果推理的影响。通过在仿真数据和真实数据上的大量实验，我们验证了该模型在面向混淆变量差异性时能准确且稳定地评估因果效应。

- **面向未知分布的测试数据，提出了因果约束的稳定学习理论方法**：为了解决预测问题中存在的变量虚假相关性，数据分布差异性和测试数据未知性等挑战，本文提出了因果约束的稳定学习理论方法框架。具体地，基于传统的逻辑回归模型，本文通过联合优化深度自动编码模型和变量全局平衡模型提出了一种深度全局平衡的回归模型，实现面向未知分布的测试数据的稳定预测。其中，变量全局平衡模型用来挖掘变量之间的因果不变关系，而深度自动编码模型用于学习变量的低维度非线性表征。在理论和实验上，我们都验证了该模型能实现面向未知分布测试数据的稳定预测。

本文提出的大数据因果推理框架为因果约束的稳定学习奠定了一定的理论基础；而提出的因果约束的稳定学习理论方法，初步解决了现有预测模型存在的不可解释性和预测不稳定性等问题，为可解释性和稳定性人工智能提供了研究思路。

关键词：因果推理；稳定学习；因果约束学习；可解释性；预测稳定性

Abstract

Owing to the era of Big Data, abundant data are accumulated and applied in various domains. By empowering machine learning with big data, the predictive power of existing predictive methods has been significantly improved, and these methods have been proved to be successful to solve practical problems in many real industrial applications. However, most of the existing predictive models are correlation-based, and their success is based on the I.I.D. hypothesis that the testing dataset and training dataset are independently and identically distributed. Correlation-based predictive models might learn the spurious correlation among variables for prediction, which leads to the lack of interpretability of the prediction. The I.I.D. hypothesis makes predictive models lose the guarantee of predictive performance on the testing data from an unknown distribution, resulting in the lack of stability of the prediction. The lack of interpretability and stability of predictive models will hardly convince users to trust them and their results, making them difficult to apply in many key areas, especially those that require decision making, such as finance, medical, military, and political. Therefore, how to improve the interpretability and stability of the predictive model is of paramount importance for both academic research and real applications.

In order to improve the interpretability and stability of the predictive model, this paper proposes a causally regularized stable learning framework, where we firstly develop a new causal inference framework under big data era, aiming to mine the invariant causal relationship between variables; then, we adopt the causal relationship to regularize predictive model learning to achieve stable and explainable prediction on testing datasets form unknown distributions. The main contributions of the thesis are as follows:

- **Data-Driven Variables Decomposition Algorithm for Causal Inference with High-dimensional Data:** In the traditional causal framework, all observed variables are regarded as confounders that affect causal effect estimation, making it difficult to apply to high-dimensional causal inference. To address this problem, this paper proposes a novel causal framework, which divides all observed variables into three parts: confounders, adjustment variables, and irrelevant variables. Based on the proposed causal framework, this paper proposes a Data-Driven Variables Decomposition for high-dimensional causal inference, where a combined orthog-

onality and sparsity regularizer is constructed to simultaneously 1) separate the confounders and adjustment variables with a data-driven approach, 2) eliminate irrelevant variables which are neither confounders nor adjustment variables to avoid overfitting, and 3) estimate the causal effect in observational studies. During estimating the causal effect, the separated confounders can effectively eliminate their confounding impact, while the adjustment variables can significantly reduce the variances of estimated causal effect through outcome adjustment. Theoretically and experimentally, we have verified the accuracy and robustness of the proposed method for high dimensional causal inference.

- **Differentiated Confounder Balancing Algorithm for Causal Inference in the Wild:** In the wild big data scenario, there are always a large number of observed variables. However, this paper theoretically proved that not all observed variables are confounders and different confounders contribute unequally to the confounding bias in data. Motivated by this, this paper proposes a Differentiated Confounder Balancing algorithm by simultaneously optimizing sample weights and confounder weights to jointly select confounders, differentiate weights of confounders and balance confounder distributions for causal effect estimation. During estimating the causal effect, the confounder weights can determine which variable is confounder and its share of contribution on confounding bias, and the sample weights are designed for confounder balancing. With extensive experiments on both synthetic and real datasets, we validate that the proposed algorithm outperforms the state-of-the-art methods on causal effect estimation in observational studies.
- **Causally Regularized Stable Learning for Stable Prediction across Unknown Testing Data:** To achieve interpretable and stable prediction, existing predictive models are still facing following challenges: spurious correlation, distribution shift among data, and unknown testing data. To address these challenges, this paper proposes a causally regularized stable learning theory framework and algorithm. Specifically, based on the traditional logistic regression model, this paper proposes a Deep Global Balancing Regression model by jointly optimizing the deep auto-encoder model and the variable global balancing model to achieve stable prediction across unknown testing data. The global balancing model constructs balancing weights that help to identify stable, causal relationships between features and outcomes, while the deep auto-encoder model is designed to reduce the dimensionality

of the feature space and learn nonlinear representation, thus making global balancing easier. We show, both theoretically and with empirical experiments, that the proposed algorithm can make stable predictions across unknown testing data.

In this paper, the proposed causal framework and algorithms for causal inference in the big data era lays a certain theoretical foundation for causally regularized stable learning. The proposed causality regularized stable learning theory framework and algorithm significantly improved the interpretability and stability of existing prediction models, which provides a possible research way for explainable and stable artificial intelligence.

Key words: Causal Inference; Stable Learning; Causality Regularized Learning; Interpretability; Stability

目 录

主要符号对照表

D^2VD	数据驱动的变量分离算法（Data-Driven Variable Decomposition）
DCB	混淆变量区分性平衡算法（Differentiated Confounder Balancing）
GBR	全局平衡回归算法（Global Balancing Regression）
DGBR	深度全局平衡回归算法（Deep Global Balancing Regression）
LASSO	最小绝对收缩和选择算子，套索算法（Least Absolute Shrinkage and Selection Operator）
RMSE	均方根误差（Root Mean Square Error）
MAE	平均绝对误差（Mean Absolute Error）
Bias	误差
SD	标准方差 (Standard Deviation）

第 1 章　引言

大数据时代，海量数据在各行各业都被充分利用起来。在数据挖掘和机器学习领域，基于海量数据的分析和建模，很多推理预测模型被用来解决工业界实际问题。通过大数据赋能机器学习，现有预测模型的准确率在各个领域都得到了很好的验证。然而当今推理预测模型大部分是数据关联驱动的，而且需要假设模型的训练数据集和测试数据集是独立同分布的。其中，关联驱动学习会使得现有预测模型误用数据中的虚假关联进行预测，导致预测缺乏可解释性；而独立同分布假设会使得现有预测模型对未知分布的测试数据集的预测能力下降，导致预测缺乏稳定性。

为了解决现有预测模型存在的不可解释和不稳定问题，本文提出一种因果约束的稳定学习方法，通过因果推理方法挖掘观测数据中变量之间的因果关系，并利用因果关系来约束预测模型的学习，最终实现对未知测试数据集的稳定预测。其中因果推理是实现可解释性分析和稳定学习的强大统计建模工具。然而大数据背景下，观测性研究中因果推理目前还存在以下挑战：一是高维变量和噪声变量；二是变量的差异性。为了实现大数据因果推理，本文提出了面向高维变量以及面向变量差异性的因果推理方法。为了实现稳定预测，本文提出利用因果知识来识别预测变量中的因果变量，并学习因果变量与结果变量之间的稳定因果关系来训练预测模型。本文针对因果约束的稳定学习算法给出了理论保证，并在仿真数据和真实数据上都验证了我们方法对未知测试数据的预测稳定性和可解释性。本章的后续部分将会介绍研究背景，简要回顾因果推理和稳定学习的相关研究，提出本文的研究问题及其所面临的挑战，阐述相应的研究贡献。

1.1　研究背景

早在 2008 年 9 月，《自然》（Nature）期刊就推出了《大数据》（Big Data）[1] 专刊，“大数据”被作为专题封面，标志着大数据时代的到来。在 2011 年 2 月，《科学》（Science）期刊推出《数据数量》（Dealing with Data）[2] 专刊，主要围绕科学研究和实际工业应用中的大数据问题展开探索和讨论，体现了大数据对于科学研究和工业应用的重要性。

大数据时代，海量数据在各行各业都被充分利用起来，大数据技术也被广泛运用于各行各业[3]。在宏观经济方面，IBM 日本分公司基于从互联网新闻中搜集影响制造业的 480 项经济大数据，建立了经济指标预测系统，实现对采购经纪人

的指数预测[①]。通过谷歌提供的心情分析工具，印第安纳大学从海量网民留言中总结出六种心情，并利用其对琼斯工业指数进行预测，预测准确率高达 87%。在金融领域，通过分析 3.4 亿调微博用户的留言数据，德温特资本市场公司提出民众情绪预测模型，挖掘出民众买卖股票规律，并基于此来决定公司股票的买入和卖出。阿里公司基于淘宝网上大量中小型企业的交易记录，预测其财务健康和诚信评分，对评分较高的企业发放无需担保的贷款，目前已放贷 300 多亿人民币，而坏账率只有 0.3%。在商业领域，沃尔玛公司通过历年销售数据，挖掘适合搭配一起出售的商品；分析顾客的购物行为，为顾客提供个性化商品推荐服务。在农业领域，硅谷某气候公司利用气象局几十年的天气数据，分析降雨量，气温，土壤和往年农作物产量的相关性，并基于此来预测农场来年的产量，向农户推荐个性化的农业保险。在医疗保健领域，“谷歌流感趋势”项目基于网上搜索内容的海量数据分析流感等疾病在全球传播的状况，与真实报告相比，其追踪疾病的精准率高达 97%。在安保领域，基于海量人脸数据学习出来的人脸识别系统被用来监控人口流动，识别在逃犯罪分子等[②]。

通过大数据赋能机器学习，现有预测模型的准确率在各个领域都得到了很好的验证。大部分现有预测模型存在以下主要特征：第一，预测模型是基于数据关联关系学习的；第二是预测模型需要假设测试数据集与训练数据集是独立同分布的。其中，关联驱动模型学习会让现有预测模型缺乏可解释性。从文章[4] 中，我们知道变量之间的关联主要有三个来源：一是因果关系（Causation），例如下雨天会导致地面湿，因此雨天跟地面湿因果相关；二是混淆关联（Confounding），例如下雨天同样会导致行人撑伞，我们在数据中会观测到撑伞与地面湿相关；第三来源是样本的内生性选择（Endogenous Selection），例如在图片识别的应用中，如果选取的数据集中 80% 的狗都在草地上，则会使得狗跟草地非常相关。这里，我们将非因果关联（混淆关联和内生性选择关联）称之为虚假关联（Spurious Correlation），因为虚假关联往往是不可解释的，例如撑伞并不能用来解释地面湿，草地也并不能用来解释狗在图片中的存在。在很多实际应用中，我们观测到的关联大部分是虚假的，导致基于数据关联驱动的预测模型缺乏可解释性。预测模型缺乏可解释性，会使得用户对其预测结果产生不信任，进而导致其在很多关键领域（例如医疗，金融，军事和政治等）都难以运用。另外，测试数据集与训练数据集独立同分布的假设是使得现有预测模型缺乏预测稳定性的主要原因。在很多实际应用中，一些自然因素（例如时间，地域等）会导致不同数据集之间的分布不一致问题，从而违反独立同分布假设。例如夏天的狗常出现在草地上，而冬天的狗则常出现在

① http://finance.sina.com.cn/roll/20120621/135712374189.shtml

② http://tech.163.com/18/0613/08/DK5TF8JP00097U80.html

雪地里。而且在实际应用中，需要预测的目标数据在预测模型训练阶段往往是未知的，导致测试数据集分布的未知性，进而导致现有预测模型的预测不稳定性。例如，基于大部分狗都在草地上的训练数据集，通过关联驱动学习的现有预测模型，会使得草地成为识别狗的重要特征。那么在预测阶段，如果未来的测试数据集也是狗在草地上，现有模型则能很好的预测；但是如果未来的测试数据集是大多数狗在雪地里，则现有预测模型不能很好的预测。预测模型缺乏预测稳定性，会使得模型对未来数据预测的结果缺乏保证，同样失去用户的信任，进而导致其很难运用于很多关键领域。因此，模型的预测稳定性和可解释性是让人类更加信任预测模型的前提。

近年来，研究者们提出可以通过大数据因果推理来提升现有预测模型的可解释性。因果推理是指基于效应发生的条件推出关于因果关系结论的过程，是用于可解释性分析的强大统计建模工具。实际上，因果推理已成为后深度学习时代的聚焦点。在 2016 年，美国科学院组织了大数据因果推理研讨会①，邀请了学术界和工业界众多科学家一起研讨如何实现大数据中的因果推理（Drawing Causal Inference from Big Data）[5]。2017 年，美国国防高级研究计划局也将因果可解释性人工智能列入 5 年研究计划②，突出了可解释人工智能（Explainable Artificial Intelligence）[6]的重要性。因果推理在给预测模型带来可解释性的同时，由于因果关系和因果知识的不变性，同时能赋予预测模型对结果预测的稳定性。因此，因果推理是实现可解释预测和稳定学习的重要基础。

在本文中，我们旨在现有预测模型的基础上，研究大数据因果推理问题，通过引入因果关系来约束预测模型的学习和训练，赋予预测模型一定的可解释性和预测稳定性，实现预测模型对于未知环境数据的稳定预测。该课题具有很强的现实意义和研究价值：首先模型的可解释性可以使得人类更好地理解预测模型，掌握模型性能的好坏情况及原因，为人类决策提供宝贵建议。而稳定性保障了预测模型对未知数据预测结果的准确性，使得预测模型能广泛地运用于很多关键领域。总的来说，可解释性和稳定性都可以帮助建立人类与预测模型之间的信任，促进人工智能领域的进一步发展。

1.2　研究主要挑战

为了实现因果约束的稳定学习与推理，本文首先研究大数据背景下因果推理问题。实现因果推理的最佳方法是进行随机对照试验，例如 A/B testing。然而在实

① http://www.nasonline.org/programs/sackler-colloquia/completed_colloquia/Big-data.html

② https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

际应用中，完全的随机对照试验价格往往非常昂贵，且由于伦理道德等因素，随机对照试验在大部分问题中是不可行的。因此，目前很多因果推理研究工作主要是基于历史观测数据进行的。大数据给观测性研究中因果推理带来了全新的机会：大量观测数据使得因果推理更精准，高维观测变量使得因果推理假设更合理。与此同时，也给因果推理带来了更大的挑战，主要包括：

（一）**高维变量和噪声变量的挑战。**在大数据应用中，我们观测到的变量成千上万，但并不是所有的观测变量都是影响因果推理的混淆变量，还包含噪声变量和其它一些有用变量，例如调整变量，其可以被用来降低因果推理的方差。然而，传统的方法主要集中在低维数据上的因果推理和效应评估，简单地将所有的观测变量都当做混淆变量。因此传统方法很难运用于大数据高维变量下的因果推理。面向高维变量数据，如何自动分离混淆变量进行因果分析是本文在大数据因果推理方面需要解决的第一个挑战。

（二）**混淆变量的差异性和变量结构的未知性的挑战。**在真实大数据场景中，除了少量的确定性变量之外几乎总是存在大量额外或者不受控制的变量，并且它们之间的结构关系在现实世界中是复杂而未知的。因此在大数据因果推理过程中，我们不能像传统方法一样对变量之间的结构进行模型假设。另外，在消除混淆变量对因果推理的影响时，不同的混淆变量对推理结果带来的偏差也是不一样的，存在明显的差异性。然而，传统方法在消除混淆变量影响时，将所有的混淆变量都同等对待，没有进行区分，导致推理的不准确性。面向变量差异性和结构未知性，如何设计无参模型实现混淆变量区分性平衡的因果推理是本文在大数据因果推理方面需要解决的第二个挑战。

基于大数据因果推理，我们通过引入因果关系来约束预测模型的训练，消除虚假关联预测，实现因果稳定预测。在实现稳定预测这块工作中，我们主要面临以下三个方面的问题：

（1）数据中的非因果（虚假）相关性。在观测数据中，变量与变量之间的关联主要有三个来源：因果关联，混淆关联和样本内生性选择关联。后两者属于非因果（虚假）关联，例如图片数据中大部分狗在草地上会使得非因果特征草地跟狗显著相关，导致非因果特征（例如草地）被预测模型学为重要的预测特征，导致模型缺乏可解释性。

（2）数据分布的差异性。在很多机器学习问题中，测试数据集都被假设成跟训练数据集是独立同分布的。然而，在实际应用中，不同数据集的分布是存在显著差异的，例如不同省份的人口特征分布，不同医院的病人数据分布等。数据分布的差异性会导致数据中的虚假关联在不同的数据中发生变化，从而导致关联驱动

的预测模型对分布不一的测试数据预测效果不稳定。

（3）测试数据的未知性。在实际应用中，预测模型通常被用来预测未来的数据结果，也就是说测试数据集在模型学习训练阶段往往是未知的。测试数据的未知性导致传统的迁移学习方法很难运用于稳定预测。为了实现稳定预测，我们需要挖掘数据中的不变性，也就是变量之间的因果关系。

如何利用大数据推理技术消除观测数据中的虚假相关，并挖掘不随数据变化的因果关联来约束预测模型的学习，最终实现稳定推理是本文需要解决的第三个挑战。

1.3　本文工作与贡献

针对以上挑战，本文提出了大数据因果推理和因果约束的稳定学习理论方法。其中，我们提出的大数据因果推理模型能基于观测的大数据准确有效地评估变量之间的因果效应；我们提出的因果约束的稳定学习推理模型能够通过挖掘变量之间不变的因果关系实现在未知测试数据上准确且稳定的预测。本文主要研究内容和创新成果如下：

面向高维数据，提出了数据驱动的混淆变量自动分离的因果推理模型：相比于传统因果推理框架，将所有的观测变量都当做混淆变量，本文提出了一个全新的面向高维数据的因果推理框架。在我们的因果推理框架中，将所有的观测变量分为三部分：混淆变量，调整变量和无关变量。基于我们的因果推理框架，我们提出了数据驱动的变量自动分离和因果推理算法（Data-Driven Variable Decomposition, D^2VD）。在我们的算法中，通过构造正交性正则项来自动分离混淆变量和调整变量，其中分离的混淆变量可以帮助我们无偏地评估因果效应，而调整变量通过简单回归帮助我们降低因果效应评估的方差；通过稀疏性正则项来消除无关变量的影响，同时防止因果推理时模型过拟合问题。在理论和实验上，我们都验证了该模型在面向高维数据时能准确地评估因果效应，且方差小。

面向混淆变量的差异性，提出了混淆变量区分性平衡的因果推理模型：为了解决混淆变量的差异性，本文提出了混淆变量区分性平衡和因果推理算法（Differentiated Confounder Balancing，DCB）。该算法是基于混淆变量平衡的推理框架，但与之前所有混淆变量平衡算法不同，我们认为不同的混淆变量给因果推理带来的影响是不一样的，且从理论上证明了这一点。在此基础上，我们提出通过同时学习混淆变量权重和样本权重来实现混淆变量区分性平衡的因果推理。其中混淆变量权重被用来选择混淆变量和区分混淆变量的影响，而样本权重则是平衡混淆变量分布来消除其对因果推理的影响。通过在仿真数据和真实数据上的大量

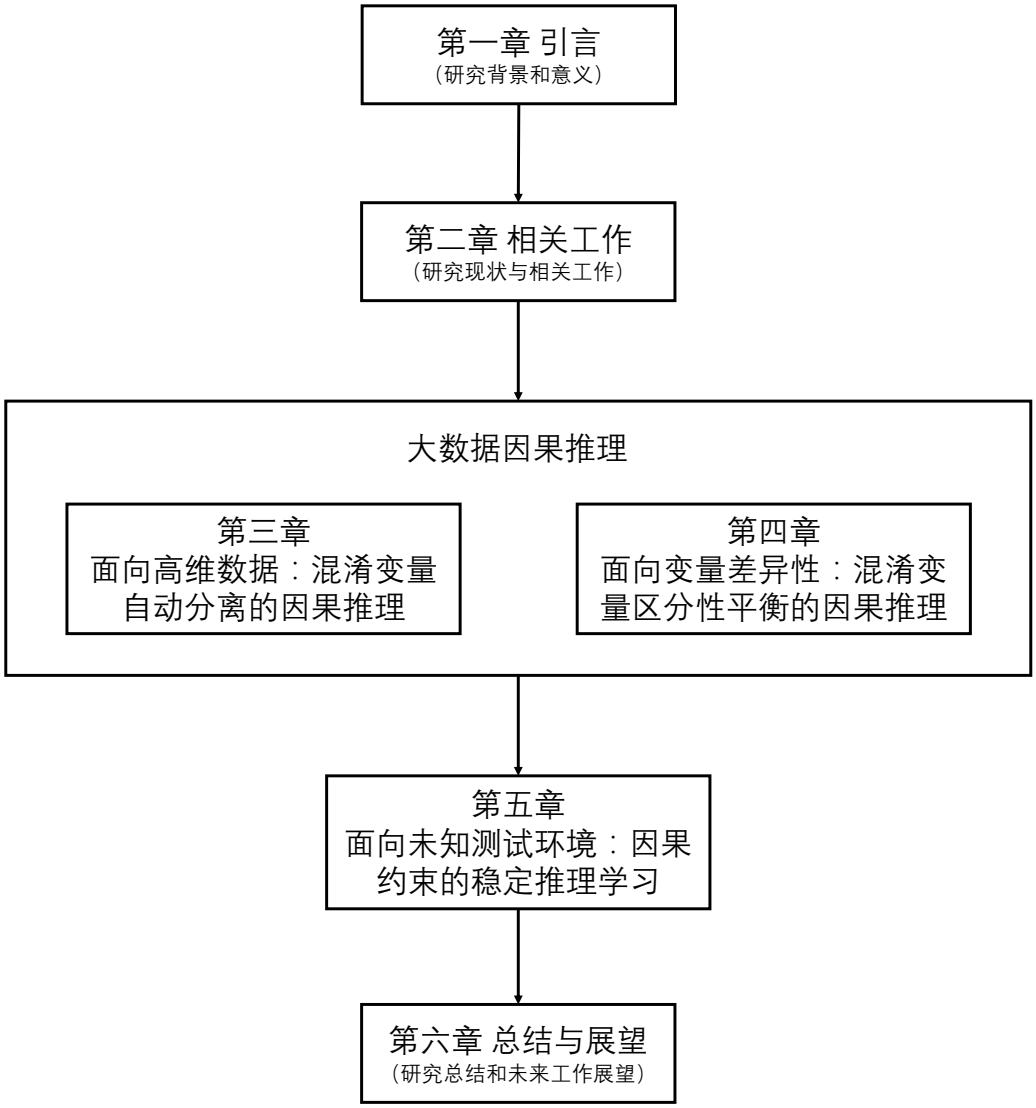


图 1.1　本文研究内容框架

实验，验证了该模型在面向混淆变量差异性时能准确且稳定地评估因果效应。

面向未知分布的测试数据，提出了因果约束的稳定学习理论方法：为了解决预测问题中存在的变量虚假相关性，数据分布的差异性和测试数据的未知性等挑战，本文提出了因果约束的稳定学习理论方法。具体地，基于传统的逻辑回归模型，提出了一种深度全局平衡回归算法（Deep Global Balanicng Regression，DGBR）。在我们算法中，通过联合优化深度自动编码模型和变量全局平衡模型，实现面向未知分布测试数据的稳定预测。其中，变量全局平衡模型用来识别和恢复预测变量和结果变量之间的因果关系，而深度自动编码模型设计用于学习变量的低纬度非线性表征，其可以使变量全局平衡更容易且噪声更小。在理论和实验上，我们都验证了该模型能实现面向未知分布测试数据的稳定预测。

1.4　本文组织结构

本文的组织结构图和各部分研究内容如图1.1所示。在第二章中，本文分别介绍因果推理，协变量迁移，迁移学习，稳定学习和可解释性学习这五个方面的国内外研究现状。为了实现因果约束的稳定学习，我们首先在第三、第四章研究了大数据中的因果推理模型。针对大数据因果推理存在的挑战，分别提出了相应的因果推理模型。在第三章中，针对高维数据带来的挑战：并不是所有的观测变量都是影响因果推理的混淆变量，我们提出了混淆变量自动分离的因果推理算法。在第四章中，针对混淆变量差异性带来的挑战：不同的混淆变量对因果推理的影响程度不一样，我们提出了混淆变量区分性平衡的因果推理算法。在第三、第四章中对大数据进行因果推理的基础上，我们在第五章将介绍如何利用因果知识对预测模型进行约束，实现面向未知分布测试数据的稳定推理和预测。在第六章中，本文将对以上工作进行全面总结，并进一步展望了未来可能的研究方向。

第 2 章　研究现状与相关工作

2.1　因果推理

在因果推理模型框架中，潜在结果模型框架（Potential Outcome Framework，也常称为 Neyman-Rubin's Framework）是其中最重要的模型之一。其核心任务是评估干预变量（Treatment）对结果变量产生的因果效应（Causal Effect，也常称为 Treatment Effect），因果效应指的是同一组样本（Unit）在接受干预（Treated）和不接受干预（Control）两种状态下的结果差异。这里我们称接受干预的样本组为干预组，而不接受干预的样本组称为对照组。潜在结果模型和方法已在各个领域都得到了广泛的应用，包括经济学[7]，流行病学[8]，社会科学[9–12]，政治学[13–15]，公共管理和政策[16–20]，广告营销[21–25] 等。

评估因果效应，最行之有效的方法是进行随机对照试验（Randomized Controlled Trial，RCT）。在随机对照试验中，接受干预和不接受干预两种状态会被随机分配给每一个样本，其中完全的随机性能保证最后接受干预的样本组与不接受干预的样本组中所有特征变量上的分布都是一致的，因此最后两组样本结果的差异性只来自于其是否接受干预，也就是干预变量对结果的因果效应。随机对照试验被广泛地运用于医学[26]，生物学[27] 和广告营销[28,29] 等。然而，在很多实际问题中，完全的随机对照试验的代价往往非常昂贵，而且在很多情况下，例如干预变量涉及到伦理道德问题，随机对照试验是不可行的。

因此，很多学者提出利用历史观测数据进行因果效应评估。基于观测数据评估因果效应的方法大致可以分为两类：基于倾向值的方法和变量直接平衡算法。

倾向值指的是在给定样本特征的条件下样本接受干预状态的概率，在 1983 年被 Rosenbaum 和 Rubin[30] 提出用来解决观测数据下因果效应评估问题。基于倾向值，很多因果效应评估的方法被提出用来解决因果效应评估问题。这些方法大致可以分为四类，包括倾向值匹配（Propensity Score Matching）、倾向值分层（Stratification on Propensity Score）、倾向值倒数加权（Inverse of Propensity Weighting）和双稳健（Doubly Robust）算法。

倾向值匹配算法[6,24,30–37] 进行因果效应评估时，需要形成具有相似倾向值分布的干预组和对照组。其中最常用的方法是样本一对一配对匹配，对于每个接受干预或不接受干预的样本，匹配一个与倾向值相似但干预状态相反的样本。一旦形成干预组和对照组之间的匹配样本，我们就可以通过比较两组样本的结果差异来评估干预变量的因果效应[30]。除了一对一匹配外，还有一对多匹配，多对一匹

配，有放回匹配和无放回匹配[6,32]。

倾向值分层算法[38–43] 引入了分层分析技术，首先基于评估的倾向值将样本进行排序，然后根据倾向值评分的分位数将所有样本划分为若干个层（一般分为 5-10 层）。分层后两组样本在各层的倾向值分布应该非常接近，则在每一层的数据可以被认为来自于近似随机对照试验。在每一层，因果效应可以通过直接比较两组样本结果的平均值得到[38]。

倾向值倒数加权算法[44–46] 利用倾向值的倒数作为样本权重，对每个样本进行加权来使得干预变量独立于样本的其它特征变量，达到近似随机对照试验（在随机对照试验中，干预变量与样本的特征变量独立）的效果。具体地，如果用 T_i 表示样本 i 是否接受干预，用 e_i 表示样本 i 的倾向值，则倾向值倒数加权算法中的样本权重可以被表示为：

$$W_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}.$$

进一步，如果用 Y_i 表示样本 i 在观测数据中被观测到的结果，则倾向值倒数加权算法可以通过以下算子来评估因果效应：

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i \cdot Y_i}{e_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1 - T_i) \cdot Y_i}{1 - e_i}, \tag{2-1}$$

其中 n 表示样本数量。

双稳健算法最早在文章[47] 中提出，其目的为了解决倾向值评估时模型假设出错的情况。双稳健算法融合了倾向值倒数加权模型和结果变量回归模型，只要两个模型中有一个模型假设对了，那么该算法就能准确地进行因果效应评估。双稳健算法[8,47,48] 评估因果效应的算子如下：

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i \cdot Y_i - (1 - e_i) \cdot m_1}{e_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1 - T_i) \cdot Y_i + (1 - e_i) \cdot m_0}{1 - e_i},$$

其中 $m_1 = E(Y|T = 1, \mathbf{X})$，$m_0 = E(Y|T = 0, \mathbf{X})$，$\mathbf{X}$ 表示样本的观测特征。

在大数据应用中，我们观测到的变量成千上万。但并不是所有的观测变量都是影响因果推理的混淆变量，还包含噪声变量和其它一些有用变量，例如调整变量。文章[49] 发现，那些不影响干预变量但对结果变量具有预测能力的非混淆变量（称为调整变量）可以被用来调整回归结果变量，实现降低因果效应评估方差的目的。然而，传统基于倾向值的方法主要集中在低维数据上的因果推理和效应评估，

简单地将所有的观测变量都当做混淆变量，忽略了观测变量中非混淆变量的作用，导致倾向值估计的不准确性。因此传统方法很难运用于大数据高维变量下的因果推理和因果效应评估。

变量直接平衡算法[50–60]的提出主要是为了解决传统因果推理方法对模型假设的依赖，例如基于倾向值的方法在评估倾向值时严重依赖模型假设。这些方法的本质目标是通过学习样本权重，使得干预变量与样本的所有特征独立，近似随机对照试验的实验效果。该类方法的主要动机是变量的矩（Moments）可以唯一决定变量的分布①。因此通过平衡干预组和对照组之间样本特征变量的矩可以使得两组之间样本特征变量的分布一致，实现干预变量与样本特征变量独立的目的。

具体地，Hainmueller 等人[53]提出熵平衡（Entropy Balancing）算法来评估观测数据中二值干预变量的因果效应。该算法通过学习样本权重，在保证干预组和对照组的特征变量矩相同的条件下，同时最大化样本权重的熵来减小样本权重的方差，提升因果效应评估的可靠性。Zhao[59]在文章中通过一系列理论分析，证明了熵平衡算法是一种双稳健（Doubly Robust）算法。Athey 等人[50]提出近似残差平衡（Approximate Residual Balancing）算法，该算法为了解决高维变量下的因果效应评估问题，将 LASSO 回归模型和矩平衡模型相结合。通过矩平衡模型学习样本权重，而 LASSO 回归模型学习特征变量对结果变量的回归系数，最后通过对结果变量回归之后的残差项加权来估计因果效应。近似残差平衡算法从理论上是双稳健算法，即只要 LASSO 回归模型跟矩平衡模型一个是无偏的，则最后对因果效应的评估就是无偏的。Zubizarreta 等人[52]将因果效应评估问题看成数据缺失问题，并提出稳定平衡（Stable Balancing）算法。该算法通过约束干预组和对照组之间变量分布一致来学习样本权重，同时最小化样本权重方差使得学习出来的样本权重更稳定。通过接合倾向值方法和变量矩平衡方法，Imai 等人[61]提出了变量平衡的倾向值（Covariate Balancing Propensity Score）算法。基于 EMM 或 EL 框架[62,63]，该方法在学习倾向值的同时，对干预组和对照组之间的矩平衡进行约束，使得最后学习出来的倾向值能平衡干预组和对照组之间的变量矩。因此该方法学习出来的倾向值称为变量平衡的倾向值。

现有基于变量矩直接平衡的因果推理算法在很多实际应用中取得了很好的效果。但是这些方法将所有的观测变量都当做了混淆变量来平衡，且没有考虑到不同混淆变量之间的差异性。实际上，在大数据应用场景中，并不是所有的观测变量都是混淆变量，而且不同的混淆变量给因果推理带来的影响是存在显著差异的。因此，为了更准确地评估大数据因果效应，我们应该对混淆变量进行精细地挑选，

① https://en.wikipedia.org/wiki/Moment_(mathematics)

并在实现混淆变量矩平衡时考虑它们的差异性，着重平衡那些对结果影响大的混淆变量。

2.2　协变量迁移

在机器学习领域，传统预测模型需要假设测试数据集和训练数据集是独立同分布的。然而，在很多实际应用中，我们无法保证需要测试的数据集跟模型训练数据的分布是一致的，也就是不同的数据集之间存在分布偏移。分布偏移使得传统预测模型在与训练集分布不同的测试数据集上的预测效果变得非常糟糕。为了解决数据分布偏移问题，协变量迁移（Covariate Shift）相关算法[64] 被提出。这类方法主要是通过调整训练数据集样本权重，使得训练数据集中协变量的分布与测试数据集中的分布一致。其中，训练样本权重为测试数据中的样本密度与训练数据中的样本密度之比，即 $p_{te}(x)/p_{tr}(x)$，其中 $p_{te}(x)$ 和 $p_{tr}(x)$ 分别表示协变量 x 在测试数据集和训练数据集出现的概率。

解决协变量迁移问题的核心是如何准确地评估测试数据和训练数据之间的样本密度比 $p_{te}(x)/p_{tr}(x)$。Shimodaira 等人[64] 发现，在通过密度比 $p_{te}(x)/p_{tr}(x)$ 加权的训练数据上学习的预测模型，在测试数据集中能达到近似最优的预测效果。为了解决密度比 $p_{te}(x)/p_{tr}(x)$ 带来的高方差问题，Sugiyama 和 Muller[65] 在密度比的基础上，通过引入参数 λ，提出利用一般化的密度比 $(p_{te}(x)/p_{tr}(x))^{\lambda}$，并在理论上证明了一般化的密度比在一定条件下能对测试数据进行近似无偏估计。除此之外，很多方法被提出用来评估样本比率，包括判别估计[66]，内核均值匹配[67]，最大熵方法[68]，最小最大优化算法[69]、强健偏差意识算法[70] 以及其它很多方法[66,71–75]。

在实际应用中，这些方法能很好地解决训练数据集和测试数据集之间协变量分布不一致的问题。然而，这一类方法对于不同的测试数据集，需要重新学习数据集样本权重，并重新训练预测模型，导致其很难运用于在线或流数据的预测问题中。除此之外，这一类方法在学习样本权重时需要以测试数据集的分布作为先验知识。然而在很多实际问题中，测试数据在模型训练阶段是未知的。相比之下，我们在本文中主要的研究问题是对于未知的各种测试数据集，如何进行准确且稳定的预测。

2.3　迁移学习

机器学习高度依赖于训练数据。在实际应用中，一些领域训练数据的不足会导致该域的任务很难取得满意的效果。为了解决该问题，迁移学习（Transfer Learning）[76] 的概念和相关方法被提出。其主要思想是将其它具有充足数据的相关域（源

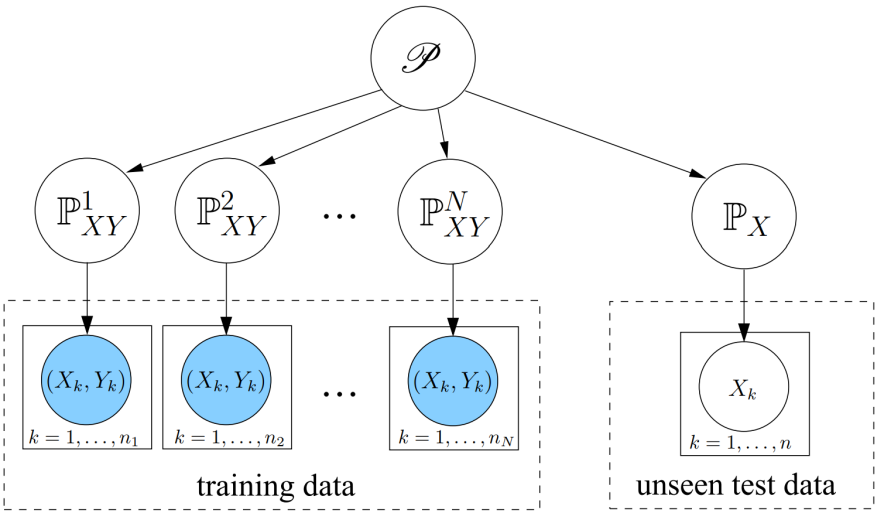


图 2.1　域自泛化框架

域）中的知识迁移到目标域，帮助目标域上的机器学习。根据迁移内容的不同，迁移学习可以分为以下四类：基于示例的迁移学习（Instabce-based Transfer Learning）、基于特征表征的迁移学习（Feature-based Transfer Learning）、基于参数的迁移学习（Parameter Transfer Learning）和基于关系知识的迁移学习（Relational Knowledge Transfer Learning）。

除了解决目标域样本稀疏问题，迁移学习也可以用于解决源域和目标域分布不一致问题。具体地，类似于协变量迁移，通过调整源域（训练集）上数据集的样本权重，使得源域数据中变量的分布与目标域（测试集）上数据分布一致。同样，通过迁移学习解决分布偏差问题，需要目标域（测试集）上数据的先验知识，然而在很多实际预测问题中，我们并不知道未来测试数据集的分布。

2.4　稳定预测

稳定预测指的是预测模型对所有的未知测试数据集都有稳定且准确的预测力，其中未知的测试数据集的分布与训练数据集的分布可能非常不同。稳定预测问题旨在提升预测模型对未知数据预测的泛化能力。为了解决稳定预测问题，一些学者提出利用不变性学习方法来训练预测模型，主要包括域自泛化方法（Domain Generazation）和因果不变性方法（Causal Invariant Prediction）。

近年来，域自泛化研究课题引起了人们的极大关注，并提出了各种解决方法[77–85]。域自泛化主要解决的问题是如何利用多个相关领域（训练数据）的知识，并将其运用到未知的领域（测试数据）中。图2.1展示了域自泛化的简化框架[79]，其目标是在给定 N 个训练数据集上，如何学习预测模型，使得其对未见到过的测试数据具有良好的预测能力。具体地，Muandet 等人[79] 提出了域不变分量分析方法，

通过最小化不同域之间的不相似性来学习特征不变性表征，并用于对未知测试数据的稳定预测。Ghifary 等人[78] 基于离散部件分析方法，提出了一个针对域自适应（Domain Adaptation）和域泛化的统一框架。与以上方法相反，Khosla 等人[77] 认为某个特征数据集上学习出来的预测模型应该有两部分，一部分是真实模型，一部分是数据偏差上学习出的模型。因此，Khosla 等人建议删除每个数据集上的偏差来统一学习数据背后的真实模型。考虑到自动编码器对特征的表征能力，Ghifary 等人[81] 提出了一种多任务自动编码模型来学习多个域中特征的不变性，并约束和指导稳定预测模型的学习，实现对未知测试数据的稳定预测。

在因果不变性这块，Peters 等人[86] 提出了一种稳定预测算法，通过探索多个训练数据集中结果条件分布的不变性来识别预测变量与结果之间的不变性。Rojas-Carulla 等人[87] 提出了一个因果转移框架，用于识别对结果变量预测能力不变的预测变量，然后用于稳定预测。原则上，域自泛化方法和因果不变性方法可以很好地运用于对未知测试数据的稳定预测，但是这些方法的有效性完全依赖于它们多个训练数据集的多样性，并且无法解决多个数据集中都存在的数据偏差和虚假关联等问题。

2.5　可解释性学习

可解释性学习指的是机器学习过程中数据，模型和结果能被人类所理解。根据可解释性对象的不同，可解释性学习可以大致分为三类：建模前的可解释性方法，模型本身的可解释性和对模型结果做出解释。建模前的可解释性方法[30,88–90] 主要针对训练数据集中的分布偏差进行预处理，消除掉由数据偏差带来的数据不可解释性。其中，Kim 等人[88] 提出的 MMD-ceritic 方法可以帮助找到数据中具有代表性或者不具有代表性的样本，帮助我们更好地理解数据情况。因果推理相关工作[89,90] 通过去除数据中混淆变量带来的偏差来恢复变量间的因果关系，提升数据本身的可解释性。例外，为了解决数据中由于样本选择偏差导致的不可解释性或者虚假相关，可以通过倾向值加权[30] 等方法来计算数据中每个样本被选择的概率，并利用其被选择概率的倒数来重新加权样本，从而消除样本选择偏差，提升数据可解释性。第二类可解释性学习是模型本身的可解释性。模型的可解释性指的是模型的判别或者推理过程可以转化成具备逻辑关系的规则。可解释性模型大致可以分为：基于单个特征的方法（例如经典的线性模型等）、基于规则的方法（例如决策树[91] 等）、基于实例的方法（例如贝叶斯实例模型[92] 等）、稀疏方法（例如 LASSO[93]、LDA 模型[94] 等）等。粗略来讲，简单的模型解释性强，复杂的模型解释性弱。简单的模型，例如线性回归模型、逻辑回归模型、决策树模型等都是解释

性很强的模型。而复杂的模型，例如深度神经网络，深度生成模型等的推理过程很难转化成具备逻辑关系的规则。正是由于深度网络等方法的高度复杂性，图灵奖得主 Judea Pearl[95] 认为这些模型只是“曲线拟合”的强大工具，可解释性非常差。从模型有效性或者推理预测准确性角度考虑，简单的模型虽然可解释性强，但是其推理预测效果远不如复杂的深度模型。因此，对模型结果做出解释是一种弥补具有黑箱性质模型可解释性的一种可行方法。对模型结果做出解释的工作主要可以被分为三类，包括隐层分析（investigation on hidden layers）方法[96,97]，模拟/代理模型（mimic/surrogate models）方法[98–101]，敏感性分析（sensitivity analysis）方法[102–107] 等。

在本文中，我们提出的因果约束的稳定学习接合了数据可解释性和模型可解释性。我们提出利用因果技术来恢复数据中的因果关联，实现了数据可解释性；并利用因果关联来指导约束简单模型的学习，实现了模型的可解释性。

第 3 章　数据驱动的变量分解和因果评估模型

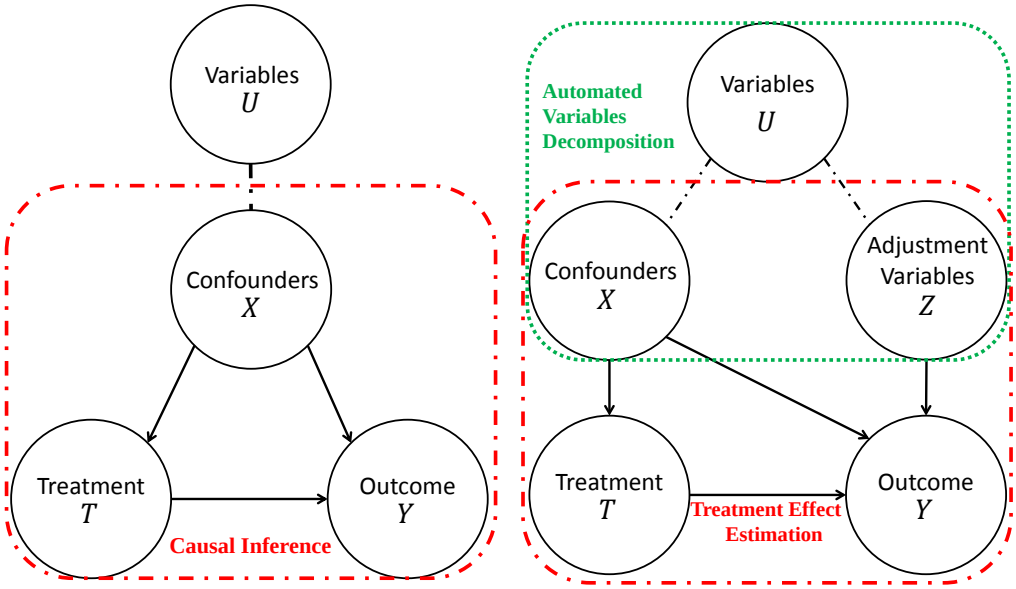
基于观测数据如何评估因果效应是因果推理中最基本的问题之一。在观测数据下，观测变量与干预变量之间可能存在混淆误差。目前消除混淆误差的一般方法是基于倾向值的方法，但是这类方法将所有的观测变量都当做了需要控制的混淆变量，从而忽略了调整变量。这里调整变量指那些对干预变量没有影响，但对结果变量具备预测能力的变量。最近，有研究证明调整变量可以用来降低因果效应评估的方差，提升因果评估的有效性。然而，在观测研究中如何自动分离混淆变量和调整变量仍然是一个悬而未决的问题，尤其是在大数据时代常见的高维变量情景中。在本章中，我们提出了一种数据驱动的变量分解（Data-Driven Variable Decomposition, D^{2}VD）算法，该算法可以：（1）通过数据驱动方法自动分离混淆变量和调整变量；（2）同时对大数据高维变量情景下的因果效应进行评估。在因果推理标准假设下，通过仿真数据和真实在线广告数据上的实验证明，相对于最新的基准方法，我们提出的 D^{2}VD 算法可以精确地实现变量分离，进一步帮助我们更准确地评估因果效应，且方差更小。

3.1　本章引言

因果推理[108] 是指基于效应的发生条件得出关于因果关系结论的过程，是用于解释性分析的强大统计建模工具。用于因果推理的黄金标准方法是随机实验，例如，A/B 测试[109]，其中不同的干预（Treatment）被随机分配到各个样本①。然而，在许多情况下，完全随机化实验的花费通常非常昂贵[110]，有时甚至是不可行的[111]。因此，基于观测数据，开发自动因果推理算法来评估观测研究中的因果效应是非常重要的。

在文献[30] 中提出了一种基于倾向值（propensity score）调整的因果效应评估的统计框架。这种框架已被广泛用于观察性因果推理，包括倾向值匹配（propensity score matching），倾向值分层（propensity score stratification），倾向值倒数加权（inverse propensity weighting）和对倾向值的回归[24,39,42]。其中倾向值倒数加权是最常用的方法，并且已经成为一大类因果模型（边缘结构模型[112,113]）的一部分。通过结合倾向值倒数加权和回归方法，文章[47] 中提出了一种双重稳健的因果效应估计算法。这些方法已被广泛应用于各个领域，包括经济学[7]，流行病学[8]，医疗保健[114]，社会科学[115] 和广告营销[21] 等。

① 样本代表干预对象。例如，在在线广告营销中，这些样本指的是广告营销中的用户。



(a) Neyman-Rubin's 因果推理框架　　(b) 我们提出的因果推理框架

图 3.1　Neyman-Rubin's 因果推理框架与我们提出的因果推理框架比较。（a）在 Neyman-Rubin's 因果推理框架中，所有的观测变量 **U** 都被当做了混淆变量（Confounders）**X**。（b）在我们提出的因果推理框架，我们将所有的观测变量 **U** 分解为三部分：(1) 混淆变量（Confounders）**X**, 指的是那些跟干预变量 T 有关且可能对结果变量 Y 有预测能力的变量；(2) 调整变量（Adjustment Variables）**Z**, 指的是那些对结果变量 Y 具有预测能力但跟干预变量 T 独立的变量；以及 (3) 无关变量 **I** (在因果推理框架中省略了), 指的是那些既跟干预变量 T 独立又跟结果变量 Y 独立的变量。

这些方法的本质是消除混淆变量对干预变量的混淆影响，从而显著提高干预变量的因果效应评估的精确度。但是，当前大多数工作在估计倾向值时将所有观察到的变量都视为混淆变量。然而，在高维变量的场景中，并不是所有的观测变量都是混淆变量，有些变量不是混淆变量且对结果变量 Y 具有预测能力，这些变量我们称之为调整变量 **Z**，如图 3.1中我们提出的因果推理框架所示。忽略调整变量将使估计的因果效应不精确并且影响评估的方差。

最近，一些学者研究了调整变量对于因果效应评估的重要性。文章[116,117] 提倡在进行因果效应评估时，应该将调整变量包含在因果推理模型中。并且文章[118] 发现虽然这些调整变量对消除因果效应评估偏差是不必要的，但可以减少因果效应评估的方差。在随机实验设定中，文章[49] 已经证明通过 LASSO 方法引入调整变量可以减少因果效应评估的方差。

然而，以上这些方法假设在观察性研究中的因果结构，即变量是否是干预变量或结果变量的原因，是先验已知的。然而，在大多数情况下，先验知识无法很好地定义因果结构，特别是在大数据时代的高维变量情景中。如何在观察研究中自动分离混淆变量和调整变量仍然是一个悬而未决的问题。

为了解决这个问题，我们提出了一种数据驱动的变量分解（Data-Driven Variable Decomposition, D^{2}VD）算法来共同优化混淆变量分离和平均因果效应（Average Treatment Effect, ATE）的估计。更具体地说，我们提出了一个正则化的综合回归模型，其中构造了正交性和稀疏性的正则因子，以同时（1）用数据驱动方法分离混淆变量和调整变量；（2）消除既不是混淆变量也不是调整变量的无关变量，防止回归过度拟合；同时（3）估计观察性研究中的 ATE。在估计 ATE 期间，分离的混淆变量可以有效地消除它们对干预变量的混淆影响，而调整变量可以通过结果变量回归显着减少 ATE 估计方差。这使我们算法能够比基准方法更准确地估计真实 ATE 并且具有更紧密的置信区间。

本章的主要贡献主要有以下几点：

- 我们研究了自动分离混淆变量和调整变量的新问题，这对于观察性研究中 ATE 估计的精确度和置信区间至关重要。
- 我们提出了一种新颖的数据驱动变量分离 D^{2}VD 算法，其中提出了一个正则化的综合回归模型，可以同时实现混淆变量分离和 ATE 估计。
- 我们提出的 D^{2}VD 算法的优点在仿真和现实世界数据中得到证明。它还可以直接应用于其它因果推理研究，如社会营销，医疗保健和公共政策。

本章其余部分的结构安排如下。在第 3.2 节我们会回顾相关工作，在第 3.3 节介绍调整的因果效应评估算子，在第 3.4 节提出变量自动分离和因果效应评估算法，在第 3.5 节介绍实验结果，最后在第 3.6 节给出本章小结。

3.2　相关工作

为了进行因果推理，黄金标准方法是随机实验，例如 A/B 测试[109]，其中不同的干预（Treatment）被随机分配给样本。然而，由于成本高或道德原因，完全随机化的实验在很多情况下经常是不可行的[110,111]。

Rosenbaum 和 Rubin[30] 提出了基于倾向值的因果推断框架来进行观察研究中的因果推理。倾向值主要是通过逻辑回归模型评估得到。后来，许多其它机器学习算法（例如 Gradient boosted machine, bagged CART 和随机森林）也被用来估计倾向值[119,120]。基于倾向值，很多因果推理的方法被提出，包括倾向值匹配，倾向值分层，倾向值倒数加权和倾向值回归[23,30,39,42]。在观察性研究中，倾向值倒数加权方法是进行因果推理和因果效应评估的常用方法。它最初是在文章[121] 中提出的，并且是一个被称为边缘结构模型的大型因果模型家族的一部分[112,113]。通过倾向值倒数加权和回归的结合，Bang 和 Robinscite[47] 提出了一个双重稳健的估计量，它在理论上被证明具有较小的渐近方差[48]。倾向值倒数加权和双重稳健方法已被

表 3.1　符号和定义

符号	定义
m	样本数量
n	变量维度
$\mathbf{X} \in \mathbb{R}^{m \times n}$	混淆变量（Confounders）
$\mathbf{Z} \in \mathbb{R}^{m \times n}$	调整变量（Adjustment Variables）
$\mathbf{U} \in \mathbb{R}^{m \times n}$	所有的观测变量, 包含 $\mathbf{X}$ 和 $\mathbf{Z}$
T	干预变量（Treatment）
Y	潜在结果（Potential Outcome）
Y^{obs}	观测到的结果（Observed Outcome）
$Y^{\star}$	转换后的结果（Transformed Outcome）
Y^{+}	调整转换后的结果（Adjusted Transformed Outcome）
S_a	T=1 的样本集合
S_b	T=0 的样本集合

广泛应用于各个领域，包括经济学[7], 流行病学[8], 医疗保健[122], 社会科学[115], 政策学[48] 和广告营销[23] 等。

现有的工作，它们假设所有变量都是混淆变量，并采用图3.1(a) 中所示的因果推理框架进行因果效应评估。与现有工作不同，我们建议将混淆变量和调整变量分开，并重新设计一个新的因果推理框架，如图 3.1(b) 所示，分离出来的混淆变量能精准地估计倾向值进而准确地评估 ATE，而分离出来的调整变量可以用来减少估计 ATE 的方差。此外，我们提出了一种数据驱动方法来自动分离混淆变量和调整变量，同时估计观察研究中干预变量对结果变量的因果效应。

我们的工作与文章[49] 密切相关，它通过用 LASSO 正则化因子引入调整变量来减少随机实验中估计的 ATE 的方差。但我们的工作与文章[49] 的不同之处在于，文章[49] 是为随机实验量身定制的，而在随机实验中无需处理混淆变量。在观察性研究中，我们需要控制混淆变量的影响并尝试在此期间通过引入调整变量来减少方差。我们论文的关键是以数据驱动的自动方式明智地分离这两组变量，进行因果推理。

3.3　调整的因果效应评估算子

在本章中，我们首先给出观察性研究中 ATE 估计的符号和假设，然后利用调整变量来减少估计 ATE 的方差，提出一种新的调整后的 ATE 估计算子。

3.3.1　符号和假设

如我们在图3.1(b) 中的因果推理框架所述，我们将干预（Treatment）定义为随机变量 T，将对应于特定干预 $T = t$ 的潜在结果定义为 $Y(t)$。在本章中，我们只考虑二元干预变量，即 $t \in \{0, 1\}$。我们定义接受干预（Treated）的样本，即 $T = 1$，作为干预样本，其它样本以 $T = 0$ 作为对照（Control）样本。然后，对于每个样本，由 $i = 1, 2 \cdots m$ 索引，我们观察到干预 T_i，结果 Y_i^{obs} 和变量向量 U_i。我们观察到样本 i 的结果 Y_i^{obs} 可表示为：

$$Y_i^{obs} = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0). \tag{3-1}$$

在观察性研究中，有三个标准假设[30] 用于因果效应估计。

假设 1：**稳定的样本干预值（Stable Unit Treatment Value）**。当给定观察到的变量时，假设一个样本的潜在结果的分布不受另一个样本的特定干预值分配的影响。

假设 2：**无混淆性（Unconfounderness）**。给定观察到的变量时，干预变量的分布与潜在的结果的分布无关。形式化地，$T \perp \big(Y(0)Y(1)\big)|\mathbf{U}$。

假设 3：**重叠性（Overlap）**。当给定观察到的变量时，每个样本都有非零概率接受干预和对照状态。形式化地，$0 < p(T = 1|\mathbf{U}) < 1$。

3.3.2　调整的因果效应评估算子

观察性研究中因果推断的重要目标是评估干预变量 T 对结果 Y 的因果效应（ATE）。ATE 代表样本在干预和对照状态下潜在结果之间的平均差异。正式地，ATE 被定义为：

$$ATE = E\big[Y(T = 1) - Y(T = 0)\big], \tag{3-2}$$

其中 $Y(T = 1)$ 和 $Y(T = 0)$ 分别代表样本在干预状态（T=1）和对照状态（T=0）下的潜在结果。$E(\cdot)$ 是指期望函数。

实际上，方程式 （3-2）是不可行的，因为对于每个样本，我们只能观察到其处于干预状态和对照状态中一个状态对应的潜在结果，不能观测到两种状态的结果。这被称为“反事实问题”[23]。

人们可以通过近似未观察到的潜在结果来解决这个反事实问题。最简单的方法是直接比较干预组和对照组之间的平均结果。然而，在观察性研究中，如果干

预变量的分配不是随机的，则直接比较两组样本的平均结果可能会产生偏差，因为没有考虑到混淆变量对干预变量的影响[23]。

为了在观察性研究中无偏地评估 ATE，必须控制混淆变量的影响。根据假设(1,2,3)[30]，我们可以通过引入了倾向值来概括控制混淆变量所需的信息。倾向值，表示为 $e(\mathbf{U})$，被定义为给定所有变量 $\mathbf{U}$ 情况下，样本被分配到干预状态（$T = 1$）的概率。实际上，只有混淆变量 $\mathbf{X}$ 与干预变量相关，因此倾向值可以被表示为:

$$e(\mathbf{U}) = p(T = 1|\mathbf{U}) = p(T = 1|\mathbf{X}) = e(\mathbf{X}). \tag{3-3}$$

基于倾向值，文章[121] 提出通过倾向值倒数加权来解决方程 (3-2) 中的反事实问题。具体地，文章[121] 提出利用倾向值的倒数对观测结果进行加权，得到转换的结果 $Y^{\star}$，并提出了倾向值倒数加权评估算法（Inverse Propensity Weight, IPW）[123]。其中转换的结果 $Y^{\star}$ 被定义为

$$Y^{\star} = Y^{obs} \cdot \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))} = Y^{obs} \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}, \tag{3-4}$$

IPW 算法对因果效应评估算子的定义如下

$$\widehat{ATE}_{IPW} = \widehat{E}(Y^{\star}) = \widehat{E}\left(Y^{obs} \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}\right). \tag{3-5}$$

然而，大多数基于倾向值的方法通常在估计倾向值时将所有观察到的变量都当做混淆变量。这将使得估计的干预效果不精确并且会影响估计方差，因为一些变量可能并不是混淆变量并且对结果变量存在直接影响。

因此，基于我们提出的因果推理框架，如图 3.1所示，我们建议将所有观察到的变量 $\mathbf{U}$ 分成三组，混淆变量 $\mathbf{X}$，调整变量 $\mathbf{Z}$ 和无关变量 $\mathbf{I}$（在图3.1中省略）。然后，我们通过引入调整变量提出了一个新的调整的因果效应评估算子，以减少估计 ATE 的方差。

假设 4：可分离性（Separateness）。观察到的变量 $\mathbf{U}$ 可以分解为三部分，即 $\mathbf{U} = \{\mathbf{X}, \mathbf{Z}, \mathbf{I}\}$，其中 $\mathbf{X}$ 是混淆变量，$\mathbf{Z}$ 是调整变量，$\mathbf{I}$ 是无关变量。

基于假设 4，我们根据 $Y^{\star}$ 引入我们调整后的转换结果 Y^{+}，定义为

$$Y^{+} = \left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}, \tag{3-6}$$

其中 $\phi(\mathbf{Z})$ 是通过利用调整变量 $\mathbf{Z}$ 和结果变量 Y 之间的相关性，来减少结果变量 Y

的方差。

最后，我们提出了调整的因果效应评估算子 $\widehat{ATE}_{adj}$，如下

$$\widehat{ATE}_{adj} = \widehat{E}(Y^{+}) = \widehat{E}\left(\left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}\right). \tag{3-7}$$

3.3.3　理论分析

在该章节，针对我们提出的调整的因果效应评估算子 $\widehat{ATE}_{adj}$，我们给出详细的理论分析，包括误差分析和方差分析。

3.3.3.1　误差分析

对于我们提出的调整的转换结果 Y^{+}（见方程 (3-6)），其具有以下属性。

定理 3.1：　基于假设 1-4, 我们可以得到

$$E(Y^{+}|\mathbf{X}, \mathbf{Z}) = E(Y(1) - Y(0)|\mathbf{X}, \mathbf{Z}). \tag{3-8}$$

证明 首先，基于假设 4,

$$\begin{aligned} & E\left(\phi(\mathbf{Z}) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}|\mathbf{X}, \mathbf{Z}\right) \\ = & \frac{\phi(\mathbf{Z})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))} \cdot E\left(T - e(\mathbf{X})|\mathbf{X}, \mathbf{Z}\right) \\ = & \frac{\phi(\mathbf{Z})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))} \cdot \left(E\left(T|\mathbf{X}\right) - e(\mathbf{X})\right) \\ = & 0. \end{aligned}$$

其次，在相关文献[124] 中，等式 $E(Y^{\star}|\mathbf{X}, \mathbf{Z}) = E(Y(1) - Y(0)|\mathbf{X}, \mathbf{Z})$ 已经被证明成立。所以，我们可以推导出

$$\begin{aligned} & E(Y^{+}|\mathbf{X}, \mathbf{Z}) \\ = & E(Y^{\star}|\mathbf{X}, \mathbf{Z}) - E\left(\phi(\mathbf{Z}) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}|\mathbf{X}, \mathbf{Z}\right) \\ = & E(Y(1) - Y(0)|\mathbf{X}, \mathbf{Z}), \end{aligned}$$

最后，我们可以得到 $E(Y^{+}|\mathbf{X}, \mathbf{Z}) = E(Y(1) - Y(0)|\mathbf{X}, \mathbf{Z})$。　□

基于定理 3.1，我们可以得到以下等式

$$E(Y^+) = E(Y(1)) - E(Y(0)). \tag{3-9}$$

因此，我们提出的调整的因果效应评估算子对因果效应的评估是无偏的。

3.3.3.2　方差分析

基于我们提出的转换结果 Y^+（见方程 (3-6)），我们可以将观测结果变量 Y^{obs} 重写为

$$Y^{obs} \quad = \quad Y^+ \cdot \frac{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}{T - e(\mathbf{X})} + \phi(\mathbf{Z}) + e^+,$$

其中 e^+ 表示的是观测结果 Y^{obs} 与转换结果 Y^+ 之间的残差。

让 σ^2_{adj} 表示我们的调整的因果效应评估算子（$\widehat{ATE}_{adj}$）的渐进方差。根据文献[49] 中的理论结果，我们可以得到

$$\sqrt{|S_a| + |S_b|}\left(\widehat{ATE}_{adj} - ATE\right) \quad \rightarrow \quad \mathcal{N}(0, \sigma^2_{adj}),$$

其中

$$\sigma^2_{adj} \quad = \quad \lim_{|S_a|+|S_b|\rightarrow\infty}\left(\frac{|S_b|}{|S_a|}\sigma^2_{e^{+(1)}} + \frac{|S_a|}{|S_b|}\sigma^2_{e^{+(0)}} + 2\sigma_{e^{+(1)}e^{+(0)}}\right),$$

和

$$\begin{aligned}
\sigma^2_{e^{+(1)}} &= \frac{1}{|S_a| + |S_b|}\sum_{i\in S_a}\left(e_i^{+(1)}\right)^2, \\
\sigma^2_{e^{+(0)}} &= \frac{1}{|S_a| + |S_b|}\sum_{i\in S_b}\left(e_i^{+(0)}\right)^2, \\
\sigma_{e^{+(1)}e^{+(0)}} &= \frac{1}{|S_a| + |S_b|}\sum_{i\in S_a\cup S_b} e_i^{+(1)}e_i^{+(0)},
\end{aligned}$$

以及 $e^{+(1)}$ 和 $e^{+(0)}$ 分别表示干预样本（$T = 1$）和控制样本（$T = 0$）的残差 e^+。

相似的，我们可以基于转换结果的表达式 Y^*（见方程 (3-4)）将观测变量 Y^{obs}

重写为:

$$Y^{obs} = Y^* \cdot \frac{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}{T - e(\mathbf{X})} + e^*,$$

其中 e^+ 表示的是观测结果 Y^{obs} 与转换结果 Y^* 之间的残差。

让 σ^2_{IPW} 表示倾向值倒数加权算子（$\widehat{ATE}_{IPW}$）的渐进方差，我们可以得到

$$\sigma^2_{IPW} = \lim_{|S_a|+|S_b|\to\infty} \left(\frac{|S_b|}{|S_a|} \sigma^2_{e^{*(1)}} + \frac{|S_a|}{|S_b|} \sigma^2_{e^{*(0)}} + 2\sigma_{e^{*(1)}e^{*(0)}} \right),$$

其中 $e^{*(1)}$ and $e^{*(0)}$ 分别表示干预样本（$T = 1$）和控制样本（$T = 0$）的残差 e^*。

对于 σ^2_{adj} 和 σ^2_{IPW}, 我们可以推出以下理论。

定理 3.2: 我们提出的调整因果效应评估算子 $\widehat{ATE}_{adj}$ 的渐进方差不会大于倾向值倒数加权算子 $\widehat{ATE}_{IPW}$ 的渐进方差:

$$\sigma^2_{adj} \le \sigma^2_{IPW}.$$

证明 σ^2_{adj} 和 σ^2_{IPW} 的不同在于:

$$\begin{aligned} \sigma^2_{adj} - \sigma^2_{IPW} &= \frac{|S_b|}{|S_a|} \lim_{|S_a|+|S_b|\to\infty} \left(\sigma^2_{e^{+(1)}} - \sigma^2_{e^{*(1)}} \right) \\ &+ \frac{|S_a|}{|S_b|} \lim_{|S_a|+|S_b|\to\infty} \left(\sigma^2_{e^{+(0)}} - \sigma^2_{e^{*(0)}} \right) \\ &+ 2 \lim_{|S_a|+|S_b|\to\infty} \left(\sigma_{e^{+(1)}e^{+(0)}} - \sigma_{e^{*(1)}e^{*(0)}} \right). \end{aligned} \tag{3-10}$$

基于方程 (3-17) 中 α 与 β 之间的正交约束项，我们可以得到 $\phi(\mathbf{Z})$ 和 $e(\mathbf{X})$ 是结果变量的正交投影。因此

$$\left(Y^* \cdot \frac{1}{W} \right)^{\mathrm{T}} \phi(\mathbf{Z}) = \left(Y^+ \cdot \frac{1}{W} \right)^{\mathrm{T}} \phi(\mathbf{Z}) = 0,$$

其中 $W = \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$。所以，我们可以推导出以下方程:

$$\begin{aligned} \sigma^2_{e^{+(1)}} - \sigma^2_{e^{*(1)}} &= ||e^{+(1)}||_2^2 - ||e^{*(1)}||_2^2 = -||\phi(\mathbf{Z})||_2^2 \le 0, \\ \sigma^2_{e^{+(0)}} - \sigma^2_{e^{*(0)}} &= ||e^{+(0)}||_2^2 - ||e^{*(0)}||_2^2 = -||\phi(\mathbf{Z})||_2^2 \le 0, \\ \sigma_{e^{+(1)}e^{+(0)}} - \sigma_{e^{*(1)}e^{*(0)}} &= \left(e^{+(1)} \right)^{\mathrm{T}} \left(e^{+(0)} \right) - \left(e^{*(1)} \right)^{\mathrm{T}} \left(e^{*(0)} \right) = -||\phi(\mathbf{Z})||_2^2 \le 0. \end{aligned}$$

最后，基于方程 (3-10), 我们可以得到 $\sigma^2_{adj} \leq \sigma^2_{IPW}$. □

因此，我们提出的调整的因果效应评估算子对观测性研究中因果效应的评估是无偏的，且方差比那些将所有观测变量都当做混淆变量的算法（例如，IPW 算法）要小。以上理论分析证明，通过分离混淆变量和调整变量，我们可以使得评估的 ATE 具有更紧密的置信区间。

3.4　变量自动分离算法

在本章中，我们提出数据驱动变量分解（D ^{2}VD）算法来自动分离混淆变量并同时估计因果效应。同时，针对基于观测数据的因果推理的“无真实值事实”问题，我们提出了参数学习的方法。

3.4.1　算法

基于我们调整的因果效应评估算子（见方程 (3-7)），我们可以通过调整的转换变量 Y^+ 回归所有的观测变量 $\mathbf{U}$ 来评估评估因果效应 ATE，并最小化以下目标函数：

$$minimize \quad \|Y^+ - h(\mathbf{U})\|^2. \tag{3-11}$$

然后，我们可以得到调整的因果效应评估算子对平均因果效应的评估值为：

$$\widehat{ATE}_{adj} = E(h(\mathbf{U}))$$

具体地，基于回归系数向量 α 和 γ，我们指定函数 $\phi(\mathbf{Z})$ 和函数 $h(\mathbf{U})$ 为线性函数[①]，也就是

$$\phi(\mathbf{Z}) = \mathbf{Z}\alpha, \tag{3-12}$$

$$h(\mathbf{U}) = \mathbf{U}\gamma, \tag{3-13}$$

基于回归系数向量 β，我们采用线性逻辑回归来评估倾向值 $e(\mathbf{X})$：

$$e(\mathbf{X}) = p(T = 1|\mathbf{X}) = \frac{1}{1 + \exp(-\mathbf{X}\beta)}. \tag{3-14}$$

① 实际上，变量之间的交叉项或者高阶项都可以被包含在观测变量 $\mathbf{U}$ 中，因此函数 $\phi(\mathbf{Z})$ 和函数 $h(\mathbf{U})$ 的线性假设不是很强。

为了简化方程和等式，我们引入了倾向值倒数加权函数 $W(\beta)$。$W(\beta)$ 是关于变量 β 的函数，表示为

$$\begin{aligned} W(\beta) &:= \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))} \\ &= (2T - \mathbf{1}_m) \odot \left(\mathbf{1}_m + \exp\left((\mathbf{1}_m - 2T) \odot \mathbf{X}\beta \right) \right) \end{aligned} \tag{3-15}$$

其中 $\odot$ 表示哈达玛积（Hadamard product），以及 $\mathbf{1}_m = [\underbrace{1, 1, \cdots, 1}_{m}]^T$。

在方程 (3-12, 3-14, 3-15)，我们假设变量分离之后的三部分是已知的，也就 $\mathbf{U} = (\mathbf{X}, \mathbf{Z}, \mathbf{I})$。然而在实际应用中，我们只知道观测变量，但并不知道观测变量中哪些是混淆变量，哪些是调整变量。因此，我们更改目标函数 (3-11)，用观测变量 $\mathbf{U}$ 来替换混淆变量 $\mathbf{X}$ 和调整变量 $\mathbf{Z}$，并提出数据驱动的变量自动分离算法来分离混淆变量和调整变量。更新后的目标函数如下：

$$\begin{aligned} \min \quad & \|(Y^{obs} - \mathbf{U}\alpha) \cdot W(\beta) - \mathbf{U}\gamma\|_2^2, \\ s.t. \quad & \|\alpha\|_1 \le \lambda, \\ & \|\beta\|_1 \le \delta, \\ & \|\gamma\|_1 \le \eta, \\ & \langle \alpha, \beta \rangle = 0. \end{aligned} \tag{3-16}$$

其中优化后的回归系数向量 α 可以用于从观测变量中分离调整变量 $\mathbf{Z}$，而 β 用于分离混淆变量 $\mathbf{X}$。最后，我们可以通过估算 $E(\mathbf{U}\gamma)$ 来评估平均因果效应 ATE。也就是说，利用我们方法优化后的回归系数向量 α，β 和 γ，可以分离混淆变量和调整变量，同时估计因果效应 ATE。

具体地，我们采用了回归系数向量 α 和 β 之间的正交约束来确保混淆变量和调整变量的分离。另外，我们对回归系数 α, β 和 γ 增加了 L_1 的稀疏约束项，用来消除无关变量 $\mathbf{I}$ 的影响，同时防止算法过拟合。

用 $\mathcal{J}(\alpha, \beta, \gamma)$ 表示我们的目标函数，则我们可以将目标函数重写为：

$$\begin{aligned} \mathcal{J}(\alpha, \beta, \gamma) &= \|(Y^{obs} - \mathbf{U}\alpha) \cdot W(\beta) - \mathbf{U}\gamma\|_2^2 \\ &\quad + \lambda\|\alpha\|_1 + \delta\|\beta\|_1 + \eta\|\gamma\|_1 + \mu\|\alpha^T\beta\|_2^2. \end{aligned} \tag{3-17}$$

为了最小化目标函数 $\mathcal{J}(\alpha, \beta, \gamma)$，我们采用近似梯度算法（proximal gradient

algorithm)[125]。因为在我们的目标函数中，我们有 L_1 范数正则化项，它是非平滑且不可导的。对于近似梯度算法中的每个迭代，我们使用近似运算符[125] 作为 L_1 正则化项。

具体地，我们首先将目标函数 $\mathcal{J}(\alpha,\beta,\gamma)$ 分解为两部分：可导部分 $f(\alpha,\beta,\gamma)$ 和不可导部分 $g(\alpha,\beta,\gamma)$ (简化为 $\mathcal{J}$, f 和 g)。也就是说：

$$\begin{aligned}\mathcal{J} &= f(\alpha,\beta,\gamma)+g(\alpha,\beta,\gamma), &(3\text{-}18)\\ f &= \|(Y^{obs}-\mathbf{U}\alpha)\cdot W(\beta)-\mathbf{U}\gamma\|_2^2+\mu\|\alpha^T\beta\|_2^2, &(3\text{-}19)\\ g &= \lambda\|\alpha\|_1+\delta\|\beta\|_1+\eta\|\gamma\|_1. &(3\text{-}20)\end{aligned}$$

然后基于近似梯度算法的操作符拆分属性[125]，在近似梯度算法的第 t 次迭代中为了得到优化的参数（例如, $\alpha^{(t+1)}$），我们可以通过优化函数 $g(\cdot)$ 的近似操作符 $prox_{\kappa^{(t)}g}$：

$$\alpha^{(t+1)}=prox_{\kappa^{(t)}g}\left(\alpha^{(t)}-\kappa^{(t)}\frac{\partial f(\cdot)}{\partial\alpha}\right) \tag{3-21}$$

其中 $\kappa^{(t)}$ 为优化步长，$\frac{\partial f(\cdot)}{\partial\alpha}$ 代表函数 $f(\cdot)$ 对参数 α 的梯度，以及

$$\begin{aligned}prox_{\kappa^{(t)}g}(x) &= (x-\kappa^{(t)}\cdot\lambda)_+-(-x-\kappa^{(t)}\cdot\lambda)_+\\ &= \begin{cases}x_i-\kappa^{(t)}\cdot\lambda & x_i\geq\kappa^{(t)}\cdot\lambda\\ 0 & |x_i|\leq\kappa^{(t)}\cdot\lambda\\ x_i+\kappa\cdot\lambda & x_i\leq-\kappa^{(t)}\cdot\lambda\end{cases} &(3\text{-}22)\end{aligned}$$

方程 (3-22) 中的 λ 指的是函数 $g(\cdot)$ 中参数 α 的系数。如果我们优化的是参数 β，则应该是系数 δ；优化的是 γ 的话，则应该是系数 η。

使用近似梯度算法，我们可以最小化方程 (4-11) 中的目标函数。具体地，从 α,β,γ 上的一些随机初始化开始，我们轮流交替优化三个参数直到收敛。显然，目标函数 $\mathcal{J}(\alpha,\beta,\gamma)$ 的下限为 0，而近似梯度搜索程序将使得 $\mathcal{J}(\alpha,\beta,\gamma)$ 单调递减，保证算法的收敛性。具体来说，函数 $f(\alpha,\beta,\gamma)$ 与所有变量 α,β,γ 的梯度如下：

$$\begin{aligned}\frac{\partial f(\cdot)}{\partial\alpha} &= -2(W(\beta)\cdot\mathbf{1}_n^{\mathrm{T}}\odot\mathbf{U})^{\mathrm{T}}\cdot\left((Y-\mathbf{U}\alpha)\odot W(\beta)-\mathbf{U}\gamma\right)+2\mu\beta,\\ \frac{\partial f(\cdot)}{\partial\beta} &= 2\left((Y-\mathbf{U}\alpha)\cdot\mathbf{1}_n^{\mathrm{T}}\odot\frac{\partial W(\beta)}{\partial\beta}\right)^{\mathrm{T}}\cdot\left((Y-\mathbf{U}\alpha)\odot W(\beta)-\mathbf{U}\gamma\right)+2\mu\alpha,\end{aligned}$$

$$\frac{\partial f(\cdot)}{\partial \gamma} = -2\mathbf{U}^{\mathrm{T}} \cdot \Big((Y - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma \Big).$$

其中

$$\frac{\partial W(\beta)}{\partial \beta} = (2T - \mathbf{1}_m) \odot \exp\Big((\mathbf{1}_m - 2T) \odot \mathbf{U}\beta \Big) \odot (\mathbf{1}_m - 2T) \cdot \mathbf{1}_n^T \odot \mathbf{U},$$

以及 $\mathbf{1}_m = [\underbrace{1, 1, \cdots, 1}_{m}]^{\mathrm{T}}$, $\mathbf{1}_n = [\underbrace{1, 1, \cdots, 1}_{n}]^{\mathrm{T}}$.

因此，基于以上近似梯度方法优化我们提出的数据驱动的变量分离算法来进行混淆变量分离和因果效应评估。我们数据驱动的变量分离算法的具体描述见算法 1.

在算法 1 中的函数 $\hat{f}_\kappa(\cdot)$ 被定义为：

$$\hat{f}_\kappa(x, y) = f(y) + (x - y) + \frac{\partial f(\cdot)}{\partial x}^{\mathrm{T}} (1/(2\kappa)) \|x - y\|_2^2. \tag{3-23}$$

我们的模型和算法可以应用于实际系统中，以处理观察性研究中的因果推理和因果效应评估问题。

3.4.2　复杂度分析

在优化的过程中，主要的计算复杂度在于计算损失函数 $\mathcal{J}$，以及更新变量 α, β, 和 γ。我们分别分析这几项的时间复杂度。对于损失函数的计算，其时间复杂度为 $O(mn)$，其中 m 为样本数量大小，n 为观测变量的维度。对于更新 α，β 和 γ，他们的时间复杂度都为 $O(mn)$。

综上所述，算法 1 中每一轮迭代中的时间复杂度为 $O(mn)$。

3.4.3　模型超参调整

在观察性研究中，因果推理算法的参数调整的主要挑战是我们没有真实 ATE 的值（Ground Truth）与我们的评估值进行比较。

为了解决这个挑战，我们使用匹配方法来评估 ATE 并将其设置为“近似 ATE 真实值”，就像 Athey 和 Imbens 在文章[124] 中所做的那样。具体来说，对于每个样本 i，在测试数据集中找到与样本 i 干预状态相反且变量之间差距最接近的样本进行匹配：

$$match(i) = \arg \min_{j:T_j = 1 - T_i} \|U_i - U_j\|_2^2. \tag{3-24}$$

Algorithm 1 数据驱动的变量分离算法（D^{2}VD）

Require: 初始化 $\mathcal{J}^{(0)} = \mathcal{J}(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})$.

Ensure: $\mathcal{J}^{(0)} \geq 0, \mathcal{J}^{(t+1)} < \mathcal{J}^{(t)}$

for $t = 1, 2, \cdots$ **do**

计算 $\frac{\partial f(\cdot)}{\partial \alpha}$, $\frac{\partial f(\cdot)}{\partial \beta}$ and $\frac{\partial f(\cdot)}{\partial \gamma}$

$\kappa = 1$

while 1 **do**

让 $\alpha^{(t+1)} = prox_{\kappa g}\left(\alpha^{(t)} - \kappa \frac{\partial f(\cdot)}{\partial \alpha}\right)$

break if $f(\alpha^{(t+1)}) \leq \hat{f}_\kappa(\alpha^{(t+1)}, \alpha^{(t)})$

更新 $\kappa = \frac{1}{2}\kappa$

end while

$\kappa = 1$

while 1 **do**

让 $\beta^{(t+1)} = prox_{\kappa g}\left(\beta^{(t)} - \kappa \frac{\partial f(\cdot)}{\partial \beta}\right)$

break if $f(\beta^{(t+1)}) \leq \hat{f}_\kappa(\beta^{(t+1)}, \beta^{(t)})$

更新 $\kappa = \frac{1}{2}\kappa$

end while

$\kappa = 1$

while 1 **do**

让 $\gamma^{(t+1)} = prox_{\kappa g}\left(\gamma^{(t)} - \kappa \frac{\partial f(\cdot)}{\partial \gamma}\right)$

break if $f(\gamma^{(t+1)}) \leq \hat{f}_\kappa(\gamma^{(t+1)}, \gamma^{(t)})$

更新 $\kappa = \frac{1}{2}\kappa$

end while

$$\mathcal{J}^{(t+1)} = \mathcal{J}(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)})$$

end for

通过比较匹配的干预组和控制组之间的平均结果，我们可以得到匹配算法对平均因果效应的评估值 $ATE_{matching}$，并将其用来近似真实的 ATE。我们把它设置为“近似的 ATE 真实值”。然后我们定义因果效应评估的误差为估计的因果效应 $\widehat{ATE}_{adj}$ 和“近似真实值”$ATE_{matching}$ 之间差距的绝对值：

$$Error = |\widehat{ATE}_{adj}(i) - ATE_{matching}(i)|.$$

使用“近似 ATE 真实值”，我们可以利用网格搜索来展开交叉验证，从而学习调整算法的超参。

3.5　实验验证

在本章，我们在仿真数据和真实在线广告数据中都验证我们算法在观测性研究中对因果效应评估的有效性。

3.5.1　基准方法

我们实现了以下基准方法与我们提出的方法进行比较：

- *直接评估算法* $\widehat{ATE}_{dir}$：该方法通过直接比较干预组样本和对照组样本之间的平均结果来评估 ATE。但是该方法忽略了混杂变量对干预变量的混淆影响。
- *IPW 评估算法* $\widehat{ATE}_{IPW}$[30]：该方法通过使用倾向值的倒数对样本进行加权来评估 ATE。但该方法将所有的观测变量都视为混淆变量而忽略了调整变量。
- *双稳健评估算法* $\widehat{ATE}_{DR}$[47]：该方法通过将 IPW 方法和回归方法接合来评估 ATE。但该方法未能区分混淆变量和调整变量。
- *变量不分离评估算法* $\widehat{ATE}_{D^2VD(-)}$：该方法是我们 D^{2}VD 算法的一个弱化版本。在目标函数 (3-17) 中，该方法将超参 μ 的值设置为 0，也就是说该方法没有分离混淆变量和调整变量。

在本章中，为了使得基准方法能应用于高维数据，我们采用了 LASSO 回归来实现 $\widehat{ATE}_{IPW}$ 和 $\widehat{ATE}_{DR}$ 算法。$\widehat{ATE}_{DR}$ 和 $\widehat{ATE}_{D^2VD-}$ 之间的区别在于是前者的估计 ATE 是分步做的，而后者是联合优化的。

3.5.2　仿真数据实验

3.5.2.1　数据描述

为了产生不同的仿真数据，我们设置数据样本大小 $m = \{1000, 5000\}$ 和观测变量维度 $n = \{50, 100, 200\}$。我们首先基于独立高斯分布（Independent Gaussian Distribution）产生观测变量 $\mathbf{U} = (\mathbf{X}, \mathbf{Z}, \mathbf{I}) = (\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}, \mathbf{z}_1, \cdots, \mathbf{z}_{n_z}, \mathbf{i}_1, \cdots, \mathbf{i}_{n_i})$：

$$\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}, \mathbf{z}_1, \cdots, \mathbf{z}_{n_z}, \mathbf{i}_1, \cdots, \mathbf{i}_{n_i} \overset{iid}{\sim} \mathcal{N}(0, 1),$$

其中 n_x, n_z 和 n_i 分别表示混淆变量 $\mathbf{X}$，调整变量 $\mathbf{Z}$ 和无关变量 $\mathbf{I}$ 的维度，以及 $n_x = 0.2 \cdot n, n_z = 0.2 \cdot n, n_i = 0.6 \cdot n.$

表 3.2　仿真数据上的实验结果：真实的 ATE 是 **1**。$Bias$ 指的是评估的因果效应和真实因果效应之间的误差绝对值，也就是 $Bias = |\widehat{ATE} - ATE|$。SD，MAE 和 RMSE 分别表示独立重复 50 实验后，评估因果效应 $\widehat{ATE}$ 的标准偏差（standard deviations），平均绝对误差（mean absolute errors）和均方根误差（root mean square errors）。

	n	$n = 50$				$n = 100$				$n = 200$			
T/m	Estimator	*Bias*	SD	MAE	RMSE	*Bias*	SD	MAE	RMSE	*Bias*	SD	MAE	RMSE
$T = T_{logit}$ $m = 1000$	$\widehat{ATE}_{dir}$	0.418	0.409	0.479	0.582	0.302	0.490	0.472	0.571	0.405	0.628	0.574	0.720
	$\widehat{ATE}_{IPW} + lasso$	0.078	0.310	0.252	0.317	0.097	0.356	0.295	0.366	0.073	0.328	0.267	0.320
	$\widehat{ATE}_{DR} + lasso$	0.060	0.181	0.152	0.189	0.067	0.190	0.155	0.199	0.081	0.181	0.169	0.190
	$\widehat{ATE}_{D^2VD(-)}$	0.053	0.138	0.124	0.146	0.064	0.130	0.117	0.144	**0.018**	0.170	0.128	0.162
	$\widehat{ATE}_{D^2VD}$	**0.045**	**0.108**	**0.091**	**0.116**	**0.019**	**0.114**	**0.093**	**0.115**	0.067	**0.144**	**0.130**	**0.152**
$T = T_{logit}$ $m = 5000$	$\widehat{ATE}_{dir}$	0.418	0.170	0.418	0.451	0.659	0.181	0.659	0.681	0.523	0.412	0.555	0.653
	$\widehat{ATE}_{IPW} + lasso$	0.036	0.201	0.163	0.202	0.034	0.222	0.194	0.213	**0.032**	0.341	0.274	0.325
	$\widehat{ATE}_{DR} + lasso$	0.051	0.079	0.071	0.094	0.106	0.075	0.114	0.127	0.055	0.084	0.086	0.096
	$\widehat{ATE}_{D^2VD(-)}$	0.112	0.080	0.118	0.137	0.114	0.102	0.121	0.150	0.164	0.076	0.164	0.179
	$\widehat{ATE}_{D^2VD}$	**0.033**	**0.072**	**0.061**	**0.078**	**0.023**	**0.073**	**0.061**	**0.073**	0.042	**0.068**	**0.062**	**0.076**
$T = T_{missp}$ $m = 1000$	$\widehat{ATE}_{dir}$	0.664	0.387	0.670	0.766	0.273	0.445	0.436	0.518	0.380	0.766	0.691	0.848
	$\widehat{ATE}_{IPW} + lasso$	0.266	0.279	0.319	0.384	0.298	0.295	0.328	0.417	0.191	0.482	0.403	0.514
	$\widehat{ATE}_{DR} + lasso$	0.138	0.187	0.174	0.231	0.253	0.197	0.269	0.320	**0.050**	0.218	0.170	0.222
	$\widehat{ATE}_{D^2VD(-)}$	0.269	0.162	0.270	0.313	0.129	0.162	0.170	0.206	0.175	0.207	0.236	0.269
	$\widehat{ATE}_{D^2VD}$	**0.066**	**0.113**	**0.102**	**0.129**	**0.019**	**0.119**	**0.101**	**0.120**	0.059	**0.177**	**0.149**	**0.184**
$T = T_{missp}$ $m = 5000$	$\widehat{ATE}_{dir}$	0.446	0.180	0.446	0.480	0.587	0.323	0.587	0.662	0.778	0.246	0.778	0.812
	$\widehat{ATE}_{IPW} + lasso$	0.148	0.133	0.161	0.198	0.172	0.167	0.199	0.239	0.142	0.224	0.206	0.263
	$\widehat{ATE}_{DR} + lasso$	0.119	0.073	0.123	0.139	0.100	0.067	0.107	0.120	0.127	0.079	0.127	0.148
	$\widehat{ATE}_{D^2VD(-)}$	0.112	0.070	0.119	0.132	0.058	**0.067**	0.069	0.086	0.068	0.055	0.073	0.086
	$\widehat{ATE}_{D^2VD}$	**0.033**	**0.055**	**0.052**	**0.063**	**0.039**	0.068	**0.066**	**0.075**	**0.032**	**0.047**	**0.049**	**0.055**

为了测试所有估计算法的稳健性，我们分别用逻辑函数（T_{logit}）和错误指定的函数（T_{missp}）来生成二元干预变量 T：

$$T_{logit} \sim Bernoulli(1/(1 + \exp(-\sum_{i=1}^{n_x} x_i))),$$

$$T_{missp} = 1 \text{ if } \sum_{i=1}^{n_x} x_i > 0.5, \text{otherwise } T_{missp} = 0.$$

结果变量 Y 的产生方式为：

$$Y = \sum_{j=\frac{n_x}{2}}^{n_x} \mathbf{x}_j \cdot \omega_j + \sum_{j=1}^{n_z} \mathbf{z}_k \cdot \rho_k + T + \mathcal{N}(0, 2),$$

在仿真数据中，我们可以发现变量 $(\mathbf{x}_{\frac{n_x}{2}}, \mathbf{x}_{\frac{n_x}{2}+1}, \cdots, \mathbf{x}_{n_x})$ 同时影响干预变量和观测变量，产生了混淆影响。在仿真数据中，真实的因果效应 ATE 是 **1**。

3.5.2.2　评估因果效应 ATE

为了评估所有方法的性能，我们独立进行了 50 次实验。根据我们估算的 ATE，我们计算 *Bias*，SD，MAE 和 RMSE，实验结果见表3.2，其中 *Bias*，SD，MAE 和 RMSE 越小越好。从表3.2中的实验结果，我们有如下发现和分析：

- 直接评估算法在所有的实验设定下都失败了，因果效应估计的 *Bias* 都非常大，这是因为该方法没有考虑观测数据中混淆变量带来的混淆误差。
- IPW 评估算法在 $T = T_{logit}$ 时，对因果效应估计比较准确（*Bias* 较小）；然而当在 $T = T_{missp}$ 时，IPW 评估算法会产生巨大错误（*Bias* 非常大），这是因为当 $T = T_{missp}$，模型假设不对导致 IPW 评估算法对倾向值估计不准确。通过联合 IPW 评估算法和回归算法，双稳健评估算法对因果效应评估的结果要优于 IPW 评估算法，尤其是当模型假设错误的时候，也就是 $T = T_{missp}$。
- 我们的 $D^2VD(-)$ 评估算法，虽然没有考虑变量分解，但是实验结果跟双稳健算法的效果相仿。相比 $D^2VD(-)$，双稳健和其它评估算法，我们提出的 D^2VD 算法通过数据驱动的分离混淆变量和调整变量，在所有的实验设定下都能显著地提升因果效应 ATE 的评估准确性（*Bias* 更小）且降低评估方差（SD 更小）。

3.5.2.3　变量分解验证

正如我们之前所描述的，使用我们算法优化的 $\hat{\alpha}$ 和 $\hat{\beta}$，我们可以将所有的观测变量 $\mathbf{U}$ 分离为混淆变量 $\mathbf{X} = \{\mathbf{U}_i : \hat{\beta}_i \neq 0\}$ 和调整变量 $\mathbf{Z} = \{\mathbf{U}_i : \hat{\alpha}_i \neq 0\}$。为了证明我们的算法的变量自动分离的有效性，我们独立地重复了 50 次实验并在表 3.3中记录真阳性率（true positive rate, TPR）和真阴性率（true negative rate, TNR）。分离混淆变量 $\mathbf{X}$ 的 TPR 和 TNR 的公式定义为：

$$\text{TPR} = \frac{\#\{\hat{\beta}_i \neq 0, \beta_i \neq 0\}}{\#\{\beta_i \neq 0\}}, \tag{3-25}$$

$$\text{TNR} = \frac{\#\{\hat{\beta}_i = 0, \beta_i = 0\}}{\#\{\beta_i = 0\}}. \tag{3-26}$$

类似地，通过参数 α，我们可以计算分离调整变量的 TPR 和 TNR。TPR 和 TNR 越接近于 1 说明变量分离结果越准确。

表 3.3 中的结果表明相对于设定 $T = T_{missp}$，我们算法在设定 $T = T_{logit}$ 下对变量分离的结果更好。这是因为在 $T = T_{logit}$ 设定下，我们算法对倾向值的模型假是对的。不过，即使是在 $T = T_{missp}$ 设定下，我们的算法仍然可以准确地分离混淆变

表 3.3　混淆变量 **X** 和调整变量 **Z** 分离的实验结果。TPR 和 TNR 越接近于 1 越好。

$\mathbf{T} = \mathbf{T}_{\mathbf{logit}}$							
		$n = 50$		$n = 100$		$n = 200$	
m		TPR	TNR	TPR	TNR	TPR	TNR
$m = 1000$	**X**	1.000	0.917	0.977	0.948	0.966	0.906
	Z	1.000	0.973	1.000	0.983	1.000	0.984
$m = 5000$	**X**	1.000	0.923	1.000	0.887	0.994	0.989
	Z	1.000	0.975	1.000	0.987	1.000	0.994
$\mathbf{T} = \mathbf{T}_{\mathbf{missp}}$							
	n	$n = 50$		$n = 100$		$n = 200$	
m		TPR	TNR	TPR	TNR	TPR	TNR
$m = 1000$	**X**	1.000	0.844	0.997	0.866	0.867	0.977
	Z	1.000	0.982	1.000	0.987	1.000	0.983
$m = 5000$	**X**	1.000	0.843	1.000	0.837	0.998	0.965
	Z	1.000	0.986	1.000	0.990	1.000	0.994

量和调整变量。这也是我们算法相对于所有基准方法能更准确评估因果效应的本质原因。

3.5.3　真实数据实验

3.5.3.1　数据描述

我们使用的真实在线广告数据集来自腾讯微信 App [①]，数据收集于 2015 年 9 月期间。在微信中，每个用户都可以在朋友圈向他/她的朋友分享帖子，并像 Twitter 和 Facebook 一样接收朋友分享的帖子。广告商可以通过将广告合并到用户的朋友圈中来向用户推送广告。用户对广告的反馈包括：“喜欢”和“不喜欢”。

我们实际用到的广告数据来自于一款 LONGCHAMP[②]女性手提包。该广告数据包括来自 14,891 位用户的“喜欢”反馈和 93,108 位用户“不喜欢”反馈。对于每个用户，我们收集了 56 维特征，包括（1）人口统计学特征，如年龄，性别，（2）朋友数量，（3）设备（iOS 或 Android），以及（4）微信用户设置，例如是否允许陌生人看他/她的朋友圈（“Share Album to Strangers”）以及是否开通了在线支付功能（“With Online Payment”）。

3.5.3.2　实验设定

在我们的实验中，我们将用户对广告的反馈设置为结果变量 Y。具体来说，当用户 i 喜欢该广告时我们设置结果 $Y_i = 1$，如果用户 i 不喜欢，则 $Y_i = 0$。我们选

① http://www.wechat.com/en/

② http://en.longchamp.com/en/womens-bags

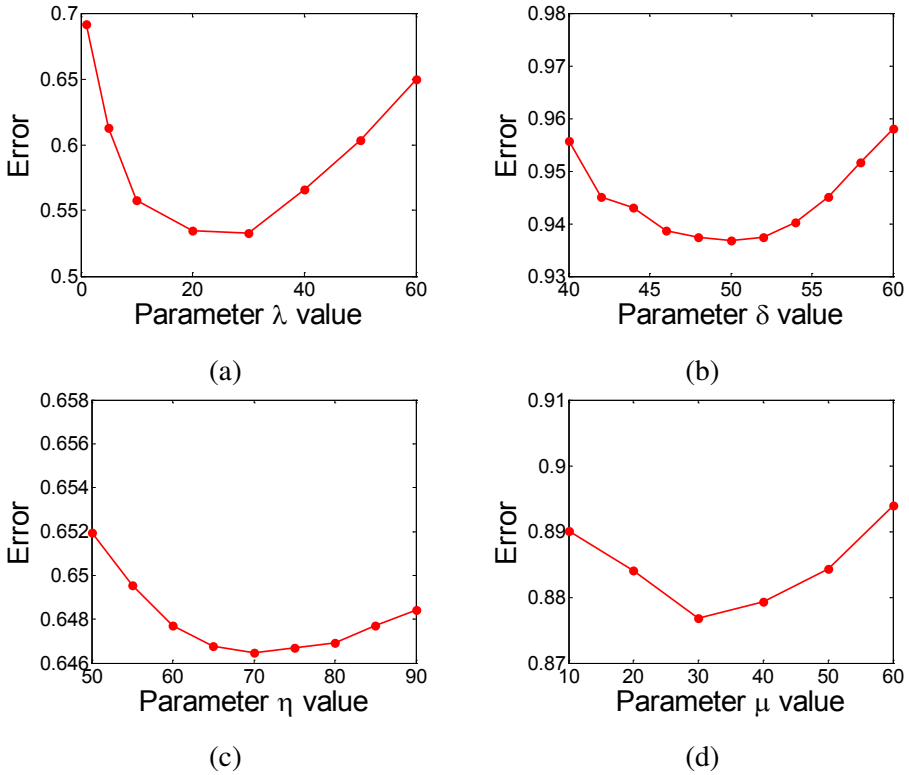


图 3.2　算法超参敏感度分析

择将其中一维特征作为干预变量 T，将其它所有特征当做观测变量 $\mathbf{U}$。这样我们就可以评估每个用户特征对结果的因果效应 ATE。

在参数调整期间，我们设置匹配阈值 $\epsilon = 5$，这使得匹配估计器接近完全匹配。通过使用网格搜索，图3.2显示了随着不同参数值变化时，因果效应评估误差（Error）的变化。最后我们得到我们算法的最佳超参为 $\lambda = 30$，$\delta = 50$，$\eta = 70$ 和 $\mu = 30$。

3.5.3.3　评估因果效应 ATE

对于每个用户特征，我们使用 D^{2}VD 算法来估计其对结果的因果效应 ATE。表 3.4展示了我们 D^{2}VD 算法跟基准方法和“近似 ATE 真实值”$ATE_{matching}$ 对因果效应的评估值。在表 3.4中，我们只列出了因果效应绝对值最大的前 10 个用户特征。值得注意的是，$ATE_{matching}$ 对样本大小有严格的要求，且需要样本之间特征近似完全匹配。对于某些用户特征，我们没有足够的数据样本能实现完全近似全匹配，导致无法推断出它们的 $ATE_{matching}$ 值。

基于表 3.4中实验结果，我们有以下几点发现：

*观察 1。相比于基准方法，我们 D^2VD 算法能更准确地评估每个特征的因果效应 ATE。*通过自动分离混淆变量和调整变量，我们的 D^{2}VD 算法估计的因果效应

表 3.4　真实数据上因果效应评估结果。在该结果中，我们值列出了在因果效应绝对值最大的 10 维特征上，我们 $\widehat{ATE}_{D^2VD}$ 算法与基准方法 $\widehat{ATE}_{IPW}$ 和 $\widehat{ATE}_{DR}$ 算法之间的比较。其中 $ATE_{matching}$ 是“近似 ATE 真实值”, “n/a”表示由于匹配样本不足，我们无法基于匹配方法计算“近似 ATE 真实值”$ATE_{matching}$。

No.	Features	$\widehat{ATE}_{D^2VD}$ (SD)	$\widehat{ATE}_{IPW}$ (SD)	$\widehat{ATE}_{DR}$ (SD)	$ATE_{matching}$
1	No. friends (> 166)	0.295 (0.018)	0.240 (0.026)	0.297(0.021)	0.276
2	Age (> 33)	-0.284 (0.014)	-0.235 (0.029)	-0.302(0.068)	-0.263
3	Share Album to Strangers	0.229 (0.030)	0.236 (0.030)	-0.034(0.021)	n/a
4	With Online Payment	0.226 (0.019)	0.260 (0.029)	0.244(0.028)	n/a
5	With High-Definition Head Portrait	0.218 (0.028)	0.203 (0.032)	0.237(0.046)	n/a
6	With WeChat Album	0.191 (0.014)	0.237 (0.021)	0.097(0.050)	n/a
7	With Delicacy Plugin	0.124 (0.038)	-0.253 (0.037)	0.067(0.051)	0.099
8	Device (iOS)	0.100 (0.024)	0.206 (0.012)	0.060(0.021)	0.085
9	Add friends by Drift Bottle	-0.098 (0.012)	0.016 (0.019)	-0.115(0.015)	-0.032
10	Gender (Male)	-0.073 (0.017)	-0.240 (0.029)	0.065(0.055)	-0.097

ATE 更接近于“近似 ATE 真实值”$ATE_{matching}$。简单地将所有观测变量都视为混淆变量导致 IPW 算法和 DR 算法在估计某些特征的 ATE 时会产生巨大误差，甚至会对 ATE 的正负性错误估计，例如 IPW 算法在估计特征 *With Delicacy Plugin* 的因果效应时，和 DR 算法在估计特征 *Gender* 的因果效应时。

*观察 2。相比于基准方法，我们 D^2VD 算法能降低因果效应 ATE 评估的方差。*通过结果变量对分离出来的调整变量进行回归，我们 D^2VD 算法相对于 IPW 和 DR 基准方法能降低因果效应评估方差，而 IPW 和 DR 算法忽略调整变量的作用，将所有的观测变量都当做了混淆变量。

*观测 3。年轻的女士喜欢 LONGCHAMP 手提包这款广告的概率更大。*我们发现用户特征 $Age(> 33)$ 的因果效应是 -0.284 和用户特征 $Gender(Male)$ 的因果效应是 -0.073，该结果表明年轻的女性用户会更喜欢这款广告。这个发现与我们的直觉一致，因为 LONGCHAMP 广告主要是为年轻女士的潜在客户而设计的。

表 3.4中的某些特征对结果的因果效应无法被直观地解释，这主要是因为存在未观察到的混淆变量或存在其中介效应（mediational effect）[126]，这在先前的因果效应研究中很常见。

3.5.3.4　变量分解验证

这里，我们给出几个研究案例来证明我们提出的因果推理算法在广告数据中能准确地分离混淆变量和调整变量。我们的研究案例包括针对两维不同的干预变量 “Add friends by Shake’’（开通摇一摇加好友功能）和 “With WeChat Album’’（使用微信朋友圈服务）。

表 3.5　干预变量时混淆变量和调整变量分离结果（特征" 开通摇一摇加好友功能" 为干预变量）。

混淆变量	调整变量
With Drift Bottle plugin	No. friends
Add friends by People Nearby	Age
Add friends by QQ Contacts	With WeChat Album
Without Friends Confirmation Plugin	Device

表 3.6　干预变量时混淆变量和调整变量分离结果（特征" 使用微信朋友圈服务" 为干预变量）。

混淆变量	调整变量
Open WeChat Album Service	Friends Count
With High-Definition Head Portrait	Age
Open to Strangers	With Drift Bottle Plugin
With Personal Information	Device

案例 1: *开通摇一摇加好友功能*。微信摇一摇 ① 是一个双向功能，在微信中两个人同时使用摇一摇功能就可以看到对方，并添加对方为好友。表 3.5 显示了当把“开通摇一摇加好友功能”这一特征当做干预变量 T 时，我们 D^{2}VD 算法对混淆变量和调整变量的分离结果。结果显示，我们算法分离出来的混淆变量跟干预变量都十分相关，是微信中其它加好友的方式，例如通过通过漂流瓶加好友功能（“With Drift Bottle Plugin’’），通过附近的人加好友功能（“Add friends by People Nearby” ），和通过 QQ 通讯录加好友功能（“Add friends by QQ Contacts” ）。实验结果说明，我们方法能准确地从观测变量中分离出混淆变量。

对于分离出来的调整变量，我们发现其跟干预变量没有显著的相关性，但对结果变量具有很强的预测能力，例如好友数量（“No. friends” ）和年龄（“Age’’）。从表 3.4中我们可以发现，好友数量（“No. friends” ）和年龄（“Age’’）这两维特征对结果变量的因果效应最大，也就说明我们算法分离出来的调整变量对结果有很强的预测能力，符合我们对调整变量的假设。该实验结果表明，我们方法能准确地分离出调整变量。

基于我们算法准确分离出来的混淆变量和调整变量，我们能提升算法对因果效应评估的准确性且降低评估的方差。

案例 2: *使用微信朋友圈服务*。在微信上，如果用户使用微信朋友圈服务，他/她就可以通过朋友圈跟好友分享新鲜事，也可以看到好友的动态。表 3.6 显示了当把“使用微信朋友圈服务”这一特征当做干预变量 T 时，我们 D^{2}VD 算法对混淆变量

① https://rumorscity.com/2014/07/25/how-to-add-friends-on-wechat-7-ways/

和调整变量的分离结果。从实验结果我们发现，我们算法分离出来的混淆变量跟干预变量都显著相关，例如开通朋友圈服务（“Open WeChat Album Service’’）和朋友圈对陌生人可见（“Open to Strangers’’）等。我们算法分离出来的调整变量对跟干预变量没有显著的相关性，但对结果变量具有很强的预测能力，例如好友数量（“No. friends”）和年龄（“Age’’）。以上结果表明，当把特征“开通微信朋友圈”当做干预变量时，我们算法能准确地从观测变量中分离混淆变量和调整变量。

3.6　本章小结

在本章中，我们将重点放在面向高维数据，如何以更精确的方式评估因果效应。面向高维数据进行因果效应评估，我们认为并不是所有的观测变量都是混淆变量。而先前大多数基于倾向值的因果方法将所有的观测变量都视为混淆变量进行因果评估，导致其很难运用于高维变量环境下准确地评估因果效应。为了解决这个问题，我们提出了全新的因果推理框架。在我们的因果推理框架中，我们建议将所有的观测变量分离为混淆变量和调整变量。基于我们的因果推理框架，我们提出了一种数据驱动的变量分解（D^2VD）算法来联合优化变量分解和 ATE 估计。仿真数据和真实数据上的大量实验验证了面向高维变量的观测数据，我们 D^2VD 算法能更准确地评估因果效应，且方差小。

第 4 章　混淆变量区分性平衡的因果效应估计

因果效应估计在许多需要决策制定的领域扮演着十分重要的角色，例如商业、医疗以及公共政策等领域。在观测性数据的因果效应估计中，最大的挑战就是如何去解决由混淆变量在干预组和对照组之间的不平衡分布所带来的混淆误差。控制高维观测变量可能会使得无混淆性（Unconfounderness）假设更为合理，但是又给倾向值估计带来了新的挑战。最近的研究尝试跳过倾向值估计，而直接优化每个样本的权重来实现混淆变量分布的平衡。但是现有的变量矩平衡方法并不能做到在大量的观测变量中选择和区分不同的混淆变量，导致其很难被运用于面向大数据高维变量的因果效应评估问题。在本章中，我们提出了一个数据驱动的混淆变量区分性平衡算法，在高维数据的设定下同时选择混淆变量、区分不同混淆变量的权重以及平衡在干预组和对照组间混淆变量的分布。我们所提出的协同学习算法能够更好地在观测性数据中减小混淆变量误差。为了验证算法的有效性，我们在仿真数据和真实数据上进行了大量的实验，实验结果清晰地表明了我们的算法在面向大数据高维变量的因果效应评估问题中效果显著优于现有方法。我们进一步展示了我们算法选取的因果效应最大的前 k 个特征能够在在线广告转化效果的预估上取得准确的预测效果。

4.1　本章引言

得益于大数据的兴起，诸如医疗和广告等多个领域都积累下了大量的数据。同时，许多机器学习和数据挖掘方法被提出来利用这些数据进行预测，旨在估计感兴趣领域中某些变量未来的变化。这些方法在面向预测的应用中已经被证明是成功的。然而，大多数预测算法缺乏可解释性使得它们在许多实际应用，尤其是需要决策的领域（如医疗保健和政策制定）中的缺乏吸引力。如何提高学习算法的解释性对于学术研究和实际应用都至关重要。

因果推理是一个用于解释性分析的强大统计建模工具。因果推理的一个根本问题是因果效应估计（Treatment Effect Estimation），其关键挑战是消除干预组和对照组之间混淆变量分布不同引起的混淆偏差。消除混淆偏差的标准方法是进行 A/B 测试等随机实验。但是完全随机化实验的开销通常是非常昂贵的[110]，有时甚至不可行的[111]。在无混淆性（Unconfounderness）假设[30] 下，许多方法直接从观测数据中估计因果效应。这些方法大多数采用倾向值来赋予样本新的权重，以消除混淆偏差[23,42,47]。虽然这些方法在许多实际应用中得到了采用，但它们需要对

干预变量的分配指定正确的模型来准确地评估倾向值。在大数据场景中，控制高维变量可能使得无混淆性假设更加合理，但是又给倾向值的估计带来了新的挑战。最近，一些研究者[50,52,54,127] 提出跳过倾向值的评估，直接通过学习优化样本的权重来平衡混淆变量的分布。但是，他们没有对混淆变量进行筛选和区分，而是平衡所有观察到的变量，从而导致在高维设定中表现不佳。总的来说，在具有模型假设和先验知识的情况下，以前的方法可以在精心设计的实验设定或观察性研究中的因果推理和效应评估取得不错的效果。

然而在真实的大数据场景中，除了少量确定的变量之外几乎总是存在大量额外的或几乎不受控制的混淆变量，并且它们之间的相关性在现实世界中是复杂而未知的[5]。因此，我们在估计观察数据中的因果效应时面临以下挑战：（1）**刻画变量之间相互作用的模型结构是未知的**：如文章[5] 所述，几乎所有的事物都与其它事物相互作用，由于现实世界的复杂性，它们的相互作用是复杂的。我们几乎不知道真实变量之间的模型结构，所以我们不能把任何假定的模型作为消除混淆变量偏差的先验知识。（2）**高维度、具有噪声的变量**：在大数据场景中，总是存在大量的观测变量，但并不是所有变量都是混淆变量，并且不同的混淆变量对数据中的混淆偏差造成的影响也是不同的。通常来说，我们没有足够的先验识证去证明纳入数百甚至数千个变量是合理的，并且如何区分混淆变量以及它们所带来的混淆偏差也是相当困难的。

为了解决这些挑战，我们提出了一种数据驱动方法，称为混淆变量区分性平衡算法。该方法是基于混淆变量平衡的框架，但是与以前所有混淆变量平衡方法不同，我们认为一些变量不应该被认为是混淆变量，并且从理论上证明了混淆变量的权重应该在混淆平衡的过程中被区别处理。在此基础上，我们提出一种联合正则化算法，同时选择混淆变量、区分混淆变量的权重以及平衡混淆变量的分布，来进行因果效应的估计。在因果效应估计过程中，被选中的混淆变量及其权重被用于调整每个样本的权重，从而可以平衡在干预组和对照组中混淆变量的分布。为了验证我们算法的有效性，我们在仿真数据和真实数据上都进行了大量实验。结果表明，我们的算法在观测数据中对于因果效应的估计要优于当前最好的方法。

在本章工作中，我们的主要贡献在于：

- 我们提出了在大数据场景下进行因果效应估计的新挑战，即高维度、具有噪声的变量以及对变量之间结构未知性，这都是以前的方法所不能解决的。
- 我们提出一种混淆变量区分性平衡算法，通过同时选择混淆变量、区分混淆变量的权重以及平衡混淆变量的分布，来实现因果效应的估计。
- 通过仿真和真实数据可以验证我们的算法的有效性，并且我们进一步展示了

我们算法产生的特征能够极大地提升在线广告转化效果估计的准确率。

该章其余部分的结构安排如下。在第 4.2 节我们会回顾相关工作，在第 4.3 节介绍区分性混淆变量平衡算子，在第 4.4 节提出准确估计因果效应的算法，在第 4.5 节介绍实验结果，最后在第 4.6 节给出该章小结。

4.2　相关工作

在观察性数据研究中，现有的基于权重的因果效应估计方法主要有两类：基于倾向值加权和直接矩平衡学习的样本权重。

倾向值首先由 Rosenbaum 和 Rubin[30] 提出，最初是通过逻辑回归来估计的。之后许多其它的机器学习算法（例如，LASSO[128,129], boosting regression[130], bagged CART and neural network[131]）也被用于倾向值估计。基于倾向值提出了各种因果效应评估算法，如倾向值匹配，倾向值倒数加权和双稳健算法[23,42,47] 等。但是这些估计方法需要对于干预变量分配指定正确的模型或需要对倾向值有精确的估计，这在许多应用中可能无法保证[50]。

最近，研究人员通过直接平衡混淆变量[50,52,54,61,127] 提出了新的基于权重的估计量。Hainmueller[127] 引入了熵平衡来将样本权重直接调整到符合指定的样本矩，同时尽可能少地调整样本权重。Athey 等人[50] 提出近似残差平衡（Approximate Residual Balancing）算法，该算法为了解决高维变量下的因果效应评估问题，将 LASSO 回归模型和矩平衡模型相结合。通过矩平衡模型学习样本权重，而 LASSO 回归模型学习特征变量对结果变量的回归系数，最后通过对结果变量回归之后的残差项加权来估计因果效应。近似残差平衡算法从理论上是双稳健算法，即只要 LASSO 回归模型跟矩平衡模型中有一个是无偏的，则最后对因果效应评估值就是无偏的。Zubizarreta[52] 通过接合最小化样本权重方差和直接调整混淆变量平衡来学习稳定的平衡权重。Chan 等人[54] 考虑了构造一个宽级校准权重，来直接获得混淆变量的平衡。Imai 等人[61] 引入协变量平衡的倾向值，在刻画干预分配的同时优化协变量平衡。这些方法中的大多数都是非参数的，并且不需要估计倾向值，但是它们把所有观测到的变量都作为混淆变量来均等地平衡，并没有对混淆变量进行区分，从而导致在高维场景下的因果效应估计中可能表现不佳。因此，通过精细的筛选和区分性对待的方法来改善因果效应评估的效率是非常可能的。为了达到这个目标，我们提出了混淆变量区分性平衡算法来联合优化混淆变量的权重和样本权重，最终达到更为精确的因果效应估计。

表 4.1　符号及定义

符号	定义
n_t (n_c)	干预组（对照组）的样本数量
p	观测变量的维度
$T \in \mathbb{R}^{n\times 1}$	干预变量
$Y \in \mathbb{R}^{n\times 1}$	结果变量
$\mathbf{X} \in \mathbb{R}^{n\times p}$	观测变量
$\mathbf{X_t} \in \mathbb{R}^{n_t\times p}$	干预组的观测变量
$\mathbf{M_t} \in \mathbb{R}^{n_t\times p}$	干预组的增广变量
$\mathbf{M_c} \in \mathbb{R}^{n_c\times p}$	对照组的增广变量
$W \in \mathbb{R}^{n_c\times 1}$	对照组的样本权重
$\beta \in \mathbb{R}^{p\times 1}$	混淆变量权重

4.3　区分性混淆变量平衡算子

在这个部分，我们首先给出相关符号的说明和问题的形式化定义，然后回顾传统的混淆变量平衡估计算子，最后通过区分性混淆变量平衡算法提出一个创新的因果效应估计算子。

4.3.1　符号和问题定义

我们的目标是基于潜在结果框架（Potential Outcome Framework）[30,89] 来估计因果效应。基于这个框架，我们将干预变量定义为随机变量 T，将潜在结果定义为对应于特定干预变量 $T = t$ 的 $Y(t)$。在本章中，我们仅仅关注二值化的干预变量，即 $t \in \{0, 1\}$。我们把接受干预 $(T = 1)$ 的样本称为干预组，其它的样本（$T = 0$）称为对照组。然后，对于编号为 $i = 1, 2, \cdots, n$ 的样本，我们可以观测到一个干预变量 T_i，一个结果变量 Y_i^{obs} 以及一系列的其它观测变量 $X_i \in \mathbb{R}^{p\times 1}$，其中第 i 个样本的结果变量 Y_i^{obs} 可以表示如下：

$$Y_i^{obs} = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0). \tag{4-1}$$

干预组和对照组样本的数量分别为 n_t 和 n_c，所有观测变量的维度为 p。在本章中，对于任意的列向量 $\mathbf{v} = (v_1, v_2, \cdots, v_m)^T$，定义 $\|\mathbf{v}\|_\infty = \max(|v_1|, \cdots, |v_m|)$，$\|\mathbf{v}\|_2^2 = \sum_{i=1}^m v_i^2$，以及 $\|\mathbf{v}\|_1 = \sum_{i=1}^m |v_i|$。

在本章中，我们假设观测性研究中的无混淆性（Unconfounderness）[30] 条件是成立的。

假设 1: 无混淆性（Unconfounderness）在给定观测变量的情况下，干预变量

的分布和潜在结果之间是独立的。形式化地，$T \perp \big(Y(0), Y(1)\big) | \mathbf{X}$.

在本章中，我们关注于估计干预组平均因果效应（Average Treatment Effect on the Treated，ATT），它代表着在干预组中，处于干预状态和对照状态下的样本之间潜在结果的平均差异。形式化地，ATT 定义如下：

$$ATT = E\big[Y(1)|T=1\big] - E\big[Y(0)|T=1\big], \tag{4-2}$$

其中 $Y(1)$ 和 $Y(0)$ 分别表示样本在不同的干预状态下的潜在结果（$T=1$ 表示干预、$T=0$ 表示对照）。本章中我们提出的方法能够被轻易地拓展到对照组平均因果效应（ATC）和总体样本上的平均因果效应（ATE）。

在公式 (4-2) 中, $E[Y(1)|T=1]$ 能够通过样本的计算 $\sum_{i:T_i=1} \frac{1}{n_t} \cdot Y_i^{obs}$ 来直接估计。但是要估计 $E[Y(0)|T=1]$ 却是非常困难的，因为我们并不能在干预组中观察到样本在对照状态的潜在结果。这就是因果推理中经典的“反事实问题”。为了解决这个问题，重新赋权的方法通过给对照组样本给予新的权值 W 使得对照组的混淆变量分布和干预组的混淆变量分布相近，从而找到可以估计无法观测到潜在结果（$Y(0)|T=1$）的替代样本。通过对照组的样本权重 W，我们可以通过如下方法估计 ATT：

$$\widehat{ATT} = \sum_{i:T_i=1} \frac{1}{n_t} \cdot Y_i^{obs} - \sum_{j:T_j=0} W_j \cdot Y_j^{obs}. \tag{4-3}$$

4.3.2　混淆变量平衡的回顾

通过公式 (4-3) 可以看出 ATT 的估计产生出了如何去学习样本权重的问题。经典的学习样本权重的方法是基于倾向值的[23,42,47]。这些方法能够取得良好效果的前提是对于干预变量的分配有正确的模型刻画以及对于倾向值有着精准的估计。但是由于变量之间模型结构的未知，这些方法在没有任何约束的观测性数据研究中经常表现糟糕。

为了减少在无约束环境下对于模型的依赖，研究者们提出了优化样本权重 W 来直接平衡混淆变量分布的方法[50,127]。这些方法背后的动机在于混淆变量的分布可以被它们的矩所唯一决定，进而可以通过平衡矩来平衡它们的分布。因此，它们通过如下的方式来学习样本权重 W：

$$W = \arg\min_{W} \|\overline{\mathbf{X}}_t - \sum_{j:T_j=0} W_j \cdot X_j\|_2^2 \tag{4-4}$$

或者也可以通过下式

$$W = \arg\min_{W} \|\overline{\mathbf{X}}_t - \sum_{j:T_j=0} W_j \cdot X_j\|_{\infty}^2, \tag{4-5}$$

其中 $\overline{\mathbf{X}}_t$ 表示在干预组中观测变量的均值。基于公式 (4-4) 或 (4-5) 直接平衡混淆变量的方法能够被应用在无约束环境下的数据。但是这些方法没有区分对待混淆变量，而是同等地平衡了所有的观测变量，这导致了它们在高维变量场景下表现不佳。

4.3.3　区分性混淆变量平衡

为了在无约束高维变量环境下准确地估计因果效应，我们采用了同时优化混淆变量权重和样本权重的方法。其中，混淆变量的权重可以决定一个变量是否是混淆变量以及它对因果效应带来的误差程度，而样本权重则是为了平衡混淆变量分布而设计的。

更具体地，我们通过解决如下带有约束的优化问题来共同优化混淆变量的权重和样本权重：

$$W = \arg\min_{W} \left(\beta^T \cdot (\overline{\mathbf{X}}_t - \sum_{j:T_j=0} W_j \cdot X_j)\right)^2, \tag{4-6}$$

其中 $W \in \mathbb{R}^{n_c \times 1}$ 表示样本权重，而 $\beta \in \mathbb{R}^{p \times 1}$ 表示混淆变量的权重。在公式 (4-6) 中，混淆变量权重 β 区分了每个混淆变量在平衡过程中所起的作用，从而在无约束观测性研究中能够更好地消除混淆变量带来的混淆偏差。

接下来，我们通过下面的命题来给出如何区分混淆变量权重的理论分析。

命题 4.1： 在观测性的研究中，并不是所有的观测变量都是混淆变量，而且不同的混淆变量以它们各自的权重在估计 ATT 时带来了不同程度的混淆偏差，并且这些权重可以通过潜在结果 $Y(0)$ 对于观测变量 $\mathbf{X}$ 的回归分析来学习。

观测变量 $\mathbf{X}$，干预变量 T 和结果变量 Y 之间的一般关系可以表示如下：

$$Y = f(\mathbf{X}) + T \cdot g(\mathbf{X}) + \epsilon, \tag{4-7}$$

其中真实的因果效应 ATT 为 $E(g(\mathbf{X}_t))$，潜在结果 $Y(0)$ 可以被表示为：

$$Y(0) = f(\mathbf{X}) + \epsilon. \tag{4-8}$$

通过以下线性假设，我们可以证明命题 4.1。

假设 2: 线性假设。潜在结果变量 $Y(0)$ 对于观测变量 $\mathbf{X}$ 的回归是线性的，即 $f(\mathbf{X}) = c + \alpha\mathbf{X}$。

在假设 2 的条件下，我们可以将估计量 $\widehat{ATT}$ 重写如下：

$$\begin{aligned}
\widehat{ATT} &= \sum_{i:T_i=1} \frac{1}{n_t} Y_i^{obs} - \sum_{j:T_j=0} W_j Y_j^{obs} \\
&= \sum_{i:T_i=1} \frac{1}{n_t}(c + \alpha X_i + g(X_i) + \epsilon_i) - \sum_{j:T_j=0} W_j(c + \alpha X_j + \epsilon_j) \\
&= E(g(\mathbf{X}_t)) + (\sum_{i:T_i=1} \frac{1}{n_t}\alpha X_i - \sum_{j:T_j=0} W_j \alpha X_j) + \phi(\epsilon) \\
&= ATT + \sum_{k=1}^{p} \alpha_k (\sum_{i:T_i=1} \frac{1}{n_t} X_{i,k} - \sum_{j:T_j=0} W_j X_{j,k}) + \phi(\epsilon).
\end{aligned}$$

其中 $\phi(\epsilon) = \sum_{i:T_i=1} \frac{1}{n_t}\epsilon_i - \sum_{j:T_j=0} W_j \epsilon_j$ 表示干预组和对照组之间噪声大小的差异，其中 $\phi(\epsilon) \simeq 0$ 如果是高斯噪声。为了减少在估计 ATT 时的偏差，我们需要控制 $\sum_{k=1}^{p} \alpha_k \cdot (\sum_{i:T_i=1} \frac{1}{n_t} X_{i,k} - \sum_{j:T_j=0} W_j X_{j,k})$ 这一项的大小，其中 $(\sum_{i:T_i=1} \frac{1}{n_t} X_{i,k} - \sum_{j:T_j=0} W_j X_{j,k})$ 表示第 k 个混淆变量的分布在干预组和对照组之间的偏差。参数 α_k 表示第 k 个混淆变量在混淆偏差中所占的权重。如果 $\alpha_k = 0$，则表明变量 X_k 不是混淆变量，不需要进行平衡；而且不同的混淆变量，其混淆误差权重 α_k 是不一样的。更重要的，α_k 是函数 $f(\mathbf{X})$ 中变量 X_k 的回归系数。因此，基于线性假设我们可以通过潜在结果 $Y(0)$ 对于观测变量 $\mathbf{X}$ 的回归分析来学习混淆变量的权重。

而事实上，由于反事实问题，潜在结果 $Y(0)$ 对于观测变量 $\mathbf{X}$ 的回归是不可行的，因为我们无法在干预组的样本中观察到潜在结果 $Y(0)$。在这里我们再次利用了样本权重 W 来促进对于干预组中潜在结果 $Y(0)$ 的替代组的构建。我们会在之后详细介绍这一点。

当函数 $f(\mathbf{X})$ 是非线性的时候，即 $f(\mathbf{X})$ 中存在观测变量的高阶项和交叉项时，在概念上是很容易通过 $f(\mathbf{X})$ 的泰勒展开来拓展上述在线性性假设下的结果，我们只需要在平衡观测变量的同时也平衡它们的高阶项和交叉项即可。因此当 $f(\mathbf{X})$ 非线性时，我们需要平衡的是增广变量 $\mathbf{M} = (\mathbf{X}, \mathbf{X}^2, X_iX_j, \mathbf{X}^3, X_iX_jX_k, \cdots)$，并且通过

回归潜在结果 $Y(0)$ 和增广变量 $\mathbf{M}$ 来学习混淆变量的权重。

4.4 优化

在这个部分中，我们会给出混淆变量区分性平衡算法的具体细节，以及介绍在观测性数据因果推断没有“事实”的情况下如何进行参数调整。

4.4.1 算法

通过命题 4.1，我们知道 ATT 估计量是受观测变量以及它们的高阶项的分布不平衡所影响的。增广变量 $\mathbf{M}$ 的含义如下：

$$\mathbf{M} = (\mathbf{X}, \mathbf{X}^2, X_iX_j, \mathbf{X}^3, X_iX_jX_k, \cdots). \tag{4-9}$$

结合公式 (4-6)&(4-9) 以及命题 4.1，我们给出了如下的目标函数，用于联合优化样本权重和混淆变量权重，并估计在观测性研究中的平均因果效应 ATT：

$$\begin{aligned} \min \quad & \left(\beta^T \cdot (\overline{\mathbf{M}}_t - \mathbf{M}_c^T W)\right)^2, \\ s.t. \quad & \sum_{j:T_j=0} (1+W_j) \cdot (Y_j - M_j \cdot \beta)^2 \le \lambda, \\ & \|W\|_2^2 \le \delta, \ \ \|\beta\|_2^2 \le \mu, \ \ \|\beta\|_1 \le \nu, \\ & \mathbf{1}^T W = 1 \quad and \quad W \ge 0, \end{aligned} \tag{4-10}$$

其中 W 表示样本权重，β 表示混淆变量权重。$\overline{\mathbf{M}}_t$ 表示在干预组中增广变量的均值。$\sum_{j:T_j=0}(1+W_j)\cdot(Y_j - M_j\cdot\beta)^2$ 表示在学习混淆变量权重的过程中，关于潜在结果变量 $Y(0)$ 的损失函数，包括了在对照组上的误差 $\sum_{j:T_j=0}(Y_j - M_j\cdot\beta)^2$ 和干预组上的误差 $\sum_{j:T_j=0} W_j\cdot(Y_j - M_j\cdot\beta)^2$，同样也是经过赋权进行替代。通过约束 $\|\beta\|_2^2 \le \mu$ 和 $\|\beta\|_1 \le \nu$，我们可以消除非混淆变量并平滑混淆变量的权重。公式 $\mathbf{1}^T W = 1$ 正则化了在对照组中样本的权重，让它们和为 1，与干预组相同。$W \ge 0$ 约束了样本权重为非负值。通过约束 $\|W\|_2^2 \le \delta$，我们可以降低样本的方差以保持稳定。

综上，我们求解的是以下的优化问题，即在满足参数 W 的约束下最小化 $\mathcal{J}(W,\beta)$。

$$\begin{aligned} \mathcal{J}(W,\beta) \quad = \quad & \left(\beta^T \cdot (\overline{\mathbf{M}}_t - \mathbf{M}_c^T W)\right)^2 \\ & + \lambda \sum_{j:T_j=0} (1+W_j)\cdot(Y_j - M_j\cdot\beta)^2 \end{aligned} \tag{4-11}$$

$$
\begin{aligned}
&+\delta\|W\|_2^2+\mu\|\beta\|_2^2+\nu\|\beta\|_1,\\
s.t.\quad &\mathbf{1}^T W=1 \quad and \quad W\geq 0.
\end{aligned}
$$

这里我们提出了一个迭代的方式来优化上述目标函数 (4-11)。

首先，我们初始化样本权重 $W=\{1/n_c,\cdots,1/n_c\}^T$ 以及混淆变量权重 $\beta=\{1/p,\cdots,1/p\}^T$。一旦初值给定，在每次迭代中，我们首先固定 W，更新 β，然后固定 β，更新 W。这些步骤如下所述：

更新 β: 当固定 W 时，问题 (4-11) 等价于优化下面的目标函数：

$$
\begin{aligned}
\mathcal{J}(\beta) \;=\; & \left(\beta^T\cdot(\overline{\mathbf{M}}_t-\mathbf{M}_c^T W)\right)^2+\mu\|\beta\|_2^2+\nu\|\beta\|_1\\
&+\lambda\sum_{j:T_j=0}(1+W_j)\cdot(Y_j-M_j\cdot\beta)^2
\end{aligned}
\tag{4-12}
$$

这是一个经典的 ℓ_1 范数正则化的最小均方误差问题，可以通过任何 LASSO 回归的程序求解。在这里我们使用了近似梯度算法[132] 来求解 (4-12) 中的目标函数。

更新 W: 通过固定 β，我们可以通过优化 (4-11) 来得到 W。这等价于优化下面的目标函数：

$$
\begin{aligned}
\mathcal{J}(W) \;=\; & \left(\beta^T\cdot(\overline{\mathbf{M}}_t-\mathbf{M}_c^T W)\right)^2+\delta\|W\|_2^2\\
&+\lambda\sum_{j:T_j=0}(1+W_j)\cdot(Y_j-M_j\cdot\beta)^2,\\
s.t.\quad &\mathbf{1}^T W=1 \quad and \quad W\geq 0.
\end{aligned}
\tag{4-13}
$$

为了保证 W 的非负性，我们令 $W=\omega\odot\omega$，其中 $\omega\in\mathbb{R}^{p\times 1}$，$\odot$ 代表哈达玛积。因此 (4-13) 可以重写为：

$$
\begin{aligned}
\mathcal{J}(\omega) \;=\; & \left(\beta^T\cdot(\overline{\mathbf{M}}_t-\mathbf{M}_c^T(\omega\odot\omega))\right)^2+\delta\|\omega\odot\omega\|_2^2\\
&+\lambda\sum_{j:T_j=0}(1+\omega_j\odot\omega_j)\cdot(Y_j-M_j\cdot\beta)^2,\\
s.t.\ &\mathbf{1}^T(\omega\odot\omega)=1.
\end{aligned}
\tag{4-14}
$$

函数 $\mathcal{J}(\omega)$ 对 ω 的偏导数为：

$$
\begin{aligned}
\frac{\partial\mathcal{J}(\omega)}{\partial\omega} \;=\; & -4(\beta^T\cdot(\overline{\mathbf{M}}_t-\mathbf{M}_c^T(\omega\odot\omega)))\cdot\mathbf{M}_c\cdot\beta\odot\omega\\
&+4\delta\omega\odot\omega\odot\omega+2\lambda\omega\odot(Y_c-\mathbf{M}_c\cdot\beta)^2.
\end{aligned}
$$

然后我们通过线性搜索来确定步长 a，并通过如下方式在第 t 轮迭代中更新 ω：

$$\omega^{(t)} = \omega^{(t-1)} - a \cdot \frac{\partial \mathcal{J}(\omega^{(t-1)})}{\partial \omega^{(t-1)}}.$$

为了满足约束 $\mathbf{1}^T(\omega \odot \omega) = 1$，我们可以根据如下的方式来正则化 $\omega^{(t)}$：

$$\omega^{(t)} = \frac{\omega^{(t)}}{\sqrt{\mathbf{1}^T(\omega^{(t)} \odot \omega^{(t)})}}.$$

然后，我们通过如下方式在第 t 轮迭代中更新 $W^{(t)}$：

$$W^{(t)} = \omega^{(t)} \odot \omega^{(t)}.$$

我们迭代式地更新 β 和 W 直到目标函数 (4-11) 收敛。整个算法的描述见算法 2。

最后，通过最优化的样本权重 W，我们可以根据公式 (4-3) 来估计平均因果效应 ATT。

Algorithm 2 混淆变量区分性平衡算法（Differentiated Confounder Balancing, DCB）

Require: 权衡参数 $\lambda > 0$，$\delta > 0$，$\mu > 0$，$\nu > 0$，干预组增广变量矩阵 $\mathbf{M}_t$，对照组增广变量矩阵 $\mathbf{M}_c$ 以及结果变量 Y。

Ensure: 混淆变量权重 β 和样本权重 W

初始化混淆变量权重 $\beta^{(0)}$ 和样本权重 $W^{(0)}$
通过公式 (4-11) 计算目标函数当前值 $\mathcal{J}(W, \beta)^{(0)} = \mathcal{J}(W^{(0)}, \beta^{(0)})$
初始化迭代变量 $t \leftarrow 0$
repeat
    $t \leftarrow t + 1$
    通过求解公式 (4-12) 中的 $\mathcal{J}(\beta^{(t-1)})$ 更新 $\beta^{(t)}$
    通过求解公式 (4-13) 中的 $\mathcal{J}(W^{(t-1)})$ 更新 $W^{(t)}$
    计算 $\mathcal{J}(W, \beta)^{(t)} = \mathcal{J}(W^{(t)}, \beta^{(t)})$
until $\mathcal{J}(W, \beta)^{(t)}$ 收敛或达到迭代上限
return β, W.

注释 4.1：　在我们的算法中产生的混淆变量权重 β 可以被用于像文章[50] 中一样进行结果变量残差调整。有了最优化的混淆变量权重和样本权重 β、W 之后，我们可以通过如下的混淆变量平衡和残差调整结合的方式估计平均因果效应 ATT：

$$\widehat{ATT} = \sum_{i:T_i=1} \frac{1}{n_t} \cdot Y_i^{obs} - \Big(\overline{\mathbf{M}}_t \cdot \beta + \sum_{j:T_j=0} W_j (Y_j^{obs} - M_i \cdot \beta)\Big). \tag{4-15}$$

4.4.2　复杂度分析

在优化的过程中，主要的计算复杂度在于计算 $\mathcal{J}(W, \beta)$，以及更新混淆变量权重 β 和样本权重 W。我们分别分析这几项的时间复杂度。对于损失函数的计算，其时间复杂度为 $O(np)$，其中 n 为样本数量大小，p 为观测变量（增广变量）的维度。对于更新 β，这是一个经典的 LASSO 回归的问题，其时间复杂度为 $O(np)$。对于更新 W，其复杂度的主导项是目标函数 $\mathcal{J}(\omega)$ 对于变量 ω 偏导数的计算。偏导数 $\frac{\partial \mathcal{J}(\omega)}{\partial \omega}$ 的计算时间复杂度为 $O(np)$。

综上所述，算法 2 在每一轮迭代中的时间复杂度为 $O(np)$。

4.4.3　超参调整

在观测性数据的因果推断研究中，对于参数调整最大的挑战是没有真实因果效应 ATT。为了解决这个问题，我们如同文章[124,133,134] 中一样采用了匹配的方式来估计 ATT，并将其设为“近似 ATT 真实值”。特别地，对于每一个干预样本 i，我们通过如下方式寻找它在对照组中的最近匹配：

$$match(i) = \arg\min_{j:T_j=0} \|X_i - X_j\|_2^2. \tag{4-16}$$

为了使匹配接近于精准匹配，对于样本 i，如果 $match(i) > \varepsilon$ 我们则丢弃它。然后我们可以通过比较匹配上的干预组和对照组之间的平均结果差异来获得“近似 ATT 真实值”。

有了“近似 ATT 真实值”之后，我们可以利用网格搜索方法通过交叉验证的方式调整算法的超参。

4.5　实验验证

在这个部分中，我们会在仿真和真实数据集上验证我们算法对大数据背景下因果推理和效应评估的有效性，并与当前最好的方法作比较。

4.5.1　基准方法

为了便于比较，我们实现了如下的基准方法。

- *直接估计量* $\widehat{ATT}_{dir}$：这个方法通过比较干预组和对照组输出结果的均值来计算 ATT，忽略了数据中的混淆偏差。
- *IPW 估计量* $\widehat{ATT}_{IPW}$[21]：这个方法通过倾向值的倒数来对样本重新加权，依赖于对倾向值计算的正确模型假设。
- *双稳健估计量* $\widehat{ATT}_{DR}$[4]：这个方法通过结合 IPW 和回归的方法来估计 ATT，依赖于对倾向值计算的正确模型假设或者对输出结果正确的回归模型假设。
- *熵平衡估计量* $\widehat{ATT}_{ENT}$[11]：这个方法通过直接平衡混淆变量和样本权重的熵误差来估计 ATT，忽略了混淆变量的权重。
- *近似残差平衡估计量* $\widehat{ATT}_{ARB}$[2]：这个方法结合了通过混淆变量平衡进行权重调整以及对于输出结果的回归来进行 ATT 估计，忽略了混淆变量权重。

在具体实验中，我们通过 LASSO 回归实现了 $\widehat{ATT}_{IPW}$ 和 $\widehat{ATT}_{DR}$，用于高维数据中进行初步混淆变量的选择。

4.5.2　仿真数据实验

在这个部分，我们将介绍如何生成一个仿真数据集，并展示我们的算法在数据集上的有效性。

4.5.2.1　数据描述

为了生成仿真数据集，我们考虑两种样本大小 $n = \{2000, 5000\}$ 和两种观测变量维度 $p = \{50, 100\}$。我们首先通过独立高斯分布来产生观测变量 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)$，形式化地：

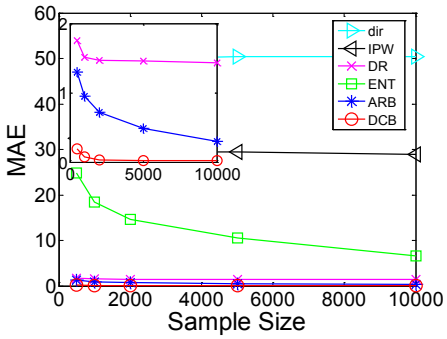
$$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p \overset{iid}{\sim} \mathcal{N}(0, 1),$$

其中 $\mathbf{x}_i$ 表示 $\mathbf{X}$ 中第 i 个变量。

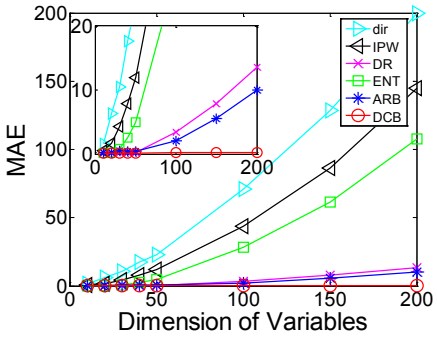
为了测试所有估计量的鲁棒性，我们通过逻辑函数 (T_{logit}) 和错误的函数 (T_{missp}) 来生成二值的干预变量 T：

$$T_{logit} \sim Bernoulli(1/(1 + \exp(-\sum_{i=1}^{p \cdot r_c} s_c \cdot x_i + \mathcal{N}(0, 1)))), and$$

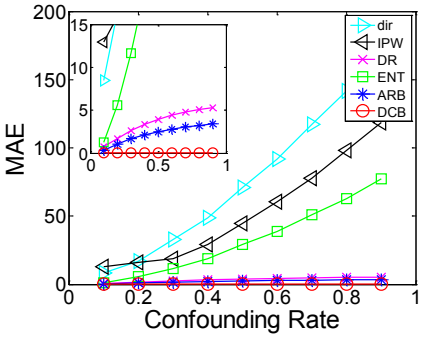
$$T_{missp} = 1 \; if \; \sum_{i=1}^{p \cdot r_c} s_c \cdot x_i + \mathcal{N}(0, 1) > 0, \; T_{missp} = 0 \; otherwise$$



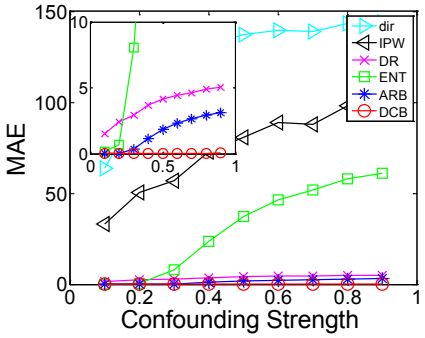
(a) 样本大小 n



(b) 观测变量维度 p



(c) 混淆变量比例 r_c



(d) 混淆变量强度 s_c

图 4.1　在不同设定下各种方法估计 ATT 的 MAE 曲线。

我们令**混淆变量比例** r_c 和 **混淆变量强度** s_c 从 0 变化到 1。其中混淆变量比例代表所有观测变量中混淆变量的比例；混淆变量强度代表混淆变量对于一个干预变量的混淆程度。

我们通过一个线性函数 (Y_{linear}) 和一个非线性函数 (Y_{nonlin}) 来生成 Y：

$$Y_{linear} = T + \sum_{j=1}^{p} \{I(mod(j,2) \equiv 0) \cdot (\frac{j}{2} + T) \cdot \mathbf{x}_j\} + \mathcal{N}(0,3),$$

$$Y_{nonlin} = T + \sum_{j=1}^{p} \{I(mod(j,2) \equiv 0) \cdot (\frac{j}{2} + T) \cdot \mathbf{x}_j\} + \mathcal{N}(0,3)$$
$$+ \sum_{j=1}^{p-1} \{I(mod(j,10) \equiv 1) \cdot \frac{p}{2} \cdot (x_j^2 + x_j \cdot x_{j+1})\},$$

其中 $I(\cdot)$ 为示性函数，$mod(x,y)$ 表示 y 对 x 的模。

在仿真实验中，对于不同的干预变量以及结果变量的设定，我们都知道真实的平均因果效应 ATT；我们通过区分性混淆变量平衡算法估计出 ATT，并与基准方法进行了比较。

4.5.2.2　实验结果

为了评价我们的方法的性能，我们在数据集上独立地进行了 100 次实验，并基于这 100 次试验进行结果分析。基于算法估计的 ATT($\widehat{ATT}$)，我们通过如下的定义计算了绝对误差（$Bias$）、标准差（SD）、平均绝对误差（MAE）和均方根误差（$RMSE$）：

$$
\begin{aligned}
Bias &= \left| \frac{1}{K} \sum_{k=1}^{K} \widehat{ATT}_k - ATT \right| \\
SD &= \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(\widehat{ATT}_k - \frac{1}{K} \sum_{k=1}^{K} \widehat{ATT}_k \right)^2} \\
MAE &= \frac{1}{K} \sum_{k=1}^{K} \left| \widehat{ATT}_k - ATT \right| \\
RMSE &= \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(\widehat{ATT}_k - ATT \right)^2}
\end{aligned}
$$

其中 K 代表独立实验的次数，$\widehat{ATT}_k$ 表示第 k 次实验 ATT 的估计值，ATT 表示真实值。

在表 4.4和4.5中，我们汇报了在各种设定下的实验结果。针对表 4.4和4.5中的实验结果，我们有以下发现和分析：

- 直接估计法 $\widehat{ATT}_{dir}$ 在混淆变量同时影响干预变量和结果变量时效果非常差，从表中我们也能看出直接估计方法的误差非常高，因为它忽略了数据中的混淆偏差。
- IPW 估计量 $\widehat{ATT}_{IPW}$ 和双稳健估计量 $\widehat{ATT}_{DR}$ 在高维变量和模型假设不正确的情况下表现糟糕，从表中我们发现在设定 3 和设定 4 中，$T = T_{missp}$ 和 $Y = Y_{nonlin}$，IPW 和双稳健算法对因果效应评估的误差非常大。
- 熵平衡估计量 $\widehat{ATT}_{ENT}$ 在混淆偏差比较小的时候表现得很好（即表中的设定 2），但是一旦混淆偏差升高其表现就会明显退化。因为它没有考虑到混淆变量不同的权重，导致其不能完全去除混淆偏差。
- 近似残差平衡估计量 $\widehat{ATT}_{ARB}$ 在绝大多数场景下都优于其它基准估计量，但其表现远远不如我们的算法，关键原因在于它相同程度地平衡了所有的观测变量，并没有考虑混淆变量的区分性。
- 我们的算法 $\widehat{ATT}_{DCB}$ 通过联合优化样本权重和混淆变量权重在各种设定下，因果效应苹果的准确性比所有基准方法都有着明显的提高。

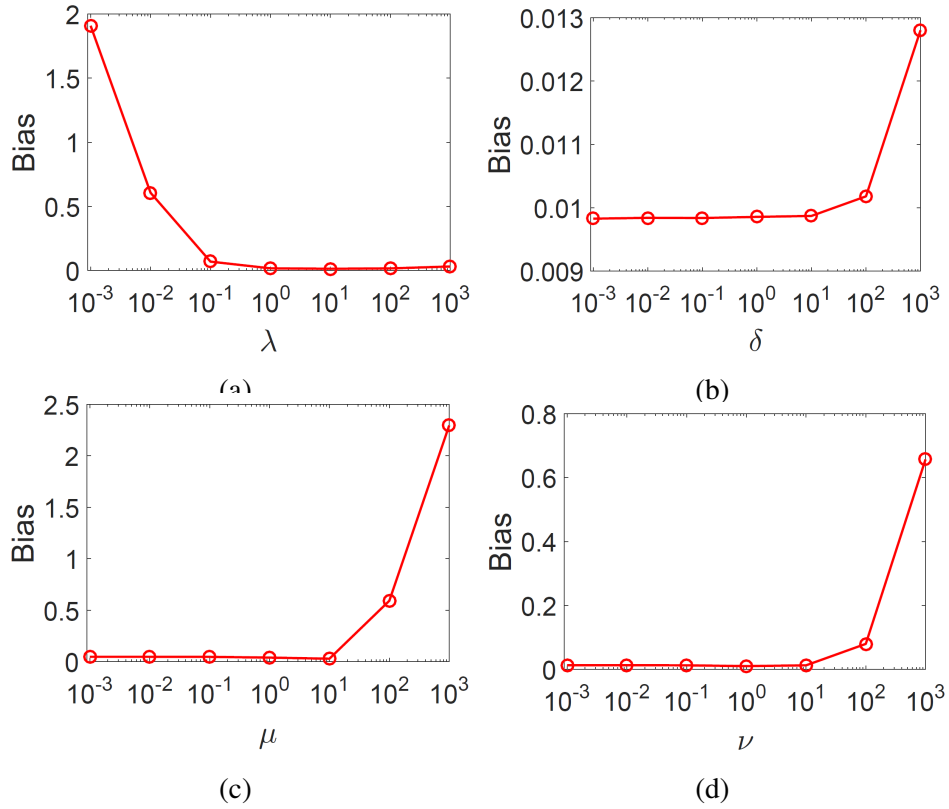


图 4.2　超参 λ, δ, μ 和 ν 对因果效应评估的影响。

此外，从图 4.1中也能看出我们的算法的鲁棒性，当我们减少 n 或者增加 p，r_c 和 s_c 时，我们的算法的 MAE 误差依旧稳定地维持在较低的水平，而其它方法所产生的误差在此情况下都有明显的提升。

4.5.2.3　超参分析

在我们的 DCB 算法中，我们有四个超参，包括 λ，δ，μ 和 ν。如前所述，我们在实验中通过网格搜索进行交叉验证来调整这些参数，并且每个参数从 {0.001,0.01,0.1,1,10,100,1000} 均匀变化。我们分别展示了 λ，δ，μ 和 ν 在 *Bias* 上对因果效应估计的影响。从图4.2可以看出，当参数 $\lambda \geq 1$ 和 $\delta, \mu, \nu \leq 1$ 时，*Bias* 变化不大且性能相对稳定。从图4.2(a)，我们可以看到 *Bias* 在参数 λ 太小时很大。主要原因是 λ 的小值会减少对混淆变量权重学习的限制，导致对混淆变量权重学习不准确，甚至会导致奇异解 $\beta = 0$。从图4.2(c)和4.2(d)，我们发现 *Bias* 随着 μ 和 ν 的增加而增加。这是因为 μ 和 ν 的大值会使得混淆变量权重趋于 0。图4.2(b)表明性能对参数 δ 不敏感。从以上超参分析，我们发现我们 DCB 算法可以很轻松得获得最佳超参。

表 4.2　LaLonde 数据集上 ATT 预估结果

Variables Set	V-RAW		V-INTERACTION	
Estimator	$\widehat{ATT}$	*Bias* (SD)	$\widehat{ATT}$	*Bias* (SD)
$\widehat{ATT}_{dir}$	-8471	10265 (374)	-8471	10265 (374)
$\widehat{ATT}_{IPW}$	-4481	6275 (971)	-4365	6159 (1024)
$\widehat{ATT}_{DR}$	1154	639 (491)	1590	204 (812)
$\widehat{ATT}_{ENT}$	1535	259 (995)	1405	388 (787)
$\widehat{ATT}_{ARB}$	1537	257 (996)	1627	167 (957)
$\widehat{ATT}_{DCB}$	1958	**164** (728)	1836	**43** (716)

4.5.3　真实数据集实验

在这一小节，我们将混淆变量区分性平衡算法应用到两个真实数据上，包括 LaLonde 数据集和在线广告数据集。

4.5.3.1　LaLonde 数据集

首先，我们将算法应用到了因果推理领域[127,135] 一个权威的数据集——LaLonde[136] ①数据集上。LaLonde 数据集的实验数据为两个部分。第一部分来源于一个基于大规模职业培训项目——NSW 的随机试验；第二部分如同文献[127] 中所做的一样，我们将随机试验中的对照组换成了由 CPS-1 采样得到的对照组，其包含的观测变量与之前的试验是相同的。在这个数据集中，作为干预变量的是参与调查者是否参加了这个特定的职业培训项目，而结果变量则是参与调查者 1978 年的收入。数据中有 10 个原始的观测变量，包括 1974 至 1975 年的收入和雇佣状态、学历状态（上学年限和是否有高中学历）、年龄、种族和婚姻状态。

总的来说，在 NSW 的数据包含 *185* 位培训项目参与者（即干预组）和 *260* 位未参与培训的人员（即对照组）；在 CPS-1 的数据中，对照组换成了 *15,992* 位未参与人员。NSW 的随机试验提供了项目的平均因果效应 ATT 的真实值，我们在 CSP-1 的数据上进行实验以对比我们的算法和基准方法。

实验设定：在实验中，我们将 CPS-1 的数据随机切分成了 6 个部分，我们用前 3 个部分用来训练我们的模型以及基准方法并进行交叉验证来调整参数，剩余的 3 个部分用来测试算法的性能。我们在两个变量集合 V-RAW 和 V-INTERACTION 上检验了我们的算法。V-RAW 集合为初始的 10 个观测变量，而 V-INTERACTION 集合包含了原始变量的交叉项和平方项。

实验结果：我们在表 4.2展示了实验结果，其中 *Bias* 和 *SD* 指标为越小越好，从表中我们可以得出如下结论：(1) 由于混淆偏差的存在，直接估计法在 LaLonde

① The dataset is available at http://users.nber.org/~rdehejia/data/nswdata2.html

表 4.3　混淆变量区分性平衡算法在 V-RAW 集合上学习的变量权重

Rank	Confounder	Weight
1	Earnings 1975	0.335
2	Earnings 1974	0.241
3	Employed 1975	0.141
4	Education	0.138
5	Employed 1974	0.050
6	Married	0.039
7	High School Degree 1975	0.017
8	Age	-0.013
9	Black	-0.003
10	Hispanic	-0.001

数据集上无法很好地估计 ATT。(2) IPW 在两个变量集合上都出现了比较大的误差，主要原因在于其模型假设的错误以及干预组对照组的样本大小不平衡。(3) 我们的方法相比于基准方法取得了最好的效果，因为我们的模型联合优化了样本权重和混淆变量权重，并且不需要模型假设。(4) 我们的方法在 V-INTERACTION 集合上取得了更好的效果，这表明了在引入观测变量的高阶项之后，混淆变量分布能够得到更好的平衡。

在表 4.3中，我们展示了 V-RAW 集合中经算法优化之后的混淆变量权重。从表中我们可以发现 1974 和 1975 年的收入状况和教育状况对于 1978 年的收入是非常重要的，但是种族对于收入的影响却很有限。这表明 1974 和 1975 年的收入状况和教育状况作为混淆变量时应该被优先平衡，因为它们对于结果影响非常大。

4.5.3.2　在线广告数据集

在在线广告的实验中，我们的数据来源于微信 APP[①]，数据收集于 2015 年 9 月期间。在微信朋友圈中，用户可以分享自己的状态，同时也可以收到来自好友的状态。广告商可以在朋友圈中插入自己的广告，形式就像一条好友的状态，每条广告都有喜欢（点赞）和不喜欢两种反馈，每个用户可以看到他的好友对于广告的点赞信息。

在本章中我们使用的广告数据是针对年轻女性的 LONGCHAMP 手袋广告[②]。这条广告收到了来自用户的 14,891 个喜欢和 93,108 个不喜欢。对于每个用户，我们有 56 维特征，包括：（1）诸如年龄性别等人口学属性（2）好友数量（3）设备的操作系统（苹果、安卓）（4）微信的一些设置诸如是否允许陌生人看朋友圈等。

① http://www.wechat.com/en/

② http://en.longchamp.com/en/womens-bags

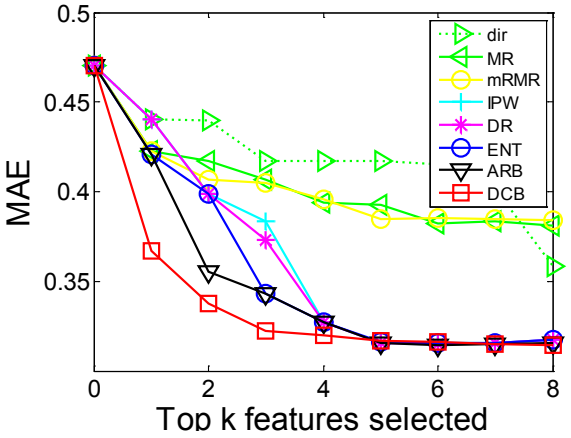


图 4.3　不同方法提取出的前 k 维特征在预测任务上的误差曲线

实验设定：在本实验中，我们将用户对于广告的反馈作为结果变量 Y。具体地，当用户 i 喜欢这个广告时 $Y_i = 1$，反之 $Y_i = 0$。我们分别将每一维用户特征当做干预变量 T，其余所有的特征设置为观测变量 $\mathbf{X}$，因此对于每一维特征我们都可以估计其对结果变量的因果效应 ATT。我们根据“近似 ATT 真实值”来调整算法和基准方法的参数。

基准方法和评价指标：在这个数据集上，我们并没有每个特征变量的真实因果效应。但我们对一件事很感兴趣，那就是根据我们的算法和基准方法得出的前 k 个重要的变量（以因果效应的绝对值大小排序），用于预测用户反馈行为时的效果好坏。在这里我们的基准方法除了以上介绍的其它因果推理方法之外，我们还比较了传统的基于相关性的特征选择方法 MRel (Maximum Relevance)[137] 和 mRMR (Maximum Relevance Minimum Redundancy)[138]。我们采用 MAE 作为评价指标，其定义如下：

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{Y}_i - Y_i|,$$

其中 m 为测试集用户的数量，$\hat{Y}_i$ 和 Y_i 表示用户 i 对广告的反馈的预测值和真实值。

实验结果：我们将实验结果画在了图 4.3中。从图中可以看到，我们的方法在选择不同数量的特征时都取得了最好的效果。同时，相比于其它基准方法我们的方法可以使用更少的特征取得最优的结果。一个主要的原因在于我们将混淆变量作了区分，通过更好地去除混淆偏差，我们可以更准确地估计每个变量的因果效应。另一个重要的发现是两种常见的特征选择方法的表现远远不如我们的方法，其误差甚至高于其它基于因果的方法。这是因为传统的基于相关性的方法无法应对训练集和测试集存在选择性偏差的情况，而基于因果的方法在一定程度上可以解决

这个问题。

实验结果表明，只要混淆变量引起的偏差问题能够很好地解决，通过估计变量的因果效应能够极大地提升预测任务的准确率。

4.6　本章小结

在本章中，我们将重点放在面向变量差异性和变量结构未知性，如何更准确地估计变量的因果效应。我们发现大多数基于加权的估计量抑或没有考虑混淆变量的区分，抑或需要正确的模型假设，导致了在高维变量的场景下表现不佳。为了解决这些问题，我们提出了混淆变量加权的概念并给出了理论的分析。我们提出了区分性混淆变量平衡算法，通过联合优化混淆变量权重和样本权重来估计因果效应。基于仿真数据集和真实数据集的大量实验表明我们的方法能够显著地超越当前最好的方法。同时我们的方法所选出的最重要的特征在广告数据集的预测任务上也取得了最好的效果。

表 4.4　仿真数据上的实验结果 1：*Bias* 指的是评估的因果效应和真实因果效应之间的误差绝对值，也就是 $Bias = |\widehat{ATT} - ATT|$。SD，MAE，和 RMSE 分别表示独立重复 100 次实验后，评估因果效应 $\widehat{ATE}$ 的标准偏差（standard deviations），平均绝对误差（mean absolute errors），和均方根误差（root mean square errors）。

设定 1: $T = T_{logit}, Y = Y_{linear}$ **and** $s_c = 1$													
	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$			$n = 5000, p = 50$			$n = 5000, p = 100$		
r_c	Estimator	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE
$r_c = 0.2$	$\widehat{ATT}_{dir}$	6.483 (3.460)	6.682	7.349	18.60 (8.859)	18.67	20.61	6.420 (2.050)	6.420	6.739	18.53 (5.148)	18.53	19.23
	$\widehat{ATT}_{IPW}$	2.220 (6.224)	4.866	6.609	8.365 (15.40)	14.47	17.52	1.907 (4.092)	3.648	4.514	8.033 (9.852)	10.52	12.71
	$\widehat{ATT}_{DR}$	0.118 (0.307)	0.253	0.329	1.591 (0.512)	1.591	1.672	0.059 (0.174)	0.145	0.183	1.446 (0.337)	1.446	1.485
	$\widehat{ATT}_{ENT}$	0.371 (0.477)	0.453	0.605	4.924 (3.167)	5.052	5.855	0.046 (0.254)	0.210	0.258	2.425 (1.229)	2.429	2.719
	$\widehat{ATT}_{ARB}$	0.074 (0.472)	0.376	0.477	0.868 (0.435)	0.881	0.971	0.027 (0.269)	0.217	0.270	0.365 (0.371)	0.447	0.520
	$\widehat{ATT}_{DCB}$	**0.014** (0.121)	**0.099**	**0.122**	**0.006** (0.119)	**0.101**	**0.119**	**0.001** (0.073)	**0.053**	**0.073**	**0.001** (0.085)	**0.067**	**0.085**
$r_c = 0.8$	$\widehat{ATT}_{dir}$	51.06 (3.725)	51.06	51.19	143.0 (9.389)	143.0	143.3	50.45 (1.900)	50.45	50.48	142.1 (5.647)	142.1	142.2
	$\widehat{ATT}_{IPW}$	29.99 (4.048)	29.99	30.26	98.24 (8.462)	98.24	98.60	29.38 (2.216)	29.38	29.46	96.86 (5.899)	96.86	97.04
	$\widehat{ATT}_{DR}$	0.345 (0.253)	0.367	0.428	4.492 (0.333)	4.492	4.504	0.338 (0.136)	0.338	0.365	4.306 (0.227)	4.306	4.312
	$\widehat{ATT}_{ENT}$	15.06 (1.745)	15.06	15.16	63.02 (4.551)	63.02	63.19	10.09 (1.473)	10.09	10.19	51.99 (3.206)	51.99	52.09
	$\widehat{ATT}_{ARB}$	0.231 (0.645)	0.553	0.685	2.909 (0.491)	2.909	2.951	0.189 (0.504)	0.428	0.538	2.259 (0.468)	2.259	2.307
	$\widehat{ATT}_{DCB}$	**0.003** (0.127)	**0.102**	**0.127**	**0.020** (0.135)	**0.114**	**0.136**	**0.003** (0.088)	**0.072**	**0.088**	**0.012** (0.088)	**0.073**	**0.089**
设定 2: $T = T_{logit}, Y = Y_{linear}$ **and** $r_c = 0.5$													
	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$			$n = 5000, p = 50$			$n = 5000, p = 100$		
s_c	Estimator	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE
$s_c = 0.2$	$\widehat{ATT}_{dir}$	11.80 (3.243)	11.80	12.24	43.38 (9.170)	43.38	44.34	11.53 (2.142)	11.53	11.73	42.64 (6.103)	42.64	43.07
	$\widehat{ATT}_{IPW}$	3.897 (2.759)	4.144	4.775	18.37 (8.317)	18.38	20.17	3.873 (2.055)	3.875	4.384	17.13 (5.971)	17.13	18.14
	$\widehat{ATT}_{DR}$	0.053 (0.150)	0.124	0.159	1.255 (0.265)	1.255	1.283	0.056 (0.104)	0.090	0.118	1.148 (0.180)	1.148	1.162
	$\widehat{ATT}_{ENT}$	0.023 (0.168)	0.128	0.170	0.174 (0.193)	0.208	0.260	**0.001** (0.116)	0.090	0.116	0.089 (0.119)	0.120	0.149
	$\widehat{ATT}_{ARB}$	**0.002** (0.170)	0.129	0.170	**0.011** (0.184)	0.151	0.185	0.004 (0.119)	0.094	0.120	0.006 (0.121)	0.093	0.122
	$\widehat{ATT}_{DCB}$	0.011 (0.107)	**0.086**	**0.107**	0.013 (0.098)	**0.080**	**0.099**	0.003 (0.065)	**0.053**	**0.065**	**0.004** (0.073)	**0.060**	**0.073**
$s_c = 0.8$	$\widehat{ATT}_{dir}$	22.81 (3.610)	22.81	23.09	69.28 (9.608)	69.28	69.94	21.91 (1.908)	21.91	21.99	68.72 (5.410)	68.72	68.93
	$\widehat{ATT}_{IPW}$	9.984 (4.878)	10.15	11.11	40.64 (12.48)	40.64	42.51	9.263 (3.615)	9.263	9.943	40.31 (7.185)	40.31	40.94
	$\widehat{ATT}_{DR}$	0.185 (0.256)	0.256	0.316	3.234 (0.449)	3.234	3.265	0.177 (0.166)	0.205	0.243	3.051 (0.245)	3.051	3.061
	$\widehat{ATT}_{ENT}$	2.805 (1.153)	2.805	3.033	23.53 (4.432)	23.53	23.94	0.742 (0.447)	0.759	0.866	15.97 (2.519)	15.97	16.16
	$\widehat{ATT}_{ARB}$	0.059 (0.564)	0.455	0.567	1.861 (0.491)	1.861	1.924	0.005 (0.408)	0.327	0.408	1.133 (0.451)	1.133	1.219
	$\widehat{ATT}_{DCB}$	**0.007** (0.124)	**0.102**	**0.124**	**0.015** (0.123)	**0.102**	**0.124**	**0.001** (0.083)	**0.067**	**0.083**	**0.017** (0.076)	**0.063**	**0.078**
设定 3: $T = T_{missp}, Y = Y_{nonlin}$ **and** $s_c = 1$													
	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$			$n = 5000, p = 50$			$n = 5000, p = 100$		
r_c	Estimator	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE
$r_c = 0.2$	$\widehat{ATT}_{dir}$	6.527 (5.367)	7.041	8.450	18.67 (14.04)	20.01	23.36	7.340 (3.425)	7.366	8.099	20.54 (9.992)	20.54	22.84
	$\widehat{ATT}_{IPW}$	5.061 (8.998)	8.542	10.32	17.31 (19.22)	21.90	25.86	6.707 (6.494)	7.934	9.336	19.81 (15.04)	21.79	24.87
	$\widehat{ATT}_{DR}$	6.334 (8.628)	8.562	10.70	23.65 (26.32)	29.16	35.38	6.493 (6.698)	7.637	9.329	23.44 (16.62)	24.77	28.74
	$\widehat{ATT}_{ENT}$	3.770 (2.166)	3.842	4.348	13.46 (5.854)	13.58	14.68	3.096 (1.285)	3.102	3.352	12.16 (3.585)	12.16	12.68
	$\widehat{ATT}_{ARB}$	0.643 (0.292)	0.647	0.706	3.757 (0.483)	3.757	3.788	0.512 (0.247)	0.517	0.569	3.288 (0.262)	3.288	3.299
	$\widehat{ATT}_{DCB}$	**0.016** (0.316)	**0.263**	**0.317**	**0.021** (0.364)	**0.294**	**0.365**	**0.017** (0.169)	**0.139**	**0.169**	**0.082** (0.214)	**0.183**	**0.230**
$r_c = 0.8$	$\widehat{ATT}_{dir}$	53.26 (5.308)	53.26	53.53	145.2 (13.47)	145.2	145.9	53.12 (3.673)	53.12	53.24	145.2 (9.247)	145.2	145.4
	$\widehat{ATT}_{IPW}$	39.46 (6.404)	39.46	39.97	113.0 (16.91)	113.0	114.3	39.04 (4.424)	39.04	39.29	111.7 (10.19)	111.7	112.1
	$\widehat{ATT}_{DR}$	15.12 (8.433)	15.40	17.31	34.07 (28.29)	37.09	44.28	14.26 (5.613)	14.28	15.33	30.92 (15.90)	31.70	34.77
	$\widehat{ATT}_{ENT}$	29.83 (1.795)	29.83	29.89	97.32 (6.507)	97.32	97.54	25.73 (1.155)	25.73	25.76	85.63 (3.114)	85.63	85.68
	$\widehat{ATT}_{ARB}$	1.342 (0.337)	1.342	1.384	7.440 (0.566)	7.440	7.462	1.102 (0.230)	1.102	1.126	6.526 (0.325)	6.526	6.535
	$\widehat{ATT}_{DCB}$	**0.076** (0.321)	**0.255**	**0.330**	**0.024** (0.388)	**0.298**	**0.389**	**0.003** (0.207)	**0.171**	**0.207**	**0.021** (0.304)	**0.248**	**0.305**
设定 4: $T = T_{missp}, Y = Y_{nonlin}$ **and** $r_c = 0.5$													
	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$			$n = 5000, p = 50$			$n = 5000, p = 100$		
s_c	Estimator	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE
$s_c = 0.2$	$\widehat{ATT}_{dir}$	18.01 (5.556)	18.01	18.84	59.49 (14.13)	59.49	61.15	18.01 (3.178)	18.01	18.29	60.34 (8.923)	60.34	60.99
	$\widehat{ATT}_{IPW}$	7.288 (6.605)	8.429	9.836	32.24 (19.66)	33.23	37.76	7.372 (4.505)	7.516	8.639	33.39 (12.87)	33.39	35.78
	$\widehat{ATT}_{DR}$	3.408 (5.953)	5.735	6.859	13.87 (21.90)	21.33	25.92	3.130 (4.146)	4.360	5.194	13.87 (12.53)	15.54	18.69
	$\widehat{ATT}_{ENT}$	1.812 (0.818)	1.812	1.988	25.54 (6.241)	25.54	26.29	0.273 (0.160)	0.282	0.317	14.49 (2.800)	14.49	14.76
	$\widehat{ATT}_{ARB}$	0.159 (0.254)	0.244	0.300	2.960 (0.385)	2.960	2.985	0.055 (0.150)	0.131	0.160	1.899 (0.241)	1.899	1.915
	$\widehat{ATT}_{DCB}$	**0.005** (0.223)	**0.178**	**0.223**	**0.011** (0.288)	**0.228**	**0.288**	**0.012** (0.120)	**0.095**	**0.120**	**0.025** (0.158)	**0.125**	**0.160**
$s_c = 0.8$	$\widehat{ATT}_{dir}$	24.58 (5.276)	24.58	25.14	72.30 (13.95)	72.30	73.63	24.10 (3.219)	24.10	24.31	71.20 (8.771)	71.20	71.74
	$\widehat{ATT}_{IPW}$	18.34 (6.819)	18.34	19.56	57.07 (18.02)	57.07	59.85	17.65 (4.755)	17.65	18.28	54.95 (9.861)	54.95	55.83
	$\widehat{ATT}_{DR}$	11.23 (8.757)	12.46	14.24	32.35 (26.22)	35.39	41.65	11.17 (5.492)	11.17	12.44	28.06 (14.24)	28.29	31.46
	$\widehat{ATT}_{ENT}$	12.88 (1.956)	12.88	13.03	48.40 (5.818)	48.40	48.75	10.46 (1.315)	10.46	10.55	40.79 (2.773)	40.79	40.88
	$\widehat{ATT}_{ARB}$	0.993 (0.343)	0.993	1.050	6.052 (0.525)	6.052	6.075	0.807 (0.255)	0.807	0.846	5.176 (0.279)	5.176	5.183
	$\widehat{ATT}_{DCB}$	**0.042** (0.310)	**0.246**	**0.313**	**0.023** (0.364)	**0.306**	**0.365**	**0.006** (0.211)	**0.167**	**0.211**	**0.013** (0.237)	**0.194**	**0.238**

表 4.5　仿真数据上的实验结果 2：*Bias* 指的是评估的因果效应和真实因果效应之间的误差绝对值，也就是 $Bias = |\widehat{ATT} - ATT|$。SD，MAE，和 RMSE 分别表示独立重复 100 次实验后，评估因果效应 $\widehat{ATE}$ 的标准偏差（standard deviations），平均绝对误差（mean absolute errors），和均方根误差（root mean square errors）。

设定 5: $T = T_{missp}$, $Y = Y_{linear}$ **and** $s_c = 1$													
	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$			$n = 5000, p = 50$			$n = 5000, p = 100$		
r_c	Estimator	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE
$r_c = 0.2$	$\widehat{ATT}_{dir}$	7.366 (2.857)	7.370	7.901	19.31 (10.17)	19.58	21.83	7.388 (2.182)	7.388	7.704	20.43 (5.003)	20.43	21.04
	$\widehat{ATT}_{IPW}$	4.335 (4.960)	5.350	6.587	11.45 (13.30)	14.54	17.55	3.132 (3.484)	3.764	4.685	12.76 (8.194)	13.11	15.17
	$\widehat{ATT}_{DR}$	0.163 (0.290)	0.266	0.333	2.309 (0.446)	2.309	2.352	0.163 (0.168)	0.185	0.234	2.265 (0.302)	2.265	2.285
	$\widehat{ATT}_{ENT}$	2.284 (1.232)	2.287	2.595	8.716 (3.705)	8.729	9.470	1.556 (0.789)	1.565	1.745	7.311 (2.202)	7.311	7.636
	$\widehat{ATT}_{ARB}$	0.077 (0.643)	0.538	0.648	1.724 (0.447)	1.724	1.781	0.094 (0.497)	0.404	0.506	1.437 (0.435)	1.437	1.501
	$\widehat{ATT}_{DCB}$	**0.005** (0.134)	**0.106**	**0.134**	**0.025** (0.117)	**0.092**	**0.120**	**0.003** (0.084)	**0.067**	**0.084**	**0.000** (0.067)	**0.052**	**0.067**
$r_c = 0.8$	$\widehat{ATT}_{dir}$	52.46 (3.347)	52.46	52.56	145.9 (8.598)	145.9	146.1	52.06 (1.963)	52.06	52.10	145.7 (5.380)	145.7	145.8
	$\widehat{ATT}_{IPW}$	35.31 (3.548)	35.31	35.49	105.3 (8.115)	105.3	105.6	34.51 (2.012)	34.51	34.57	104.5 (5.467)	104.5	104.6
	$\widehat{ATT}_{DR}$	0.437 (0.251)	0.442	0.504	4.885 (0.348)	4.885	4.897	0.396 (0.132)	0.396	0.417	4.649 (0.252)	4.649	4.656
	$\widehat{ATT}_{ENT}$	23.72 (1.416)	23.72	23.77	76.10 (3.331)	76.10	76.17	20.70 (1.059)	20.70	20.72	68.32 (2.304)	68.32	68.36
	$\widehat{ATT}_{ARB}$	0.357 (0.528)	0.514	0.637	3.534 (0.488)	3.534	3.567	0.276 (0.539)	0.457	0.605	3.034 (0.421)	3.034	3.063
	$\widehat{ATT}_{DCB}$	**0.005** (0.128)	**0.106**	**0.128**	**0.034** (0.124)	**0.105**	**0.129**	**0.004** (0.084)	**0.066**	**0.084**	**0.002** (0.086)	**0.068**	**0.086**
设定 6: $T = T_{missp}$, $Y = Y_{linear}$ **and** $r_c = 0.5$													
	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$			$n = 5000, p = 50$			$n = 5000, p = 100$		
s_c	Estimator	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE
$s_c = 0.2$	$\widehat{ATT}_{dir}$	18.00 (3.090)	18.00	18.26	58.80 (9.332)	58.80	59.54	17.70 (1.734)	17.70	17.79	59.63 (5.402)	59.63	59.88
	$\widehat{ATT}_{IPW}$	5.868 (3.710)	5.905	6.943	27.90 (10.52)	27.91	29.82	5.612 (2.314)	5.612	6.071	27.45 (6.236)	27.45	28.15
	$\widehat{ATT}_{DR}$	0.093 (0.197)	0.182	0.218	2.191 (0.347)	2.191	2.218	0.099 (0.125)	0.129	0.160	1.972 (0.206)	1.972	1.983
	$\widehat{ATT}_{ENT}$	0.106 (0.237)	0.215	0.260	5.948 (1.987)	5.948	6.271	0.041 (0.147)	0.122	0.153	0.540 (0.279)	0.540	0.607
	$\widehat{ATT}_{ARB}$	0.007 (0.237)	0.190	0.237	0.444 (0.383)	0.484	0.586	0.002 (0.148)	0.118	0.148	0.017 (0.230)	0.190	0.231
	$\widehat{ATT}_{DCB}$	**0.003** (0.099)	**0.080**	**0.099**	**0.007** (0.124)	**0.098**	**0.124**	**0.002** (0.070)	**0.057**	**0.070**	**0.002** (0.075)	**0.063**	**0.075**
$s_c = 0.8$	$\widehat{ATT}_{dir}$	23.99 (3.322)	23.99	24.22	71.72 (8.267)	71.72	72.19	24.25 (1.828)	24.25	24.32	72.19 (5.520)	72.19	72.40
	$\widehat{ATT}_{IPW}$	14.18 (3.898)	14.18	14.71	47.86 (9.081)	47.86	48.72	14.00 (2.514)	14.00	14.23	47.90 (6.710)	47.90	48.37
	$\widehat{ATT}_{DR}$	0.356 (0.244)	0.367	0.431	3.910 (0.466)	3.910	3.937	0.280 (0.141)	0.282	0.314	3.830 (0.268)	3.830	3.839
	$\widehat{ATT}_{ENT}$	9.040 (1.216)	9.040	9.122	35.08 (3.207)	35.08	35.22	6.990 (0.981)	6.990	7.058	30.22 (2.387)	30.22	30.32
	$\widehat{ATT}_{ARB}$	0.214 (0.579)	0.494	0.617	2.756 (0.528)	2.756	2.806	0.110 (0.530)	0.439	0.542	2.417 (0.420)	2.417	2.454
	$\widehat{ATT}_{DCB}$	**0.003** (0.123)	**0.099**	**0.123**	**0.013** (0.123)	**0.098**	**0.123**	**0.000** (0.073)	**0.057**	**0.073**	**0.003** (0.077)	**0.065**	**0.077**
设定 7: $T = T_{logit}$, $Y = Y_{nonlin}$ **and** $s_c = 1$													
	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$			$n = 5000, p = 50$			$n = 5000, p = 100$		
r_c	Estimator	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE
$r_c = 0.2$	$\widehat{ATT}_{dir}$	6.639 (5.061)	7.128	8.348	18.60 (14.50)	19.95	23.59	6.301 (3.417)	6.403	7.168	16.44 (9.478)	16.80	18.98
	$\widehat{ATT}_{IPW}$	3.132 (10.15)	8.573	10.62	13.00 (26.69)	22.22	29.69	2.793 (6.610)	5.872	7.176	11.82 (15.47)	15.43	19.47
	$\widehat{ATT}_{DR}$	1.646 (8.908)	7.640	9.059	12.98 (25.93)	22.17	29.00	2.516 (6.266)	5.410	6.752	12.90 (15.89)	15.99	20.47
	$\widehat{ATT}_{ENT}$	1.908 (1.659)	2.062	2.529	10.88 (6.590)	11.19	12.72	0.780 (0.672)	0.835	1.029	7.509 (3.069)	7.552	8.112
	$\widehat{ATT}_{ARB}$	0.310 (0.305)	0.371	0.435	2.854 (0.464)	2.854	2.892	0.150 (0.228)	0.215	0.273	2.100 (0.267)	2.100	2.117
	$\widehat{ATT}_{DCB}$	**0.000** (0.251)	**0.204**	**0.251**	**0.004** (0.314)	**0.257**	**0.314**	**0.015** (0.160)	**0.129**	**0.160**	**0.023** (0.175)	**0.139**	**0.176**
$r_c = 0.8$	$\widehat{ATT}_{dir}$	49.87 (5.283)	49.87	50.15	143.6 (15.26)	143.6	144.4	50.13 (3.167)	50.13	50.23	143.5 (10.02)	143.5	143.9
	$\widehat{ATT}_{IPW}$	31.81 (6.563)	31.81	32.48	105.5 (16.47)	105.5	106.8	32.58 (4.659)	32.58	32.91	104.6 (11.58)	104.6	105.2
	$\widehat{ATT}_{DR}$	10.86 (8.339)	11.57	13.69	24.45 (22.87)	28.56	33.48	11.63 (5.477)	11.67	12.86	28.23 (15.73)	29.17	32.32
	$\widehat{ATT}_{ENT}$	23.27 (2.175)	23.27	23.37	89.07 (5.759)	89.07	89.26	17.79 (1.395)	17.79	17.85	73.90 (3.717)	73.90	74.00
	$\widehat{ATT}_{ARB}$	1.032 (0.367)	1.032	1.096	6.783 (0.529)	6.783	6.804	0.774 (0.274)	0.774	0.821	5.697 (0.342)	5.697	5.707
	$\widehat{ATT}_{DCB}$	**0.033** (0.308)	**0.246**	**0.310**	**0.040** (0.395)	**0.324**	**0.397**	**0.026** (0.185)	**0.147**	**0.186**	**0.156** (0.251)	**0.246**	**0.295**
设定 8: $T = T_{logit}$, $Y = Y_{nonlin}$ **and** $r_c = 0.5$													
	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$			$n = 5000, p = 50$			$n = 5000, p = 100$		
s_c	Estimator	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE	*Bias* (SD)	MAE	RMSE
$s_c = 0.2$	$\widehat{ATT}_{dir}$	10.85 (5.138)	10.87	12.01	41.68 (13.38)	41.68	43.77	11.53 (3.348)	11.53	12.01	41.02 (9.734)	41.02	42.16
	$\widehat{ATT}_{IPW}$	3.970 (4.980)	5.027	6.369	18.16 (14.94)	19.72	23.51	4.525 (3.891)	4.988	5.968	17.96 (10.84)	18.03	20.98
	$\widehat{ATT}_{DR}$	1.175 (4.624)	3.740	4.771	3.810 (15.86)	12.57	16.31	1.482 (3.303)	2.970	3.620	4.847 (9.978)	8.739	11.09
	$\widehat{ATT}_{ENT}$	0.154 (0.188)	0.203	0.243	9.315 (3.602)	9.315	9.987	0.101 (0.124)	0.133	0.160	2.035 (0.733)	2.035	2.163
	$\widehat{ATT}_{ARB}$	0.009 (0.172)	0.139	0.172	1.035 (0.302)	1.035	1.078	0.002 (0.113)	0.092	0.113	0.406 (0.151)	0.406	0.433
	$\widehat{ATT}_{DCB}$	**0.004** (0.159)	**0.121**	**0.159**	**0.006** (0.192)	**0.152**	**0.192**	**0.008** (0.112)	**0.089**	**0.112**	**0.015** (0.142)	**0.116**	**0.143**
$s_c = 0.8$	$\widehat{ATT}_{dir}$	21.48 (5.483)	21.48	22.17	71.82 (14.83)	71.82	73.33	21.98 (3.225)	21.98	22.21	69.41 (9.158)	69.41	70.01
	$\widehat{ATT}_{IPW}$	10.64 (9.112)	11.94	14.01	49.78 (22.10)	50.77	54.47	12.22 (5.276)	12.23	13.31	47.33 (11.83)	47.33	48.79
	$\widehat{ATT}_{DR}$	5.907 (8.284)	8.353	10.17	21.41 (22.76)	25.53	31.24	6.601 (6.164)	7.591	9.031	21.09 (14.99)	22.45	25.87
	$\widehat{ATT}_{ENT}$	7.549 (2.105)	7.549	7.837	40.91 (7.127)	40.91	41.53	4.448 (1.021)	4.448	4.564	30.59 (3.313)	30.59	30.76
	$\widehat{ATT}_{ARB}$	0.596 (0.312)	0.607	0.672	4.843 (0.449)	4.843	4.864	0.356 (0.235)	0.369	0.426	3.884 (0.288)	3.884	3.894
	$\widehat{ATT}_{DCB}$	**0.012** (0.282)	**0.221**	**0.282**	**0.042** (0.325)	**0.253**	**0.327**	**0.007** (0.153)	**0.121**	**0.154**	**0.014** (0.188)	**0.145**	**0.188**

第 5 章　因果约束的稳定预测

在许多重要的机器学习应用中，用于学习概率分类器的训练数据分布不同于分类器将用于进行预测的测试数据分布。传统方法通过使用测试和训练数据之间的样本密度比率加权训练数据来纠正分布变化。然而，在许多应用中，我们很难知道未来测试数据分布的先验知识。最近，一些学者提出了通过学习潜在的因果结构来解决这种分布变化的方法，但是这些方法依赖于由多个训练数据集的多样性，并且它们算法复杂度使得其很难应用于高维数据。在本章中，我们提出了一种新的深度全局平衡回归（Deep Global Balancing Regression，DGBR）算法，以联合优化用于特征表征和选择的深度自动编码器模型和用于在未知环境中进行稳定预测的全局平衡模型。全局平衡模型构建平衡权重，通过估计特征对结果变量的因果效应（保持固定所有其它特征），来识别特征和结果之间稳定的因果关系。深度自动编码器模型旨在降低特征空间的维度，从而使全局平衡更容易。在理论上和经验实验上，我们都证明了我们的算法可以在未知环境中进行稳定的预测。在仿真和真实数据集上的大量实验表明，我们的算法可以在未知环境中进行稳定预测，预测效果优于所有的基准方法。

5.1　本章引言

使用在训练数据集上学习的模型来预测未知测试数据的结果是常见的统计问题。当测试数据和训练数据来自相同的分布时，很多机器学习和数据挖掘方法都能非常成功地对未来测试数据的结果进行预测。然而，在很多实际应用中，我们很难保证未来的测试数据集的分布跟我们训练模型时用的训练数据集来自于同一分布，也就是说未来测试数据集的特征联合分布跟训练数据集的分布不一致。这里我们定义特征的联合分布为环境。因此，我们需要打破训练数据集和测试数据集之间的独立同分布假设，开发对测试环境变化具有鲁棒性的预测算法，实现对未来所有可能的测试环境进行稳定预测。

最近，很多方法被提出来解决训练数据集和测试数据集分布不一致的问题[64,66–68,70]。这些方法的主要思想是学习训练数据和测试数据之间的密度比率来加权训练数据样本，使得训练数据的分布与测试数据分布一致。这些方法对于校正训练数据和测试数据之间的特征分布变化具有良好的性能，但是在估计密度比率时它们需要测试数据的分布作为先验知识。

面对未知测试数据的稳定预测，一些研究人员提出通过在多个训练数据集中

进行训练来学习多个数据集上的不变性，实现稳定预测。Peters 等人[86] 提出了一种识别因果特征不变性的算法，而 Rojas-Carulla 等人[87] 提出了一个因果变换框架来学习不变结构。类似地，域泛化方法[79] 尝试学习数据特征表征的不变性。这些方法的性能依赖于其多个训练数据的多样性，并且它们无法解决所有训练数据中都未出现的分布变化。此外，这些方法中的大多数都非常复杂，在最坏的情况下，训练复杂度随着特征空间的维度呈指数增长，使得其很难运用于高维数据应用中。

在本章中，我们提出利用变量之间的因果关系来约束预测模型的学习，从而解决未知测试环境的挑战。为了解决稳定预测问题，我们假设在给定所有观测特征的条件下结果变量的条件分布在所有环境中都是稳定（不变）的。此外，我们假设所有的观测特征可以分为两类：对于第一类特征，结果变量的条件期望对其具有非零依赖性，也就是说这一类特征对结果变量有因果效应，我们称之为“因果”变量或稳定特征。例如，耳朵，鼻子和胡须是猫的因果特征，其在可以拍摄动物图像的不同环境中是稳定的。第二类特征被称为非因果变量或者非稳定特征，它们是与因果变量，结果变量中一者或两者相关的变量，但它们本身不会对结果产生因果影响。这里，我们没有先验知识去区分哪些变量是因果变量，哪些变量不是。

在这种设定下，提高预测算法稳定性的一种可能方法是分别隔离计算每个单独特征对结果变量的影响。在因果推理文献中，我们发现可以通过变量平衡算法来隔离变量，并且评估每个单独变量对结果的因果效应。但因果推理的方法与预测推理模型的的优化目标非常不同，因果推理方法主要研究的是基于训练数据评估干预变量对结果变量的因果效应，而不是预测。这些方法适用于分析人员知道哪个变量对结果变量具有因果效应，所有分析的重点在于控制其它变量（混淆变量）对干预变量和结果变量的影响。实际上，只有控制住了混淆变量的影响之后，干预变量和结果变量之间的关系才能被解释为因果效应。在存在混淆变量的情况下，一种有效的因果效应评估算法是变量平衡算法。这类算法通过学习样本权重，并利用样本权重加权样本来使得干预组和对照组样本的混淆变量分布达到一致。它们要么使用倾向值[24,30,39,42,134]，要么直接优化平衡权重[50,52,53,139]。这些方法提供了一种在观察性研究中评估少量干预变量对结果变量的因果效应的有效方法，但是他们无法处理可能存在许多因果变量并且分析人员不知道哪些变量是因果变量的实际场景，而且现有的变量平衡方法不能很直接地扩展到一般的稳定预测问题。

受到因果推理文献中变量平衡方法的启发，我们提出了一种用于稳定预测的深度全局平衡回归（Deep Global Balancing Regression，DGBR）算法。算法的框架如图 5.2所示，它由三个（联合优化的）子模型组成：（1）深度自动编码模型，用于学习特征在低维度的表征来降低特征的维数；（2）全局变量平衡模型，用于学

习样本权重使得变量之间能相互隔离（变量之间不相互影响）。通过变量隔离学习出来的样本权重加权样本，每个变量与结果变量之间的相关性就只来源于其因果相关；以及（3）稳定学习模型，利用来自于深度自动编码模型的特征低维表征和全局变量平衡模型的样本权重来恢复变量与结果变量之间的因果关系，并学习稳定预测模型。

由于我们提出的算法考虑了通过变量平衡来恢复预测变量与结果变量之间的因果关系，而因果关系比关联关系在所有的环境中更稳定。因此我们算法对未知测试环境的预测能力比起那些纯预测模型更稳定。通过经验实验和理论分析，我们证明了我们算法对于未知的测试环境能够实现稳定预测。在仿真数据集和真实数据集上，我们进行了大量实验。实验结果表明，在稳定预测问题中，我们方法的稳定性优于所有的基准方法。

在该章节中，我们的贡献主要包括以下几点：

- 我们研究了未知环境中稳定预测的问题，其中测试数据的分布可能与训练数据分布非常不同，而且在模型训练期间不可知。
- 我们提出了新颖的 DGBR 算法，可用于联合优化深度自动编码模型（用于降维）和全局变量平衡模型（用于估计因果效应），以及同时解决稳定的预测问题。
- 我们对我们提出的算法进行理论分析，证明我们的算法可以通过全局平衡在未知环境中进行稳定的预测。
- 我们 DGBR 算法的优点在仿真数据集和真实数据集上都得到了验证。

本章的其余部分安排如下。第 5.2 节回顾了相关工作。在第 5.3 节中，我们给出问题的描述并介绍我们的 DGBR 算法。第 5.4 节给出了我们算法的优化和讨论。第 5.5 节给出了实验结果。最后，第 5.6 节小结。

5.2　相关工作

在本章中，我们介绍之前的相关工作，包括协变量偏移，变量平衡和不变性学习。

协变量偏移[64] 主要研究的问题在于训练数据分布与测试数据分布不同。为了纠正分布差异，文章[64] 通过测试数据中的样本密度与训练数据中的样本密度之比来加权训练样本，使得训练数据的分布和测试数据分布一致。之后，很多方法被提出用来估计密度比，包括判别估计[66]，内核均值匹配[67]，最大熵方法[68]，最小最大优化算法[69] 和强健偏差意识算法[70]。这些方法对于校正训练数据和测试数据之间的协变量分布不一致具有良好的性能，但是在估计密度比率时它们需要测试

数据的分布作为先验知识。相比之下，我们主要的研究问题是未知测试环境中的稳定预测。

控制混淆变量是观察性研究中评估因果效应的关键挑战，很多变量平衡的方法[30,50–53,134,139,140] 被提出用来控制混淆变量的影响。在一篇开创性的论文中，Rosenbaum 和 Rubin[30] 建议在观测数据通利用倾向值的倒数来加权样本，从而达到变量平衡的目的。Kuang 等人[134] 提出了一种数据驱动的变量分离算法来实现混淆变量分离和平衡。Li 等人[51] 通过匹配非线性表示来实现变量平衡。Hainmueller[53] 通过变量的矩来实现变量分布平衡，并提出了熵平衡方法。Athey 等人[50] 提出了近似残差平衡算法，该算法将使用 LASSO 的结果建模与构建的平衡权重相结合，以近似平衡干预组和对照组之间的协变量。Kuang 等人[139] 通过联合优化样本权重和变量权重来实现对混淆变量的选择和区分，实现了混淆变量区分性平衡，大幅度提升了变量平衡的效果。这些方法提供了一种有效的方法来估计观察性研究中的因果效应，但它们仅限于估计一个变量的因果效应，而不是针对具有许多因果变量的情况而设计的，导致其很难直接运用于稳定预测问题。

最近，有些学者提出通过不变学习方法来实现对未知测试数据集进行稳定预测。Peters 等人[86] 提出了一种稳定预测算法，通过探索多个训练数据集的结果条件分布的不变性来识别预测变量与结果之间的不变性。Rojas-Carulla 等人[87] 提出了一个因果转移框架，用于识别对结果变量预测能力不变的预测变量，然后用于稳定预测。类似地，域泛化[79] 方法通过学习多个训练集上数据特征表征的不变性来实现对未知预测数据集的预测。原则上，不变学习方法可以很好地运用于对未知测试数据集的预测，但这些方法的性能依赖于其多个训练数据的多样性，并且它们无法解决多个数据集中都未出现的分布变化。

5.3　稳定预测问题

让 $\mathcal{X}$ 表示观察到的特征的空间，$\mathcal{Y}$ 表示结果空间。为简单起见，我们考虑特征具有有限支持的情况，所有特征在没有损失一般性的情况下可以表示为一组二进制特征：$\mathcal{X} = \{0,1\}^p$。我们假设结果空间也是二进制的情况：$\mathcal{Y} = \{0,1\}$。我们将**环境**定义为 $\mathcal{X} \times \mathcal{Y}$ 上的联合分布 P_{XY}，并让 $\mathcal{E}$ 表示所有环境的集合。在 $e \in \mathcal{E}$ 的每个环境中，我们有数据集 $D^e = (\mathbf{X}^e, Y^e)$，其中 $\mathbf{X}^e \in \mathcal{X}$ 是预测变量，$Y^e \in \mathcal{Y}$ 是结果变量。特征和结果的联合分布 P_{XY} 可能因环境而异，也就是说：对于所有 $e, e' \in \mathcal{E}$，如果 $e \neq e'$，则 $P^e_{XY} \neq P^{e'}_{XY}$ 。

在本章中，我们的目标是学习一种预测模型可以在未知环境中进行稳定预测。在给出问题定义之前，我们首先定义预测模型在各种环境下的平均预测误差

Average_Error 和预测稳定性误差 *Stability_Error*：

$$Average\_Error=\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}Error(D^e), \tag{5-1}$$

$$Stability\_Error=\sqrt{\frac{1}{|\mathcal{E}|-1}\sum_{e\in\mathcal{E}}(Error(D^e)-Average\_Error)^2}, \tag{5-2}$$

其中 $|\mathcal{E}|$ 表示环境的数量，$Error(D^e)$ 表示预测模型在环境 e 中的数据集 D^e 上的预测误差。

在本章中，我们通过 *Stability_Error* 定义模型的稳定性[141]。预测模型的 *Stability_Error* 越小，表明模型的预测能力更稳定。然后，我们定义稳定预测问题如下：

问题 5.1 (**稳定预测**)： **给定**一个训练环境 $e \in \mathcal{E}$ 以及相应的训练数据集 $D^e = (\mathbf{X}^e, Y^e)$，任务是 **学习**一个预测模型，用于预测未知环境 $\mathcal{E}$ 下的结果，不仅需要 *Average_Error* 小，还需要 *Stability_Error* 小。

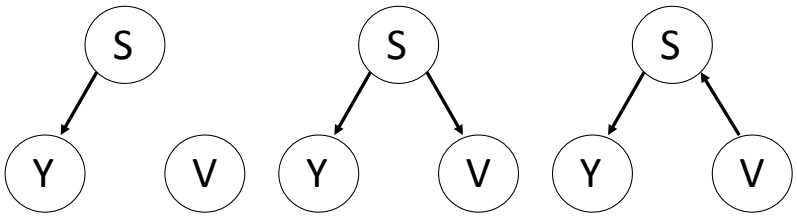
假设 $\mathbf{X} = \{\mathbf{S}, \mathbf{V}\}$。我们将 $\mathbf{S}$ 定义为稳定特征，定义其它特征 $\mathbf{V} = \mathbf{X}\backslash\mathbf{S}$ 为不稳定特征。以下假设给出了它们的定义属性：

假设 5.1： 存在概率质量函数 $P(y|s)$，使得对于所有环境 $e \in \mathcal{E}$, $Pr(Y^e = y|\mathbf{S}^e = s, \mathbf{V}^e = v) = Pr(Y^e = y|\mathbf{S}^e = s) = P(y|s)$.

基于假设 5.1，我们知道结果变量 Y 的值完全由稳定特征 $\mathbf{S}$ 决定，跟不稳定特征 $\mathbf{V}$ 没有任何关系。通过假设 5.1，我们可以通过构建一个学习稳定函数 $P(y|s)$ 的模型来解决稳定预测问题。为了更好地理解假设 5.1表达的内容，在不失一般性的情况下，我们可以在环境 e 中使用稳定特征 s 构建结果单元 i 的生成模型，其中 $h(\cdot)$ 是一个已知函数来解释 Y 的离散性：

$$Y_i^e(s) = h(g(s) + \epsilon_{si}^e)和Y_i^e = Y_i^e(\mathbf{S}_i) = h(g(\mathbf{S}_i) + \epsilon_{\mathbf{S}_i i}^e).$$

$Y_i^e(s)$ 是环境 e 中样本 i，当输入等于 s 的结果。如果我们允许 ϵ_{si}^e 以任意方式与样本的特征 $\mathbf{X}_i$ 相关联，假设 5.1可能会不成立，例如，如果 $\mathbf{V}_i^e$ 与 ϵ_{si}^e 正相关，然后 $\mathbf{V}_i^e$ 值较高的样本将高于 Y_i^e 的平均值，因此 $\mathbf{V}_i^e$ 在给定环境中将是一个对结果变量有用的预测特征，但这种关系可能因环境而异，从而导致不稳定。如果我们强加一个条件，例如，对于每个 s，在给定 $\mathbf{S}_i^e$ 的条件下，ϵ_{si}^e 独立于 $\mathbf{V}_i^e$。基于这个假设



(a) $\mathbf{S} \perp \mathbf{V}$　　(b) $\mathbf{S} \rightarrow \mathbf{V}$　　(c) $\mathbf{V} \rightarrow \mathbf{S}$

图 5.1　稳定特征 $\mathbf{S}$，不稳定特征 $\mathbf{V}$，和结果变量 Y 之间的三种结构图。

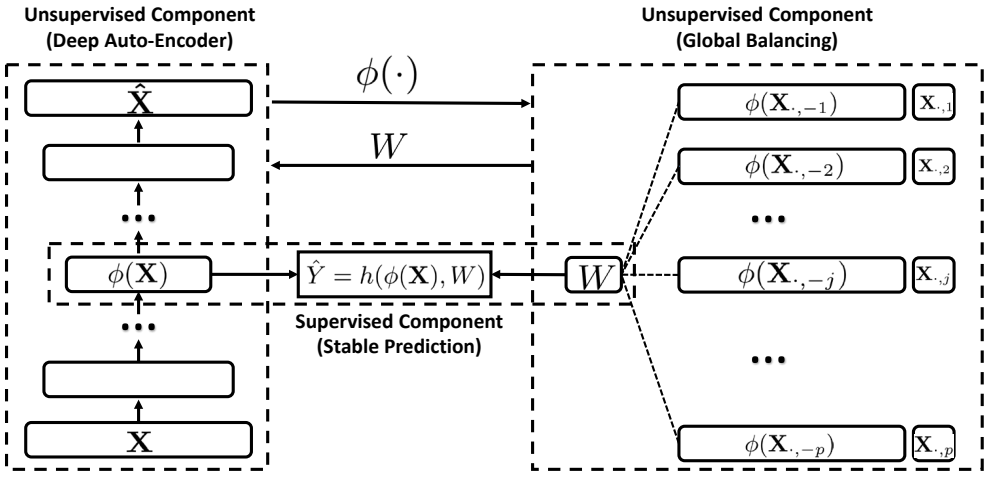


图 5.2　我们提出的 DGBR 算法框架。

下，当给定了 $\mathbf{S}_i^e$，$\mathbf{V}_i^e$ 就对预测结果变量没有任何作用了。要是我们强加第二个条件，例如，对于每个 s，在给定 $\mathbf{V}_i^e$ 的条件下，ϵ_{si}^e 独立于 $\mathbf{S}_i^e$。在这个假设下，ϵ_{si}^e 的在所有环境中的不稳定分布不会影响 $Pr(Y^e = y|\mathbf{S}^e = s\mathbf{V}^e = v)$。保持第一个条件，第二个条件不仅能保证假设 5.1的成立，而且可以帮助我们通过采用因果分析的技术实现函数 $g(\cdot)$ 的一致估计。在事先不知道哪些特征是稳定的，哪些特征是不稳定的情况下，我们在本章中提出了估算 g 的方法，以实现稳定预测。我们还观察到一个更强但更简单的条件可以取代第二个条件来保证假设5.1成立，即 ϵ_{si}^e 的分布不随 $\{e, s\}$ 的变化而改变。图5.1说明了预测变量 $\mathbf{X^e} = \{\mathbf{S}^e\mathbf{V}^e\}$ 和结果变量 Y^e 在符合假设5.1下的三种关系，包括 $\mathbf{S} \perp \mathbf{V}$，$\mathbf{S} \rightarrow \mathbf{V}$ 和 $\mathbf{V} \rightarrow \mathbf{S}$。

5.4　因果约束的稳定推理模型

5.4.1　模型框架和算法

我们提出深度全局平衡回归（Deep Global Balancing Regression，DGBR）算法来识别稳定特征并捕获其与结果变量之间的结构以进行稳定预测。框架具体细节如图 5.2所示。为了识别稳定的特征，我们提出了一个全局平衡模型。该模型通过

变量平衡来学习全局样本权重，通过样本重新加权来评估每个特征对结果变量的因果效应。由于稳定特征和非稳定特征之间最本质的区别是稳定特征对结果变量有因果效应，而非稳定特征没有。因此通过全局平衡模型，我们可以初步识别观测变量中的稳定特征。为了捕获稳定特征和结果变量之间的非线性结构，我们采用深度自动编码模型。该模型由多个非线性映射函数组成，将输入特征映射到非线性和低维空间。同时，将输入特征映射到低维空间可以降低全局平衡模型的复杂度。最后，基于全局平衡模型学来的全局样本权重和深度自动编码器模型学来的低维非线性表征，我们提出使用正则化回归学习，实现对结果变量的稳定预测模型。

5.4.1.1　全局平衡回归算法

在本章中，我们将介绍如何构建全局平衡权重。为了自成一体，我们简要回顾一下变量平衡技术的关键思想。变量平衡技术通常用于观察性研究中的因果效应评估。在观察性研究中，由于干预状态的非随机分配，导致干预组和对照组之间的协变量分布不一致。变量分布不一致会使得因果效应评估出现误差，因此，为了更准确地评估因果效应，我们需要平衡干预组和对照组之间的变量分布。现有大多数变量平衡方法利用变量的矩来代表其分布，并通过调整样本权重 W 来平衡它们在干预组和对照组之间的平衡：

$$W = \arg\min_{W} \left\| \frac{\sum_{i:T_i=1} W_i \cdot \mathbf{X}_i}{\sum_{i:T_i=1} W_i} - \frac{\sum_{i:T_i=0} W_i \cdot \mathbf{X}_i}{\sum_{i:T_i=0} W_i} \right\|_2^2. \tag{5-3}$$

给定一个干预变量 T，$\frac{\sum_{i:T_i=1} W_i \cdot \mathbf{X}_i}{\sum_{i:T_i=1} W_i}$ 和 $\frac{\sum_{i:T_i=0} W_i \cdot \mathbf{X}_i}{\sum_{i:T_i=0} W_i}$ 表示样本加权后干预组 ($T = 1$) 和对照组 ($T = 0$) 中变量 $\mathbf{X}$ 的一阶矩（当然，可以同时平衡变量的二阶矩或更高阶矩）。通过从公式 (5-3) 中学习的 W 重新加权样本，我们可以通过比较干预组和对照组之间的 Y 的平均差异来估计干预变量对结果变量的因果效应。在高维问题中，近似平衡可用于在一些额外假设下进行稳定地因果效应评估[50]。

在低维度中，可以采用类似的方法来估计 $Pr(Y = y|\mathbf{X} = x)$。但是，当 p 很大时，可能没有足够的数据这样做。因此近似变量平衡技术可能在实践中表现更好，并且帮助从高维度 $\mathbf{X}$ 中识别出稳定特征。

基于近似平衡技术，我们提出全局平衡正则因子来学习全局样本权重。具体地，我们先后将每个变量视为干预变量，并通过全局样本权重来平衡其它所有变

量。我们提出的全局平衡正则因子表示如下：

$$\sum_{j=1}^{p}\left\|\frac{\mathbf{X}_{\cdot,-j}^{T}\cdot(W\odot\mathbf{X}_{\cdot,j})}{W^{T}\cdot\mathbf{X}_{\cdot,j}}-\frac{\mathbf{X}_{\cdot,-j}^{T}\cdot(W\odot(1-\mathbf{X}_{\cdot,j}))}{W^{T}\cdot(1-\mathbf{X}_{\cdot,j})}\right\|_{2}^{2}, \tag{5-4}$$

其中 W 是全局样本权重，$\mathbf{X}_{\cdot j}$ 是 $\mathbf{X}$ 中的第 j 个变量，$\mathbf{X}_{\cdot -j}=\mathbf{X}\backslash\mathbf{X}_{\cdot j}$ 表示删除 $\mathbf{X}$ 中的第 j 个变量后剩下的所有变量[①]。公式（5-4）中的被加数表示当把变量 $\mathbf{X}_{\cdot j}$ 视为干预变量时，其余变量的分布不平衡带来的损失。$\odot$ 表示哈达玛积（Hadamard Product）。注意，在公式 (5-4) 中，我们只考虑变量的一阶矩，高阶矩也可以很轻易地引入到我们目标函数。

通过从公式 (5-4) 中学习 W 的样本重新加权，我们可以通过检查 Y 和 $\mathbf{X}$ 之间是否存在任何关联来识别稳定特征 $\mathbf{S}$。因为，正如我们在下面所示，在样本重新加权后，只有稳定的特征与 Y 相关。

基于传统的逻辑回归模型，利用公式 (5-4) 中的全局平衡正则化因子，我们提出了一种全局平衡回归 (Global Balancing Regression，GBR) 算法。该算法通过联合优化全局样本权重 W 和回归系数 β 进行稳定预测模型的学习：

$$\begin{aligned}
\min\quad & \sum_{i=1}^{n}W_{i}\cdot\log(1+\exp((1-2Y_{i})\cdot(\mathbf{X}_{i}\beta))), \\
s.t.\quad & \sum_{j=1}^{p}\left\|\frac{\mathbf{X}_{\cdot,-j}^{T}\cdot(W\odot\mathbf{X}_{\cdot,j})}{W^{T}\cdot\mathbf{X}_{\cdot,j}}-\frac{\mathbf{X}_{\cdot,-j}^{T}\cdot(W\odot(1-\mathbf{X}_{\cdot,j}))}{W^{T}\cdot(1-\mathbf{X}_{\cdot,j})}\right\|_{2}^{2}\leq\lambda_{1},\quad W\geq 0, \\
& \|W\|_{2}^{2}\leq\lambda_{2},\ \ \|\beta\|_{2}^{2}\leq\lambda_{3},\ \ \|\beta\|_{1}\leq\lambda_{4},\ \ (\sum_{k=1}^{n}W_{k}-1)^{2}\leq\lambda_{5}
\end{aligned} \tag{5-5}$$

其中 $\mathbf{X}_i$ 表示 $\mathbf{X}$ 中的第 i 个样本，$\sum_{i=1}^{n}W_i\cdot\log(1+\exp((1-2Y_i)\cdot(\mathbf{X}_i\beta)))$ 是逻辑回归模型的加权损失。约束项 $W\geq 0$ 约束每个样本的权重为非负数。使用范数 $\|W\|_2^2\leq\lambda_2$，我们可以减少样本权重的方差。范数 $\|\beta\|_2^2\leq\lambda_3$ 和 $\|\beta\|_1\leq\lambda_4$ 有助于避免预测模型过度拟合。约束项 $(\sum_{k=1}^{n}W_k-1)^2\leq\lambda_5$ 能帮助避免所有样本权重为 0。

5.4.1.2　深度全局平衡回归算法

方程 (5-5) 中提出的 GBR 算法可以帮助识别稳定的特征并进行稳定的预测。但是高维数据和非线性函数会给 GBR 算法带来全新的挑战，高维数据使得 GBR 算法复杂度非常高，而且 GBR 算法不能评估变量之间的非线性关系。

① 我们在实验中通过在 $\mathbf{X}$ 中设置第 j 个变量的值为 0 来获得 $\mathbf{X}_{-j}$。

为了应对高维和非线性的挑战，我们提出了深度全局平衡回归（DGBR）算法。该算法在全局平衡回归算法（GBR）的基础上增加了深度自动编码模型，用于特征非线性降维。遵循标准方法[142]，深度自动编码模型由多个非线性映射函数组成，将输入数据映射到低维空间，同时捕获底层特征交互信息。深度自动编码模型是一种无监督模型，由编码器和解码器两部分组成。编码器将输入数据映射到低维表征，而解码器基于低维表征重建原始输入特征。给定输入特征 $\mathbf{X}_i$，深度自动编码模型的每个隐藏层的结构表示如下：

$$\begin{aligned}\phi(\mathbf{X}_i)^{(1)} &= \sigma(\mathbf{A}^{(1)}\mathbf{X}_i + b^{(1)}) \\ \phi(\mathbf{X}_i)^{(k)} &= \sigma(\mathbf{A}^{(k)}\phi(\mathbf{X}_i)^{(k-1)} + b^{(k)}), k = 2, \cdots, K\end{aligned}$$

其中 K 是隐藏层的数量。$\mathbf{A}^{(k)}$ 和 $b^{(k)}$ 是分别 k 层的回归系数矩阵和截断值。$\sigma(\cdot)$ 表示非线性激活函数[①]。

在获得表示 $\phi(\mathbf{X}_i)^{(K)}$ 之后，我们可以通过利用参数 $\hat{\mathbf{A}}^{(k)}$ 和 $\hat{b}^{(k)}$ 和反转编码器的计算过程来获得输入特征的重构 $\hat{\mathbf{X}}_i$。深度自动编码模型的目标是最小化输入 $\mathbf{X}_i$ 和重建 $\hat{\mathbf{X}}_i$ 之间的重建误差，目标函数如下：

$$\mathcal{L} = \sum_{i=1}^{n} \|(\mathbf{X}_i - \hat{\mathbf{X}}_i)\|_2^2. \tag{5-6}$$

将方程 (5-6) 中的深自动编码模型的损失函数与方程 (5-5) 中的 GBR 算法相结合，我们给出了深度全局平衡回归算法的目标函数：

$$\begin{aligned}\min \quad & \textstyle\sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (\phi(\mathbf{X}_i)\beta))), \\ s.t. \quad & \textstyle\sum_{j=1}^{p} \left\| \frac{\phi(\mathbf{X}_{\cdot,-j})^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\phi(\mathbf{X}_{\cdot,-j})^T \cdot (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^T \cdot (1 - \mathbf{X}_{\cdot,j})} \right\|_2^2 \leq \lambda_1, \\ & \|(W \cdot \mathbf{1}) \odot (X - \hat{X})\|_F^2 \leq \lambda_2, \;\; W \geq 0, \;\; \|W\|_2^2 \leq \lambda_3, \\ & \|\beta\|_2^2 \leq \lambda_4, \;\; \|\beta\|_1 \leq \lambda_5, \;\; (\textstyle\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_6 \\ & \textstyle\sum_{k=1}^{K} (\|A^{(k)}\|_F^2 + \|\hat{A}^{(k)}\|_F^2) \leq \lambda_7,\end{aligned} \tag{5-7}$$

其中为了简化，我们让 $\phi(\cdot) = \phi(\cdot)^{(K)}$。$\|(W \cdot \mathbf{1}) \odot (X - \hat{\mathbf{X}})\|_F^2$ 表示输入特征 $\mathbf{X}$ 和重建特征 $\hat{\mathbf{X}}$ 之间的重建错误。W 是全局样本权重。$\sum_{k=1}^{K} \|\mathbf{A}^k\|_F^2 + \|\hat{\mathbf{A}}^k\|_F^2 \leq \lambda_7$ 约束深度自动编码器模型中的参数，防止过拟合。

① 我们使用 sigmoid 函数 $\sigma(x) = \frac{1}{1+\exp(-x)}$ 作为非线性激活功能。

5.4.2　理论分析

在本章中，我们对我们提出的算法进行了理论分析，并证明它可以在未知环境中进行稳定预测。该方法工作的关键要求是重叠性（Overlap）假设成立，这是因果推理文献中常见的假设[50]。

假设 5.2 (重叠性 (Overlap))：对于任何一个变量 $\mathbf{X}_{\cdot,j}$，如果将其视为干预变量，那么 $\forall j, 0 < P(\mathbf{X}_{\cdot,j} = 1|\mathbf{X}_{\cdot,-j}) < 1$。

然后我们可以推导出以下引理和定理：

引理 5.1：如果 $\forall j, 0 < P(\mathbf{X}_{\cdot,j} = 1|\mathbf{X}_{\cdot,-j}) < 1$, 且 $\mathbf{X}$ 是二值的，那么 $\forall i, 0 < P(\mathbf{X}_i = x) < 1$，其中 $\mathbf{X}_i$ 是 X 中的第 i 行。

证明 假设干预变量 $T = \mathbf{X}_{i,j}$，那么 $\mathbf{X}_{i,-j}$ 就是潜在的混淆变量。我们知道倾向值的取值范围是 $(0, 1)$，以及 $\exists(x_1^0, \cdots, x_{j-1}^0, x_{j+1}^0, \cdots, x_p^0)$, $P(\mathbf{X}_{i,-j} = (x_1^0, \cdots, x_{j-1}^0, x_{j+1}^0, \cdots, x_p^0)) > 0$，和

$$
\begin{aligned}
& P(\mathbf{X}_i = (x_1^0, \cdots, x_{j-1}^0, x_j, x_{j+1}^0, \cdots, x_p^0)) \\
= \; & P(\mathbf{X}_{i,-j} = (x_1^0, \cdots, x_{j-1}^0, x_{j+1}^0, \cdots, x_p^0)) P(\mathbf{X}_{i,j} = x_j | \mathbf{X}_{i,-j} = (x_1^0, \cdots, x_{j-1}^0, x_{j+1}^0, \cdots, x_p^0))
\end{aligned}
$$

我们可以得出对于 $x_j = 0$ 或者 $x_j = 1$。

$$
0 < P(\mathbf{X}_i = (x_1^0, \cdots, x_{j-1}^0, x_j, x_{j+1}^0, \cdots, x_p^0) < 1 \tag{5-8}
$$

□

基于不等式 (5-8)，我们接下来证明 $\forall x$ (x 是二值的)，

$$
0 < P(\mathbf{X}_i = x) < 1.
$$

让 $k \neq j$，由于

$$
\begin{aligned}
& P(\mathbf{X}_i = (x_1^0, \cdots, x_{j-1}^0, x_j, x_{j+1}^0, \cdots, x_p^0)) \\
= \; & P(\mathbf{X}_{i,-k} = (x_1^0, \cdots, x_{k-1}^0, x_{k+1}^0, \cdots, x_p^0)) P(\mathbf{X}_{i,k} = x_k^0 | X_{i,-k} = (x_1^0, \cdots, x_{k-1}^0, x_{k+1}^0, \cdots, x_p^0))
\end{aligned}
$$

和 $0 < P(\mathbf{X}_i = (x_1^0, \cdots, x_{j-1}^0, x_j, x_{j+1}^0, \cdots, x_p^0)) < 1$，我们可以得到

$$
P(\mathbf{X}_{i,-k} = (x_1^0, \cdots, x_{k-1}^0, x_{k+1}^0, \cdots, x_p^0)) > 0
$$

由于 $\mathbf{X}_{i,k}$ 可以被视为干预变量，所以

$$0 < P(\mathbf{X}_{i,k} = x_k^0 | \mathbf{X}_{i,-k} = (x_1^0, \cdots, x_{k-1}^0, x_{k+1}^0, \cdots, x_p^0)) < 1,$$

因此

$$0 < P(\mathbf{X}_{i,k} = 1 - x_k^0 | \mathbf{X}_{i,-k} = (x_1^0, \cdots, x_{k-1}^0, x_{k+1}^0, \cdots, x_p^0)) < 1$$

不失一般性，我们假设 $k < j$，则我们可以推导出 $\forall x_k, x_j$

$$0 < P(\mathbf{X}_i = (x_1^0, \cdots, x_{k-1}^0, x_k, x_{k+1}^0, \cdots, x_{j-1}^0, x_j, x_{j+1}^0, \cdots, x_p^0) < 1$$

对每一维变量我们都将其视为干预变量并重复以上证明，我们可以得出 $\forall x$,

$$0 < P(\mathbf{X}_i = x) < 1$$

定理 5.1： 让 $X \in \mathbb{R}^{n \times p}$。在引理5.1满足的情况下，如果变量维度 p 是有限的，那么 $\exists W$ 以概率 1 满足

$$\lim_{n \to \infty} \sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{-j}^T (W \odot \mathbf{X}_{\cdot,j})}{W^T \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{-j}^T (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^T (1 - \mathbf{X}_{\cdot,j})} \right\|_2^2 = 0 \tag{5-9}$$

例如，$W_i^* = \frac{1}{P(\mathbf{X}_i = x)}$ 就是满足等式 (5-9) 的一个 W 解。

证明 因为 $\|\cdot\| \geq 0$，公式 (5-9) 可以被简化为 $\forall j, \forall k \neq j$

$$\lim_{n \to \infty} \left(\frac{\sum_{i:\mathbf{X}_{i,k}=1, \mathbf{X}_{i,j}=1} W_i}{\sum_{i:\mathbf{X}_{i,j}=1} W_i} - \frac{\sum_{i:\mathbf{X}_{i,k}=1, \mathbf{X}_{i,j}=0} W_i}{\sum_{i:\mathbf{X}_{i,j}=0} W_i} \right) = 0$$

以概率 1 成立。基于引理 5.1我们知道 $0 < P(\mathbf{X}_i = x) < 1$，那么对于 W^*，$\forall x, \forall i$, $t = 1$ 或 0,

$$\begin{aligned} \lim_{n \to \infty} \frac{1}{n} \sum_{i:\mathbf{X}_{i,j}=t} W_i^* &= \lim_{n \to \infty} \frac{1}{n} \sum_{x:x_j=t} \sum_{i:\mathbf{X}_i=x} W_i^* \\ &= \lim_{n \to \infty} \sum_{x:x_j=t} \frac{1}{n} \sum_{i:\mathbf{X}_i=x} \frac{1}{P(\mathbf{X}_i = x)} \end{aligned}$$

$$= \lim_{n\to\infty} \sum_{x:x_j=t} P(\mathbf{X}_i = x) \cdot \frac{1}{P(\mathbf{X}_i = x)} = 2^{p-1}$$

以概率 1 成立 (大数定理)。由于特征都是二值的

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=1} W_i^* = 2^{p-2}$$
$$\lim_{n\to\infty} \frac{1}{n} \sum_{i:\mathbf{X}_{i,j}=0} W_i^* = 2^{p-1}, \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i:\mathbf{X}_{i,k}=1,X_{i,j}=0} W_i^* = 2^{p-2}$$

因此，我们可以以概率 1 得到以下等式：

$$\lim_{n\to\infty} \left(\frac{\mathbf{X}_{\cdot,k}^T (W^* \odot \mathbf{X}_{\cdot,j})}{W^{*T} \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,k}^T (W^* \odot (1 - \mathbf{X}_{\cdot,j}))}{W^{*T} (1 - \mathbf{X}_{\cdot,j})} \right) = \frac{2^{p-2}}{2^{p-1}} - \frac{2^{p-2}}{2^{p-1}} = 0.$$

以下理论结果表明，如果有足够的数据使得 x 的所有值都出现在数据中，则我们算法可以求出能实现变量全局精准平衡的样本权重。随后，我们展示了在这种情况下，$\mathbf{X}$ 在重新加权的数据中是相互独立的。在现实世界的数据集中，可能无法得到准确平衡的样本权重，但理论结果仍然表明近似平衡权重依旧可以减少变量之间的协方差。

命题 5.1：　对于所有的 x，如果 $0 < \hat{P}(\mathbf{X}_i = x) < 1$，其中 $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n} \sum_i \mathbb{I}(\mathbf{X}_i = x)$，那么存在一个全局样本权重 W^* 使得公式 (5-4) 等于 0，而且在通过样本权重 W^* 加权之后的数据上 $\mathbf{X}$ 中的变量是相互独立的。

证明 如果 $0 < \hat{P}(\mathbf{X}_i = x) < 1$，由理论 5.1我们得知 $W_i^* = \frac{1}{\hat{P}(\mathbf{X}_i = x)}$ 可以使得公式 (5-4) 等于 0。下面我们证明数据经过 W^* 加权之后，$\mathbf{X}$ 中的变量相互独立。将矩阵 $\mathbf{X} \in \mathbb{R}^{n\times p}$ 中的每一行 $\mathbf{X}_i$ 都重复了 $W_i^* = \frac{1}{\hat{P}(\mathbf{X}_i = x)}$ 次，我们可以得到矩阵 $\mathbf{X} \in \mathbb{R}^{n\times p}$ 的增广矩阵，表示为 $\tilde{\mathbf{X}}$[①]。用 $\tilde{n}$ 表示矩阵 $\tilde{\mathbf{X}}$ 中的行数。当 $0 < \hat{P}(\mathbf{X}_i = x) < 1$ 时，我们可以得到

$$\sum_i W_i^* = \tilde{n} \sum_x \frac{1}{\tilde{n}} \sum_{i:\tilde{\mathbf{X}}_i = x} W_i^* = \tilde{n} \sum_x \hat{P}(\tilde{\mathbf{X}}_i = x) \cdot \frac{1}{\hat{P}(\tilde{\mathbf{X}}_i = x)} = \tilde{n} \cdot 2^p$$

相似地，$\sum_{i:\tilde{\mathbf{X}}_{i,j}=1} W_i^* = \tilde{n} \cdot 2^{p-1}$，$\sum_{i:\tilde{X}_{i,j}=0} W_i^* = \tilde{n} \cdot 2^{p-1}$，和 $\sum_{i:\tilde{\mathbf{X}}_{i,j}=x} W_i^* = \tilde{n}$。因此，对

① W_i^* 不需要是整数。

于 $x = (x_1, \cdots, x_p)$，我们可以推出

$$\hat{P}(\tilde{\mathbf{X}}_i = (x_1, \cdots, x_p)) = \frac{\sum_{i:\tilde{\mathbf{x}}_{i,j}=x} W_i^*}{\sum_i W_i^*} = \frac{1}{2^p}$$

和 $\forall j, \hat{P}(\tilde{\mathbf{X}}_{i,j} = x_j) = \frac{\sum_{i:\tilde{\mathbf{x}}_{i,j}=j} W_i^*}{\sum_i W_i^*} = \frac{1}{2}$，最后我们可以得出

$$\hat{P}(\tilde{\mathbf{X}}_i = (x_1, \cdots, x_p)) = \hat{P}(\tilde{\mathbf{X}}_{i,1} = x_1) \cdots \hat{P}(\tilde{\mathbf{X}}_{i,p} = x_p)$$

上述结果表明 $\tilde{\mathbf{X}}$ 中的变量是相互独立的，等价于数据经过 W^* 加权之后，$\mathbf{X}$ 中的变量相互独立。 □

命题 5.2： 如果对于环境 e 中所有的 x 都满足 $0 < \hat{P}(\mathbf{X}_i^e = x) < 1$，那么对于环境 e'，当给定其通过全局样本权重 W^* 加权之后的 $(\mathbf{X}^{e'}, Y^{e'})$ 联合分布，其中全局样本权重 W^* 是基于环境 e 学习得到的，那么变量 $Y^{e'}$ 和 $\mathbf{V}^{e'}$ 则会相互独立。也就是说 $p^{e'}(x, y) = p^e(y|x) \cdot (1/|\mathcal{X}|)$。

证明 显然，我们首先可以得到 $Pr(Y^{e'} = y|\mathbf{X}^{e'} = x) = Pr(Y^e = y|\mathbf{X}^e = x)$。再结合假设 5.1，我们可以得到 $Pr(Y^{e'} = y|\mathbf{X}^{e'} = x) = Pr(Y^{e'} = y|\mathbf{S}^{e'} = s)$。由命题 5.1，我们得知 $(\mathbf{S}^{e'}, \mathbf{V}^{e'})$ 是相互独立的。因此，我们可以推导出

$$\begin{aligned} Pr(Y^{e'} = y|\mathbf{V}^{e'} = v) =& E_{\mathbf{S}^{e'}}[Pr(Y^{e'} = y|\mathbf{S}^{e'}, \mathbf{V}^{e'} = v)|\mathbf{V}^{e'} = v] \\ =& E_{\mathbf{S}^{e'}}[Pr(Y^{e'} = y|\mathbf{S}^{e'})|\mathbf{V}^{e'} = v] \\ =& Pr(Y^{e'} = y). \end{aligned}$$

因此，$Y^{e'}$ 和 $\mathbf{V}^{e'}$ 独立。 □

命题5.1和5.2表明我们的 GBR 算法可以在满足假设 5.1的未知环境中进行稳定预测，因为在经过样本重新加权后，只有稳定特征跟结果变量相关，且 $p(y|s)$ 在不同的环境中始终不变。GBR 算法的目标函数等价于逻辑回归的对数似然目标。即使正则化约束会对估计的 $p(y|s)$ 产生一些偏差，但偏差会随着样本大小 n 增大而减小。

下面我们考虑深度全局平衡回归（DGBR）算法的性质：

1. *DGBR 算法除了拥有 GBR 算法以上的所有属性之外，还可以使得重叠性假设更容易满足且能降低全局样本权重的方差。*通过 Johnson-Lindenstrauss (JL) 引理[143]，

Algorithm 3 深度全局平衡回归算法（Deep Global Balancing Regression algorithm）

Require: 特征矩阵 $\mathbf{X}$ 和结果变量 Y。

Ensure: 更新参数 W，β，θ。

初始化 $W^{(0)}$，$\beta^{(0)}$ 和 $\theta^{(0)}$,
基于参数 $(W^{(0)}, \beta^{(0)}, \theta^{(0)})$ 计算损失函数,
初始化迭代变量 $t \leftarrow 0$,
repeat
    $t \leftarrow t + 1$,
    固定参数 β 和 θ，通过梯度下降算法更新参数 $W^{(t)}$,
    固定参数 W and θ，通过梯度下降算法更新参数 $\beta^{(t)}$,
    固定参数 W and β，通过梯度下降算法更新参数 $\theta^{(t)}$,
    基于参数 $(W^{(t)}, \beta^{(t)}, \theta^{(t)})$ 计算损失函数，
until 损失函数收敛或者到达最大迭代次数,
return W, β, θ.

我们知道对于任何 $0 < \epsilon < 1/2$ 和 $x_1, \cdots, x_n \in \mathbb{R}^p$，存在一个映射函数 $f : \mathbb{R}^p \to \mathbb{R}^k$，其中 $k = O(\epsilon^{-2} \log n)$，使得$\forall i, j$ $(1-\epsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2$。基于 JL 引理，我们可以将高维数据转换到较低的合适维度空间，同时近似保留点与点之间的原始距离，因此我们 DGBR 算法可以近似保留 GBR 算法以上的所有属性。另外，我们的 DGBR 算法减少了特征维数，因此使得总体重叠假设更容易满足。特征维度降低使得我们算法在保证特征近似平衡的条件下，尽可能降低全局样本权重的方差。

2. *DGBR 算法可以更准确地估计条件概率函数* $p(y|s)$。DGBR 算法中多个非线性映射函数，使得其可以更容易地捕获稳定特征和结果变量之间的非线性关系，从而更准确地估计 $p(y|s)$。

5.5　算法优化和讨论

5.5.1　算法优化

为了在方程 (5-7) 中优化我们的 DGBR 模型，我们提出了一种迭代方法，在算法 3中进行了详细描述。首先随机初始化 W，β 和 $\theta = \{\mathbf{A}^{(k)}, \hat{\mathbf{A}}^{(k)}, b^{(k)}, \hat{b}^{(k)}\}_{k=1}^{K}$，之后每次迭代我们都固定两个参数，并通过梯度下降算法更新第三个参数，直到损失函数收敛或者到达最大迭代次数。

5.5.2　复杂度分析

在优化过程中，主要的时间成本是计算损失函数和更新参数 W，β 和 θ。为计算损失函数，其复杂度为 $O(npd)$，其中 n 是样本大小，p 是变量的维度，d 是深度编码器模型中隐藏层的最大维度。对于更新参数 W，其复杂性也是 $O(npd)$。对于更新参数 β，它是标准的 LASSO 问题，其复杂性为 $O(nd)$。对于更新 θ，其复杂性为 $O(npd)$。

总的来说，算法（Algorithm） 3 中每次迭代的复杂度是 $O(npd)$。

5.5.3　超参调整

为了调整算法和基准方法的参数，我们需要多个分布不同的验证数据集，并且不同于训练数据集的分布。在我们的实验中，我们通过对训练数据进行非随机数据重采样生成此类验证数据集 $\mathcal{E}$。通过选 $RMSE$ 作为公式 (5-1) 和 (5-2) 中评估预测算法的 $Error$ 项，并计算我们算法在所有验证数据集上的 $Average\_Error$ 和 $Stability\_Error$。对于所有的方法，包括我们算法和基准方法，我们利用网格搜索加交叉验证方法来搜索使得验证数据集上的 $Average\_Error + \alpha \cdot Stability\_Error$ 最小的超参。我们在实验中设置 $\alpha = 5$。

验证数据集的构建

构建验证数据的关键是构建协变量的联合分布在不同环境中发生变化的数据集。具体地，我们可以通过改变不稳定特征和稳定特征或者结果变量之间的条件分布，从而改变协变量的分布来产生分布不同的验证数据集。但是，在实际应用中，我们没有关于哪些特征是不稳定特征的先验知识。幸运的是，我们可以通过因果效应评估的算法选出那些对结果变量没有因果效应的特征作为不稳定特征。基于使用经验识别的不稳定特征，我们可以生成验证数据集，这些数据集可以改变不稳定特征的分布并运用于参数调整。

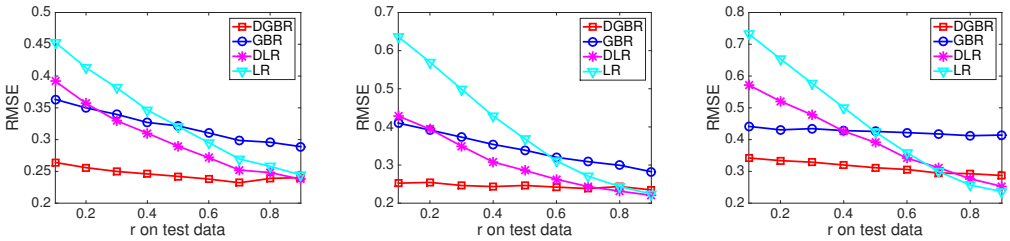
5.6　实验验证

在本章中，我们在仿真数据和真实数据上评估我们算法的性能，并与以下基准方法进行比较。

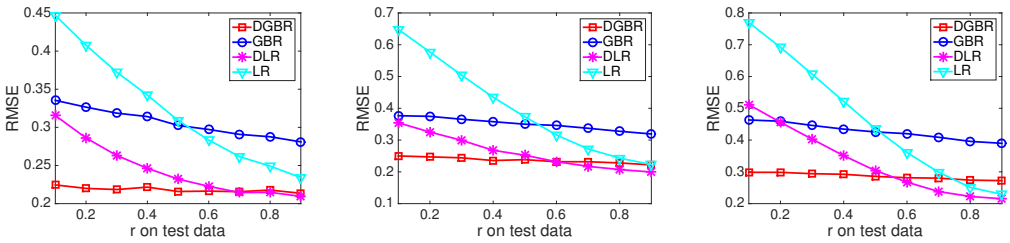
5.6.1　基准方法

我们实现了以下基准方法与我们方法进行比较，包括：

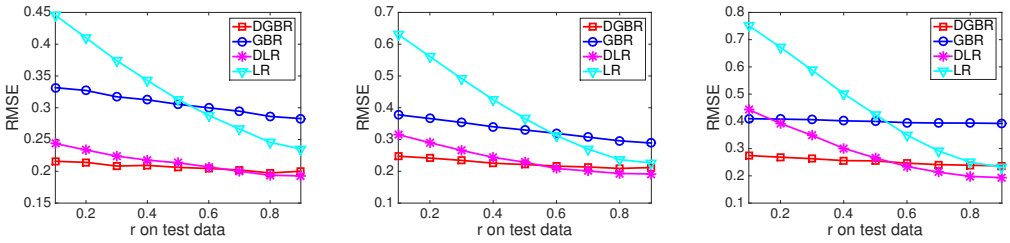
- 逻辑回归模型 *(Logistic Regression，LR)*[144]



(a) Trained on $n = 1000, p = 20, r = 0.65$ (b) Trained on $n = 1000, p = 20, r = 0.75$ (c) Trained on $n = 1000, p = 20, r = 0.85$



(d) Trained on $n = 2000, p = 20, r = 0.65$ (e) Trained on $n = 2000, p = 20, r = 0.75$ (f) Trained on $n = 2000, p = 20, r = 0.85$



(g) Trained on $n = 4000, p = 20, r = 0.65$ (h) Trained on $n = 4000, p = 20, r = 0.75$ (i) Trained on $n = 4000, p = 20, r = 0.85$

图 5.3　$\mathbf{S} \perp \mathbf{V}$ 设定下实验结果。当改变训练数据集的样本大小 n（纵向）和训练偏差率 r（横向）时，各个算法的预测结果（RMSE）。每个子图中横坐标表示测试数据集中的偏差率 r，不同的偏差率代表不同分布的测试数据。

- *深度逻辑回归模型 (Deep Logistic Regression，DLR)*[145]: 接合了深度自编码模型和逻辑回归模型。
- *全局平衡回归模型 (GBR)*: 如目标函数 (5-5) 所示，该方法接合了全局变量平衡正则项和加权的逻辑回归模型。

由于我们的算法是基于逻辑回归模型提出的，所以我们在这里主要与逻辑回归相关的方法进行比较。对于其它预测方法，也可以基于我们提出的变量平衡正则项提出相应的（深度）全局平衡回归算法，并进行比较。我们将其留作将来工作。

5.6.2　仿真数据实验

5.6.2.1　数据描述

在生成仿真数据时，我们考虑由图5.1中所示的所有三种情况，包括：$\mathbf{S} \perp \mathbf{V}$，$\mathbf{S} \to \mathbf{V}$ 和 $\mathbf{V} \to \mathbf{S}$。

S ⊥ V: 在该设定下，**S** 和 **V** 相互独立。基于图5.1，通过独立高斯分布，我们如下生成预测变量 $\mathbf{X} = \{\mathbf{S}_{\cdot,1}, \cdots, \mathbf{S}_{\cdot,p_s}, \mathbf{V}_{\cdot,1}, \cdots, \mathbf{V}_{\cdot,p_v}\}$：

$$\tilde{\mathbf{S}}_{\cdot,1}, \cdots, \tilde{\mathbf{S}}_{\cdot,p_s}, \tilde{\mathbf{V}}_{\cdot,1}, \cdots, \tilde{\mathbf{V}}_{\cdot,p_v} \overset{iid}{\sim} \mathcal{N}(0,1),$$

其中 $p_s + p_v = p$，$\mathbf{S}_{\cdot,j}$ 表示 **S** 中的第 j^{th} 个变量。为了二值化 **X**，我们让 $\mathbf{X}_{\cdot,j} = 1$ 如果 $\tilde{\mathbf{X}}_{\cdot,j} \geq 0$，否则 $\mathbf{X}_{\cdot,j} = 0$。

S → V: 在该设定下，稳定特征 **S** 是不稳定特征的 **V** 原因。我们首先通过独立高斯分布产生稳定特征 $\tilde{\mathbf{S}}$，并让 $\mathbf{S}_{\cdot,j} = 1$ 如果 $\tilde{\mathbf{S}}_{\cdot,j} \geq 0$，否则 $\mathbf{S}_{\cdot,j} = 0$。之后，我们基于稳定特征 $\tilde{\mathbf{S}}$，如下产生不稳定特征 $\tilde{\mathbf{V}} = \{\tilde{\mathbf{V}}_{\cdot,1}, \cdots, \tilde{\mathbf{V}}_{\cdot,p_v}\}$：

$$\tilde{\mathbf{V}}_{\cdot,j} = \tilde{\mathbf{S}}_{\cdot,j} + \tilde{\mathbf{S}}_{\cdot,j+1} + \mathcal{N}(0,2),$$

并且让 $\mathbf{V}_{\cdot,j} = 1$ 如果 $\tilde{\mathbf{V}}_{\cdot,j} > 1$，否则 $\mathbf{V}_{\cdot,j} = 0$。

V → S: 在该设定下，不稳定变量 **V** 是稳定变量 **S** 的原因。我们首先通过独立高斯分布产生不稳定特征 $\tilde{\mathbf{V}}$，并让 $\mathbf{V}_{\cdot,j} = 1$ 如果 $\tilde{\mathbf{V}}_{\cdot,j} \geq 0$，否则 $\mathbf{V}_{\cdot,j} = 0$。之后，我们基于不稳定特征 $\tilde{\mathbf{V}}$，如下产生稳定特征 $\mathbf{S} = \{\mathbf{S}_{\cdot,1}, \cdots, \mathbf{S}_{\cdot,p_s}\}$：

$$\tilde{\mathbf{S}}_{\cdot,j} = \tilde{\mathbf{V}}_{\cdot,j} + \tilde{\mathbf{V}}_{\cdot,j+1} + \mathcal{N}(0,2),$$

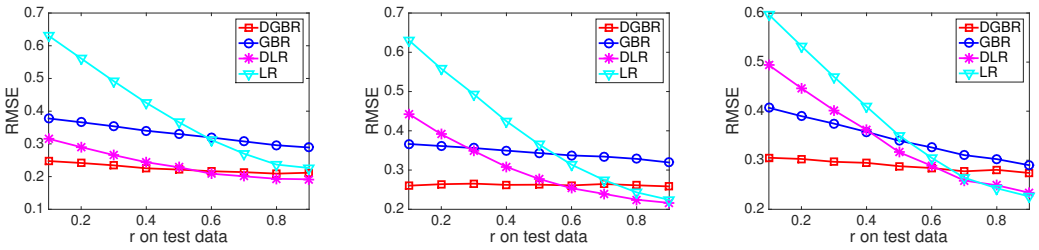
并且让 $\mathbf{S}_{\cdot,j} = 1$ 如果 $\tilde{\mathbf{S}}_{\cdot,j} > 1$，否则 $\mathbf{S}_{\cdot,j} = 0$。

最后，我们使用相同的函数 g 为以上三个设置生成响应变量 Y，如下所示：

$$Y = 1/(1 + \exp(-\sum_{\mathbf{X}_{\cdot,i} \in \mathbf{S}_l} \alpha_i \cdot \mathbf{X}_{\cdot,i} - \sum_{\mathbf{X}_{\cdot,j} \in \mathbf{S}_n} \beta_j \cdot \mathbf{X}_{\cdot,j} \cdot \mathbf{X}_{\cdot,j+1})) + \mathcal{N}(0, 0.2),$$

其中我们将所有的稳定特征 **S** 分成两部分，线性部分 $\mathbf{S}_l$ 和非线性部分 $\mathbf{S}_n$。$\alpha_i = (-1)^i \cdot (i\%3 + 1) \cdot p/3$ 以及 $\beta_j = p/2$。为了二值化结果变量 Y，我们让 $Y = 1$ 如果 $Y \geq 0.5$，否则 $Y = 0$。

为了测试所有算法的稳定性，我们需要生成一系列测试环境 e，每个测试环境都有不同的特征联合分布。在假设5.1下，由于 $(\mathbf{S}, \mathbf{V})$ 的联合分布在不同的环境中不同而导致预测不稳定，也就说明 $P(Y|\mathbf{V})$ 的分布随环境而变化。为了生成与假设5.1一致的测试环境，我们应该改变 $(\mathbf{S}, \mathbf{V})$ 的联合分布，同时保持 Y 和 **V** 的条件独立性。然而，为了创建一组更具挑战性的环境，我们在这里考虑协变量分布在



(a) 训练环境 $n = 4000, p = 20, r = 0.75$　(b) 训练环境 $n = 4000, p = 40, r = 0.75$　(c) 训练环境 $n = 4000, p = 80, r = 0.75$

图 5.4　$\mathbf{S} \perp \mathbf{V}$ 设定下实验结果。当改变训练数据集的变量维度 p 时，各个算法的预测结果（RMSE）。

环境中变化的导致假设5.1不成立的测试环境。这样的设定更能凸显了我们方法对稳定预测的有效性。

具体地，我们引入偏差率 $r \in (0, 1)$ 来对样本进行有偏选择，从而达到改变 $P(Y|\mathbf{V})$ 分布的目的。对于无偏整体样本中的每一个样本，如果其不稳定特征的值与结果变量的值相同，即 $\mathbf{V} = Y$，那么我们选取该样本的概率为 r，否则选取概率为 $1 - r$。值得注意的是，当 $r > .5$ 意味着在我们有偏选取的样本中，不稳定特征 $\mathbf{V}$ 跟结果变量 Y 之间存在虚假的正相关性。

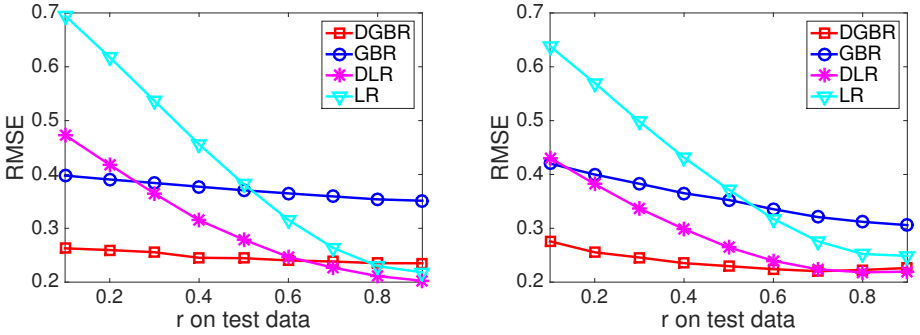
在有偏差的样本选择之后，由于选择偏差问题，即使给定了 $\mathbf{S}$，$\mathbf{V}$ 还是可能与结果变量 Y 相关联。但是，因为 $\mathbf{S}$ 是确定 Y 的一个重要因素，所以通过控制 $\mathbf{S}$ 也能减少 $\mathbf{V}$ 与 Y 之间的虚假相关性。

5.6.2.2　实验结果

我们通过改变样本量 $n = \{1000, 2000, 4000\}$，变量维度 $p = \{20, 40, 80\}$ 和偏差率 $r = \{0.65, 0.75, 0.85\}$，来生成不同的仿真数据。基于设定 $\mathbf{S} \perp \mathbf{V}$ 的实验结果见图 5.3 和 5.4。对于设定 $\mathbf{S} \to \mathbf{V}$ 和 $\mathbf{V} \to \mathbf{S}$ 的实验结果，我们只汇报了一部分，见表 5.5。

对这些实验结果，我们有以下发现和分析：

- 在所有的实验设定下，LR 和 DLR 方法都不能解决稳定预测问题。因为它们无法在模型训练期间消除不稳定特征与结果变量之间的虚假相关性，导致它们通常将不稳定特征学习成预测结果的重要特征，从而导致跨环境预测的不稳定性。
- 相比于基准方法，我们提出的方法能更好地解决稳定预测问题，在所有的设定下，我们方法的预测能力都比较稳定。我们的 GBR 方法比 LR 方法更稳定，DGBR 方法比 DLR 方法更稳定。主要原因是我们方法中的全局变量平衡正则项可以帮助我们去除不稳定特征与结果变量之间的虚假关联，并准确地评估稳定特征对结果变量的影响。



(a) 设定 $\mathbf{S} \to \mathbf{V}$ 和训练环境 $n = 2000, p = 20, r = 0.75$ 下的预测结果　(b) 设定 $\mathbf{V} \to \mathbf{S}$ 和训练环境 $n = 2000, p = 20, r = 0.75$ 下的预测结果

图 5.5　设定 $\mathbf{S} \to \mathbf{V}$ 和 $\mathbf{V} \to \mathbf{S}$ 下的部分实验结果。

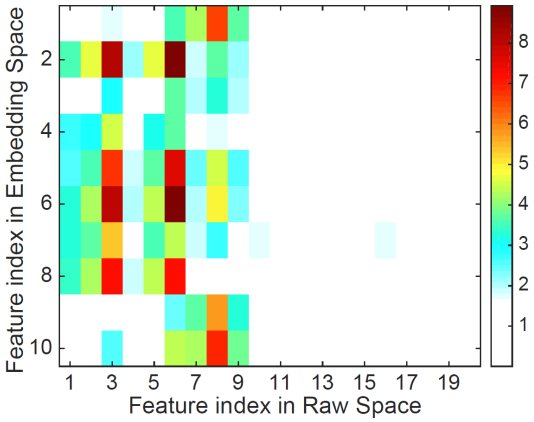
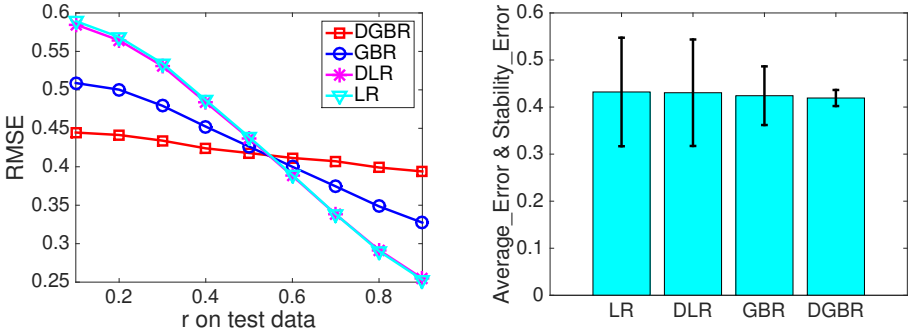


图 5.6　我们 DGBR 算法对特征的表征权重，其中 $\mathbf{X}_{\cdot,1}, \cdots, \mathbf{X}_{\cdot,9}$ 是稳定特征 $\mathbf{S}$，其它特征为非稳定特征 $\mathbf{V}$。我们 DGBR 算法对观测变量 $\mathbf{X}$ 的表征中几乎没有来自非稳定特征 $\mathbf{V}$ 的信息。

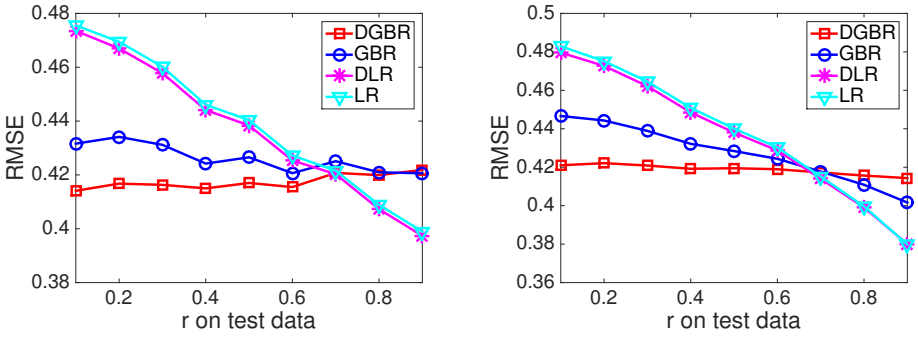
- 相比于 GBR 方法，我们的 DGBR 方法能带来更精准且更稳定的预测。这是因为在 DGBR 算法中，深度自动编码模型通过非线性降低特征维度使得全局变量平衡更容易实现，且能更好地拟合稳定特征与结果变量之间的非线性关系。
- 通过改变训练环境的样本大小 n，变量维度 p 和训练的偏差率 r，我们发现我们 DGBR 算法对未知测试环境的预测错误都很小，而且很稳定。

图 5.6展示了我们 DGBR 算法对特征的表征几乎没有来自不稳定特征 $\mathbf{V}$ 的信息。这表明，通过算法中的联合优化，DGBR 在降低协变量空间的维数时优先考虑稳定特征，从而使得算法能实现稳定预测。



(a) RMSE　　(b) Average_Error & Stability_Error

图 5.7　设定 1 中各个算法在真实广告数据上的预测结果。



(a) 变量 *mail plugin*　　(b) 变量 *bottle plugin*

图 5.8　基于某个变量改变训练数据集上的训练偏差率 r 时，各个算法在所有未知测试数据集上的预测结果。

5.6.3　真实数据实验

5.6.3.1　在线广告数据实验

我们使用的真实在线广告数据来自腾讯微信 App [①]，主要集收集于 2015 年 9 月期间。在微信中，每个用户都可以在朋友圈向他/她的朋友分享帖子，并像 Twitter 和 Facebook 一样接收朋友分享的帖子。广告商可以通过将广告合并到用户的朋友圈中来向用户推送广告。用户对广告的反馈包括：“喜欢”和“不喜欢”。

我们实际用到的广告数据来自于一款 LONGCHAMP[②]女性手提包。该广告数据包括来自 14,891 位用户的“喜欢”反馈和 93,108 位用户“不喜欢”反馈。对于每个用户，我们收集了 56 维特征，包括（1）人口统计学特征，如年龄，性别，（2）朋友数量，（3）设备（iOS 或 Android），以及（4）微信用户设置，例如是否允许陌生人看他/她的朋友圈（“Share Album to Strangers”）以及是否开通了在线支付功能（“With Online Payment”）。

实验设定。在我们实验中，我们让 $Y_i = 1$ 如果用户 i 喜欢这条广告，否则 $Y_i = 0$。

① http://www.wechat.com/en/

② http://en.longchamp.com/en/womens-bags

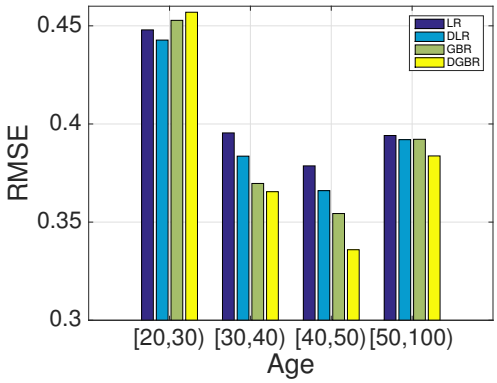


图 5.9　设定 2：不同测试数据集上预测结果。所有模型在年龄属于 [20, 30) 的用户数据集上训练，在所有的 4 个不同用户年龄的数据集测试。

对于非二值的用户特征，我们利用其均值将特征二值化。考虑到重叠性假设（假设5.2），我们只保留了满足条件 $0.2 \leq \frac{\#\{x=1\}}{\#\{x=1\}+\#\{x=0\}} \leq 0.8$ 的用户特征。在我们实验中，所有的预测特征跟结果变量都是二值的。

为了测试我们提算法在稳定预测问题上的性能，我们使用两种不同的设定进行实验。第一个实验设定与仿真数据集上的设定类似。我们通过偏差率 r 选择有偏差的样本，从而来产生不同的环境。在此设置中，我们选择那些与结果无关的特征作为偏差样本选择的不稳定特征。在第二个实验设置中，我们通过用户某些特征的不同来分离数据集并产生各种测试环境。具体来说，我们根据用户的年龄将整个数据集分为 4 个部分，包括年龄属于 [20, 30)，[30, 40)，[40, 50) 和 [50, 100) 四个区间。

设定 1 的实验结果。在该实验设定中，我们先选取了 4 维不稳定特征，实验结果见图 5.7 和图 5.8。基于该 4 维不稳定特征，我们通过不同的偏差率 r 产生了不同的测试环境，用于测试我们方法和基准方法预测的稳定性，实验结果见图5.7(a)。为了突出我们方法在稳定预测中的优势，我们在图5.7(b)中画出了所有方法在所有未知的测试环境下的 $Average\_Error$ 和 $Stability\_Error$。进一步，我们通过改变其它变量与结果变量之间的分布产生了多个测试环境，实验结果见图 5.8。图 5.8(a) 和图 5.8(b) 表明我们 DGBR 方法，相对于所有的基准方法，能在所有的测试环境中实现最稳定的预测。总的来说，在该设定下的实验结果及其解释与仿真数据上的实验非常相似。

设定 2 上的实验结果。

在该实验设定中，我们根据用户的年龄将整个数据集分为 4 个部分，包括年龄属于 [20, 30)，[30, 40)，A[40, 50) 和 [50, 100) 四个区间。我们在数据 [20, 30) 上训练所有的模型，并在所有的 4 部分数据上检测方法的预测稳定性，实验结果见

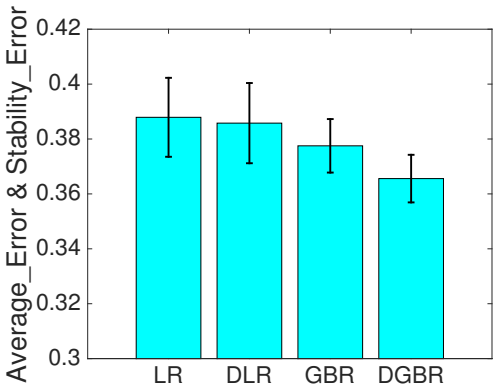


图 5.10　设定 2：*Average_Error* 和 *Stability_Error*

图 5.9。

当测试数据集来自于数据 $[20, 30)$ 时，我们 DGBR 方法的预测效果要比基准方法差一些，这是因为测试数据集的分布与训练数据集分布一致，使得不稳定变量与结果变量之间的虚假相关会帮助提升基准方法的预测能力，而我们方法去除了变量之间的虚假相关信息。当测试数据集来自于其它三个部分时，其分布与训练数据集不同，我们的 DGBR 获得最佳预测性能。因为这时不稳定变量与结果变量之间的虚假相关联会给基准方法的预测起反作用。

从图5.9中我们可以推断出 DGBR 算法的稳定性不如基准方法的好，这是由结果变量的分布 $P(Y)$ 在这四种环境中变化导致的。利用对全局数据集上的 $P(Y = 1) = \frac{14,891}{14,891+93,108}$ 的结果对测试数据集进行数据采样来修复 $P(Y)$ 后，我们在图5.10中汇报了所有算法在四个环境中的 *Average_Error* 和 *Stability_Error*。从实验结果我们发现，当 $P(Y)$ 稳定时，我们的 DGBR 算法的预测稳定性优于所有基准方法。

5.6.3.2　图片分类数据实验

为了验证我们方法在图片识别应用中的有效性，我们在 *YFCC100M*[146] 数据集上进行了实验。*YFCC100M* 数据集中包含 100 万张图片，而且对于每一张都存在多个标签，包含主要的类别标签和上下文标签。为了模拟真实应用中，样本选择偏差导致的不同数据集之间分布不一致的情形，我们基于 *YFCC100M* 数据集手工构建了一个小型数据集。在我们数据集中一共有 10 个类别的物体，对于每个类别的物体（例如狗）图片，都存在 5 个上下文标签（或背景标签，例如，草地，沙滩，雪地等），数据集的具体细节可以参考表 5.1。

实验设定

在实验中，为了增大训练数据集和测试数据集之间的分布差异，我们基于每个类别的 5 个上下文标签分割训练数据集、验证数据集和测试数据集。具体地，我

表 5.1　我们数据集包含 10 个类别，每个类别的数据都各自包含 5 个不同上下文标签及相应每个标签的图片数量。

	Context 1	Context 2	Context 3	Context 4	Context 5	Total
bird	duck(210)	gull(200)	hawk(200)	heron(200)	parrot(190)	1000
bridge	san francisco(160)	london(110)	nyc(110)	street(100)	sydney(180)	660
car	art(114)	bmw(120)	classic(200)	ferrari(200)	racing(180)	814
cat	black(180)	house(120)	kitten(200)	tabby(200)	white(240)	940
church	basilica(94)	catholic(83)	gothic(104)	orthodox(100)	roman(81)	462
dog	beach(200)	car(150)	grass(200)	home(200)	snow(190)	940
flower	blossom(200)	lily(240)	orchid(240)	rose(220)	tulip(190)	1090
horse	dressage(260)	equestrian(206)	jumping(200)	pony(50)	racing(140)	856
train	diesel(250)	locomotive(230)	metro(100)	station(68)	steam(150)	798
tree	christmas(140)	leaves(220)	palm(170)	snow(160)	spring(170)	860

们用类别在前三个上下文标签（标签 1，2，3）的数据用于模型训练，第 4 个上下文标签的数据用于模型验证，第 5 个上下文标签的数据用于模型测试。此外，在构造训练数据集时，我们采取了非均匀采样的方式来增加数据中的样本选择偏差。具体地，对于不同的上下文标签，选取的比例差别很大，我们让第 1 个上下文标签占绝大多数。其中，第 1，2，3 个上下文标签在训练数据集中所占的比例分别为 66%,17%,17%。对于图片特征提取，我们采用了视觉词（visual words）[147] 来提取图片特征。

实验结果

在本实验中，我们采用 F1 值作为预测算法对图片识别有效性的评价指标。在图 5.11中我们汇报了具体的实验结果。

在图 5.11中为了更明显地展示我们算法相对于基准方法的性能提升，并刻画数据集偏差程度与算法性能提升之间的关系，我们通过 EMD（Earth Moving Distance）距离度量了每个类别图片在训练集和测试集之间特征向量分布的差距，并把这个差距与每个类别上我们的方法对于最好的基准方法 F1 值的相对提升进行比较。从图 5.11结果所示，我们发现当随着数据集之间偏差程度越来越严重，我们方法相对于基准方法的提升越显著。

因果特征的可视化和模型解释性

由于我们算法能基于数据之间的关联关系恢复变量之间的因果关系，我们算法可以从所有特征中选择因果特征进行预测，初步实现可解释性预测。为了展示我们算法的解释能力，我们分别可视化了我们的算法和逻辑回归算法所提取出的 5 维最重要的特征（根据特征的回归系数进行排序）。如图 5.12所示，我们可以看到我们的算法所选取的特征几乎都位于目标物体之上，也就是因果特征。而相对的，逻辑回归则选择了很多位于背景的上下文特征。基于因果特征进行预测，我

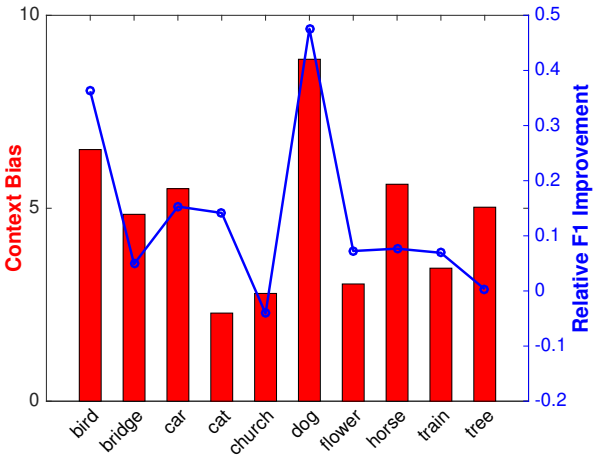


图 5.11　数据集偏差程度 V.S. 算法性能提升。

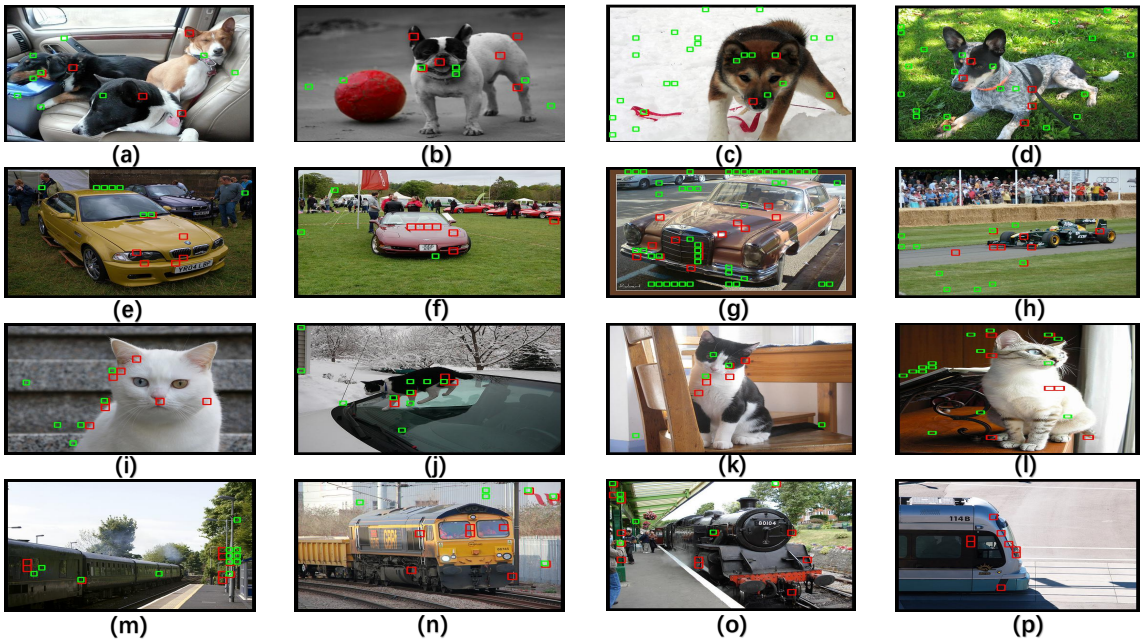
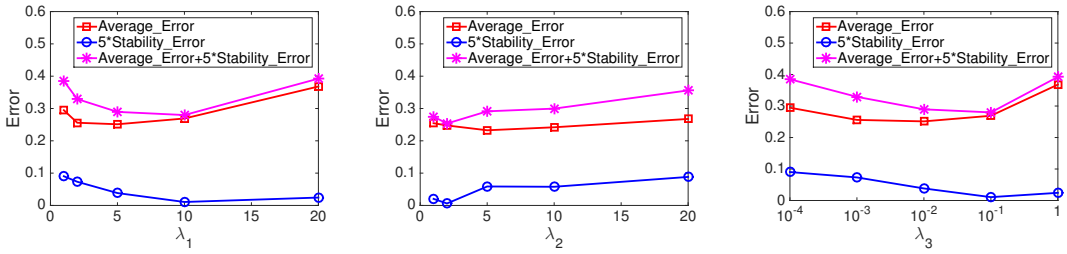


图 5.12　我们的算法和逻辑回归算法分别所提取的前 5 维最重要的特征，红框表示我们的算法提取的特征，绿框表示逻辑回归算法提取的特征。

们模型使得算法使用者能更好地理解我们算法预测的机制，增强了预测模型的可解释性。

5.6.4　超参分析

在我们的 DGBR 算法中，我们有一些超参数，例如 λ_1 用于约束全局变量平衡的偏差，λ_2 约束自动编码器的损失函数，λ_3 约束全局样本的方差等。在本章中，我们将研究这些超参数如何影响预测结果。我们根据构建的验证数据，通过网格搜索进



(a) λ_1　　(b) λ_2　　(c) λ_3

图 5.13　超参 λ_1，λ_2 和 λ_3 的敏感度分析。

行交叉验证，在实验中调整了这些参数。我们在 $\mathbf{S} \perp \mathbf{V}$，$n = 2000$ 和 $p = 20$ 的设定下汇报了 $Average\_Error$，$5 * Stability\_Error$ 和 $Average\_Error + 5 * Stability\_Error$ 的值。

预测效果和变量平衡之间的权衡：在图 5.13(a)中，我们首先展示超参数 λ_1 如何影响预测模型的性能。其中 λ_1 的参数限制了全局变量平衡的偏差。我们可以看到，当 λ_1 的值增加时，$Average\_Error$ 和 $Stability\_Error$ 的值刚开始都会减少。这很直观，因为通过 λ_1 的增加可以使得数据更加平衡，并且平衡数据可以帮助识别稳定的特征并去除一些噪声以进行更精确的预测。但是，当 λ_1 的值进一步增加时，$Stability\_Error$ 的值会减小，但 $Average\_Error$ 的值开始缓慢增加。λ_1 的大值使得算法集中在全局变量平衡正则项上而忽略了预测模型的重要性。预测效果和全局平衡约束对于稳定预测都是必不可少的。

特征表征：在图 5.13(b)中，我们展示了超参 λ_2 对预测模型性能的影响。$Average\_Error$ 的值随 λ_2 增大而减小，这是因为 λ_2 越高会导致越准确的预测。最初，$Stability\_Error$ 随 λ_2 增大而减少，但当 $\lambda_2 \geq 5$ 时，$Stability\_Error$ 则随 λ_2 增大而增大。选择合适的 λ_2 值对特征表征很重要，但我们的方法对此参数不是很敏感。

全局样本权重的方差：图5.13(c)中展示了超参 λ_3 对结果的影响。当 λ_3 的值增加时，$Average\_Error$ 和 $Stability\_Error$ 的值都会减小。因为对全局样本权重的方差进行适当约束可能会阻止某些样本在整个数据中占主导地位，从而有助于提高预测的精度和稳健性。但是，当 λ_3 的值变得太大时，这些错误会增加。因为 λ_3 的值太大可能导致学习的全局样本权重无法在平衡和预测之间进行适当的权衡。

5.7　本章小结

在本章中，我们关注如何在未知环境中进行稳定预测，其中未知环境的数据分布可能与训练数据的分布非常不同。我们认为，大多数先前用于解决稳定预测的方法都是不完善的，因为它们需要将测试数据的分布作为先验知识，或者依赖

于来自不同环境的训练数据集的多样性。因此，我们提出了一种深度全局平衡回归算法，通过联合优化深度自动编码器模型和全局平衡模型，在未知环境中进行稳定预测。全局平衡模型可以识别预测变量和结果变量之间的因果关系，而深度自动编码器模型设计用于捕获变量之间的非线性结构，使全局平衡更容易，噪声更小。从理论分析和经验实验中，我们证明了我们的算法可以做出稳定的预测。仿真数据集和现实数据集上的大量实验结果表明，我们的 DGBR 算法预测性能优于所有基准方法，可以在未知环境中进行稳定预测。

第 6 章　研究总结和工作展望

6.1　研究总结

本文针对现有预测模型存在的不可解释性和预测不稳定性，研究了因果约束的稳定学习理论方法，通过大数据因果推理方法挖掘变量之间不变的因果关系，并利用因果关系对预测模型学习进行约束，使得预测模型能对未知分布的测试数据集实现稳定预测。研究过程中，面对观测数据下因果推理中大数据带来的挑战，我们提出了全新的因果推理框架和因果效应评估算法，且在大量真实数据上验证了我们方法的有效性。通过融合大数据因果推理算法和预测推理模型，我们提出了因果约束的稳定学习算法来提升预测模型的可解释性和预测稳定性，并且在理论和实验上验证了我们稳定学习框架和算法的有效性。

具体来说，各个工作总结如下：

首先，面向高维数据中并不是所有的观测变量都是混淆变量的问题，本文提出了混淆变量自动分离的因果推理模型。该模型通过数据驱动的方法将观测数据中所有的观测变量分为三部分：混淆变量，调整变量和无关变量。并通过构造混淆变量和调整变量之间的正交性约束项，本文进一步提出了数据驱动的变量自动分离和因果推理算法。其中分离的混淆变量可以帮助无偏地评估因果效应，而调整变量通过回归可以帮助低因果效应评估的方差。在理论和实验上，我们都验证了该模型在面向高维数据时因果推理的准确性和鲁棒性。

其次，面向混淆变量差异性，本文从理论上证明了在因果推理问题中不同的混淆变量给因果推理带来的误差影响是不一样的，并提出了混淆变量区分性平衡的因果推理模型。该模型通过同时学习混淆变量权重和样本权重来实现混淆变量区分性平衡。该模型中的混淆变量权重用于选择混淆变量和区分混淆变量的影响，而样本权重则用于平衡混淆变量分布来消除其对因果推理的影响。通过在仿真数据和真实数据上的大量实验，我们验证了该模型在面向混淆变量差异性时能准确且稳定地评估因果效应。

最后，为了解决预测问题中存在的变量虚假相关性，数据分布差异性和测试数据未知性等挑战，本文提出了因果约束的稳定学习理论方法框架。具体地，基于传统的逻辑回归模型，本文通过联合优化深度自动编码模型和变量全局平衡模型提出了一种深度全局平衡的回归模型，实现面向未知分布的测试数据的稳定预测。该模型中变量全局平衡模型用来挖掘变量之间的因果不变关系，而深度自动编码模型用于学习变量的低维度的非线性表征。在理论和实验上，我们都验证了该模

型能实现面向未知分布的测试数据的稳定预测。

6.2　未来工作展望

本文通过融合因果推理和机器学习，提出了因果约束的稳定学习框架，初步解决了现有预测模型存在的不可解释和预测不稳定等问题，但依然存在很多可以改进的空间有待未来工作进一步研究解决。另外，通过融合因果推理，也可以弥补现有机器学习方法其他方面的一些缺陷，例如算法偏差等。因此，关于因果约束的机器学习这个课题，未来还有以下可以研究的方向：

1. 连续干预变量的因果推理和效应估计：本文研究的大数据因果推理模型主要针对的是二值的干预变量，例如病人吃药和不吃药之间的差别。然而在很多实际应用中，我们关心的往往是某些连续变量的因果效应，例如病人吃多少片药效果最好；或者商品售价为多少时盈利最大等等。基于大量观测数据，如何评估连续变量对最后结果的因果效应也是重要的研究方向。

2. 基于时序数据的稳定学习理论方法研究：本文提出的因果约束的稳定学习框架主要针对的是基于对静态数据的预测模型，然而很多实际场景中的数据往往是具有时序性的。如何针对时序数据设计因果约束的稳定学习框架也是下一个研究方向。

3. 跨模态的稳定学习理论方法研究：本文提出的因果约束的稳定学习框架主要考虑的是单模态数据下的稳定学习，然而在实际应用中，我们的数据往往是多模态的，例如文本数据，图片数据和视频数据等。如何融合多模态数据表征，多模态因果推理和机器学习，提出多模态因果约束的稳定学习理论方法是极具挑战性和研究价值的问题。

4. 去除算法不公平性的机器学习方法研究：在现有机器学习算法中，很多机器学习方法由于数据中存在的偏差问题，导致算法模型的不公平性。例如，由于训练数据集中，黑人犯罪数据比例显著高于白人犯罪数据，大部分现有机器学习会将肤色作为预测是否犯罪的重要特征，导致模型的不公平性。因为在我们认为，肤色对于判断一个人是否会犯罪应该没有任何因果效应。因此，如何利用因果推理技术来去除有数据偏差带来的算法不公平问题是非常具有研究意义和现实意义的问题。

参考文献

[1] Big data[J]. Nature, 2008, 455(7209):1-136.

[2] Dealing with data[J]. Science, 2011, 331(6024):639-806.

[3] 邬贺铨. 大数据时代的机遇与挑战[J]. 求是, 2013, 4(9).

[4] Elwert F, Winship C. Endogenous selection bias: The problem of conditioning on a collider variable[J]. Annual review of sociology, 2014, 40:31-53.

[5] Shiffrin R M. Drawing causal inference from big data[J]. Proceedings of the National Academy of Sciences, 2016, 113(27):7308-7309.

[6] Gunning D. Explainable artificial intelligence (xai)[J]. Defense Advanced Research Projects Agency (DARPA), 2017.

[7] Stuart E A. Matching methods for causal inference: A review and a look forward[J]. Statistical science: a review journal of the Institute of Mathematical Statistics, 2010, 25(1):1.

[8] Funk M J, Westreich D, Wiesen C, et al. Doubly robust estimation of causal effects[J]. American journal of epidemiology, 2011, 173(7):761-767.

[9] Winship C, Morgan S L. The estimation of causal effects from observational data[J]. Annual review of sociology, 1999, 25(1):659-706.

[10] Winship C, Sobel M, Hardy M, et al. Causal inference in sociological studies[J]. Handbook of Data Analysis, 2004.

[11] Morgan S L, Winship C. Counterfactuals and causal inference[M]. : Cambridge University Press, 2014

[12] Sobel M E. Causal inference in the social and behavioral sciences[M]//Handbook of statistical modeling for the social and behavioral sciences. : Springer, 1995: 1-38

[13] Morton R B, Williams K C. Experimental political science and the study of causality: From nature to the lab[M]. : Cambridge University Press, 2010

[14] Druckman J N, Green D P, Kuklinski J H, et al. Cambridge handbook of experimental political science[M]. : Cambridge University Press, 2011

[15] Kittel B, Luhan W J, Morton R B. Experimental political science: Principles and practices[M]. : Palgrave Macmillan Basingstoke, 2012

[16] Margetts H Z. Experiments for public management research[J]. Public Management Review, 2011, 13(2):189-208.

[17] Baekgaard M, Baethge C, Blom-Hansen J, et al. Conducting experiments in public management research: A practical guide[J]. International Public Management Journal, 2015, 18(2):323-342.

[18] Anderson D M, Edwards B C. Unfulfilled promise: Laboratory experiments in public management research[J]. Public Management Review, 2015, 17(10):1518-1542.

[19] Bouwman R, Grimmelikhuijsen S. Experimental public administration from 1992 to 2014: A systematic literature review and ways forward[J]. International Journal of Public Sector Management, 2016, 29(2):110-131.

[20] James O, Jilke S R, Van Ryzin G G. Experiments in public management research: Challenges and contributions[M]. : Cambridge University Press, 2017

[21] Sun W, Wang P, Yin D, et al. Causal inference via sparse additive models with application to online advertising.[C]//AAAI. 2015: 297-303.

[22] Wang P, Sun W, Yin D, et al. Robust tree-based causal inference for complex ad effectiveness analysis[C]//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015: 67-76.

[23] Chan D, Ge R, Gershony O, et al. Evaluating online ad campaigns in a pipeline: causal models at scale[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010: 7-16.

[24] Kuang K, Jiang M, Cui P, et al. Steering social media promotions with effective strategies[C]// 2016 IEEE 16th International Conference on Data Mining (ICDM). 2016: 985-990.

[25] Kuang K, Jiang M, Cui P, et al. Effective promotional strategies selection in social media: A data-driven approach[J]. IEEE Transactions on Big Data, 2018, 4(4):487-501.

[26] Keteyian S J, Levine A B, Brawner C A, et al. Exercise training in patients with heart failure a randomized, controlled trial[J]. Annals of internal medicine, 1996, 124(12):1051-1057.

[27] Chang E L, Wefel J S, Hess K R, et al. Neurocognition in patients with brain metastases treated with radiosurgery or radiosurgery plus whole-brain irradiation: a randomised controlled trial [J]. The lancet oncology, 2009, 10(11):1037-1044.

[28] Gerber A S, Gimpel J G, Green D P, et al. How large and long-lasting are the persuasive effects of televised campaign ads? results from a randomized field experiment[J]. American Political Science Review, 2011, 105(1):135-150.

[29] Panagopoulos C, Green D P. Field experiments testing the impact of radio advertisements on electoral competition[J]. American Journal of Political Science, 2008, 52(1):156-168.

[30] 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.

[31] Rosenbaum P R, Rubin D B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score[J]. The American Statistician, 1985, 39(1):33-38.

[32] Hill J, Reiter J P. Interval estimation for treatment effects using propensity score matching[J]. Statistics in Medicine, 2006, 25(13):2230-2256.

[33] Dehejia R H, Wahba S. Propensity score-matching methods for nonexperimental causal studies [J]. Review of Economics and statistics, 2002, 84(1):151-161.

[34] Sekhon J S, et al. Multivariate and propensity score matching software with automated balance optimization: The matching package for r[J]. Journal of Statistical Software, 2011, 42(i07).

[35] Rubin D B, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates[J]. Journal of the American Statistical Association, 2000, 95(450): 573-585.

[36] Baser O. Too much ado about propensity score models? comparing methods of propensity score matching[J]. Value in Health, 2006, 9(6):377-385.

[37] Abadie A, Imbens G W. Matching on the estimated propensity score[J]. Econometrica, 2016, 84(2):781-807.

[38] Rosenbaum P R, Rubin D B. Reducing bias in observational studies using subclassification on the propensity score[J]. Journal of the American statistical Association, 1984, 79(387):516-524.

[39] Lunceford J K, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study[J]. Statistics in medicine, 2004, 23(19):2937-2960.

[40] Senn S, Graf E, Caputo A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure[J]. Statistics in medicine, 2007, 26(30): 5529-5544.

[41] Cao X. Exploring causal effects of neighborhood type on walking behavior using stratification on the propensity score[J]. Environment and Planning A, 2010, 42(2):487-504.

[42] Austin P C. An introduction to propensity score methods for reducing the effects of confounding in observational studies[J]. Multivariate behavioral research, 2011, 46(3):399-424.

[43] Luellen J K, Shadish W R, Clark M. Propensity scores: An introduction and experimental test [J]. Evaluation Review, 2005, 29(6):530-558.

[44] Glynn A N, Quinn K M. An introduction to the augmented inverse propensity weighted estimator [J]. Political analysis, 2010, 18(1):36-56.

[45] Curtis L H, Hammill B G, Eisenstein E L, et al. Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases[J]. Medical care, 2007, 45 (10):S103-S107.

[46] Austin P C, Stuart E A. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies[J]. Statistics in medicine, 2015, 34(28):3661-3679.

[47] Bang H, Robins J M. Doubly robust estimation in missing data and causal inference models[J]. Biometrics, 2005, 61(4).

[48] Dudík M, Langford J, Li L. Doubly robust policy evaluation and learning[C]//International Conference on Machine Learning. 2011: 1097-1104.

[49] Bloniarz A, Liu H, Zhang C H, et al. Lasso adjustments of treatment effect estimates in randomized experiments[J]. PNAS, 2016.

[50] Athey S, Imbens G W, Wager S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions[J]. Journal of the Royal Statistical Society Series B, 2018, 80(4):597-623.

[51] Li S, Fu Y. Matching on balanced nonlinear representations for treatment effects estimation[C]// Advances in Neural Information Processing Systems. 2017: 930-940.

[52] Zubizarreta J R. Stable weights that balance covariates for estimation with incomplete outcome data[J]. Journal of the American Statistical Association, 2015, 110(511):910-922.

[53] Hainmueller J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies[J]. Political Analysis, 2012, 20(1):25-46.

[54] Chan K C G, Yam S C P, Zhang Z. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2015.

[55] Deville J C, Särndal C E. Calibration estimators in survey sampling[J]. Journal of the American statistical Association, 1992, 87(418):376-382.

[56] Graham B S, de Xavier Pinto C C, Egel D. Inverse probability tilting for moment condition models with missing data[J]. The Review of Economic Studies, 2012, 79(3):1053-1079.

[57] Graham B S, Pinto C C d X, Egel D. Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast)[J]. Journal of Business & Economic Statistics, 2016, 34(2):288-301.

[58] Hellerstein J K, Imbens G W. Imposing moment restrictions from auxiliary data by weighting [J]. Review of Economics and Statistics, 1999, 81(1):1-14.

[59] Zhao Q, Percival D. Entropy balancing is doubly robust[J]. Journal of Causal Inference, 2017, 5(1).

[60] Zhao Q, et al. Covariate balancing propensity score by tailored loss functions[J]. The Annals of Statistics, 2019, 47(2):965-993.

[61] Imai K, Ratkovic M. Covariate balancing propensity score[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 76(1):243-263.

[62] Hayashi F. Econometrics. 2000[J]. Princeton University Press. Section, 2000, 1:60-69.

[63] Owen A B. Empirical likelihood[M]. : Chapman and Hall/CRC, 2001

[64] Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function[J]. Journal of statistical planning and inference, 2000, 90(2):227-244.

[65] Sugiyama M, Müller K R. Model selection under covariate shift[C]//International Conference on Artificial Neural Networks. 2005: 235-240.

[66] Bickel S, Brückner M, Scheffer T. Discriminative learning under covariate shift[J]. Journal of Machine Learning Research, 2009, 10(Sep):2137-2155.

[67] Huang J, Gretton A, Borgwardt K M, et al. Correcting sample selection bias by unlabeled data [C]//Advances in neural information processing systems. 2007: 601-608.

[68] Dudík M, Phillips S J, Schapire R E. Correcting sample selection bias in maximum entropy density estimation[C]//Advances in neural information processing systems. 2006: 323-330.

[69] Wen J, Yu C N, Greiner R. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification[C]//International Conference on Machine Learning. 2014: 631-639.

[70] Liu A, Ziebart B. Robust classification under sample selection bias[C]//Advances in neural information processing systems. 2014: 37-45.

[71] Kanamori T, Hido S, Sugiyama M. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection[C]//Advances in neural information processing systems. 2009: 809-816.

[72] Sugiyama M, Nakajima S, Kashima H, et al. Direct importance estimation with model selection and its application to covariate shift adaptation[C]//Advances in neural information processing systems. 2008: 1433-1440.

[73] Yamada M, Suzuki T, Kanamori T, et al. Relative density-ratio estimation for robust distribution comparison[J]. Neural computation, 2013, 25(5):1324-1370.

[74] Zadrozny B. Learning and evaluating classifiers under sample selection bias[C]//Proceedings of the twenty-first international conference on Machine learning. 2004: 114.

[75] Bickel S, Brückner M, Scheffer T. Discriminative learning for differing training and test distributions[C]//Proceedings of the 24th international conference on Machine learning. 2007: 81-88.

[76] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on knowledge and data engineering, 2009, 22(10):1345-1359.

[77] Khosla A, Zhou T, Malisiewicz T, et al. Undoing the damage of dataset bias[C]//European Conference on Computer Vision. 2012: 158-171.

[78] Ghifary M, Balduzzi D, Kleijn W B, et al. Scatter component analysis: A unified framework for domain adaptation and domain generalization[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(7):1414-1430.

[79] Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature representation [C]//International Conference on Machine Learning. 2013: 10-18.

[80] Xu Z, Li W, Niu L, et al. Exploiting low-rank structure from latent domains for domain generalization[C]//European Conference on Computer Vision. 2014: 628-643.

[81] Ghifary M, Bastiaan Kleijn W, Zhang M, et al. Domain generalization for object recognition with multi-task autoencoders[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2551-2559.

[82] Gan C, Yang T, Gong B. Learning attributes equals multi-source domain generalization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 87-97.

[83] Li D, Yang Y, Song Y Z, et al. Deeper, broader and artier domain generalization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5542-5550.

[84] Blanchard G, Deshmukh A A, Dogan U, et al. Domain generalization by marginal transfer learning[J]. arXiv preprint arXiv:1711.07910, 2017.

[85] Ding Z, Fu Y. Deep domain generalization with structured low-rank constraint[J]. IEEE Transactions on Image Processing, 2018, 27(1):304-313.

[86] Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals[J]. Journal of the Royal Statistical Society: Series B, 2016, 78 (5):947-1012.

[87] Rojas-Carulla M, Schölkopf B, Turner R, et al. Invariant models for causal transfer learning[J]. The Journal of Machine Learning Research, 2018, 19(1):1309-1342.

[88] Been Kim R K, Koyejo S. Examples are not enough, learn to criticize! criticism for interpretability[C]//Advances in Neural Information Processing Systems. 2016.

[89] Imbens G W, Rubin D B. Causal inference in statistics, social, and biomedical sciences[M]. : Cambridge University Press, 2015

[90] Pearl J. Causality[M]. : Cambridge university press, 2009

[91] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology[J]. IEEE transactions on systems, man, and cybernetics, 1991, 21(3):660-674.

[92] Kim B, Rudin C, Shah J A. The bayesian case model: A generative approach for case-based reasoning and prototype classification[C]//Advances in Neural Information Processing Systems. 2014: 1952-1960.

[93] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1996:267-288.

[94] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan):993-1022.

[95] Pearl J. Theoretical impediments to machine learning with seven sparks from the causal revolution[J]. arXiv preprint arXiv:1801.04016, 2018.

[96] Bau D, Zhou B, Khosla A, et al. Network dissection: Quantifying interpretability of deep visual representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6541-6549.

[97] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. 2014: 818-833.

[98] Hendricks L A, Akata Z, Rohrbach M, et al. Generating visual explanations[C]//European Conference on Computer Vision. 2016: 3-19.

[99] Ba J, Caruana R. Do deep nets really need to be deep?[C]//Advances in neural information processing systems. 2014: 2654-2662.

[100] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. stat, 2015, 1050: 9.

[101] Bucilu C, Caruana R, Niculescu-Mizil A. Model compression[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006: 535-541.

[102] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.

[103] Smilkov D, Thorat N, Kim B, et al. Smoothgrad: removing noise by adding noise[J]. arXiv preprint arXiv:1706.03825, 2017.

[104] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks[C]//International Conference on Machine Learning. 2017: 3319-3328.

[105] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]//International Conference on Machine Learning. 2017: 1885-1894.

[106] Adler P, Falk C, Friedler S A, et al. Auditing black-box models for indirect influence[J]. Knowledge and Information Systems, 2018, 54(1):95-122.

[107] Erhan D, Bengio Y, Courville A, et al. Visualizing higher-layer features of a deep network[J]. University of Montreal, 2009, 1341(3):1.

[108] Holland P W. Statistics and causal inference[J]. Journal of the American statistical Association, 1986, 81(396):945-960.

[109] Lewis R, Reiley D. Retail advertising works! measuring the effects of advertising on sales via a controlled experiment on yahoo![J]. 2009.

[110] Kohavi R, Longbotham R. Unexpected results in online controlled experiments[J]. ACM SIGKDD Explorations Newsletter, 2011, 12(2):31-35.

[111] Bottou L, Peters J, Candela J Q, et al. Counterfactual reasoning and learning systems: the example of computational advertising.[J]. Journal of Machine Learning Research, 2013, 14(1): 3207-3260.

[112] Hernán M Á, Brumback B, Robins J M. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men[J]. Epidemiology, 2000, 11(5):561-570.

[113] Hernán M A, Brumback B A, Robins J M. Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures[J]. Statistics in medicine, 2002, 21.

[114] Reis D, Landeiro V, Culotta, et al. Using matched samples to estimate the effects of exercise on mental health from twitter[C]//AAAI. 2015: 182-188.

[115] Lechner M. Earnings and employment effects of continuous gff-the-job training in east germany after unification[J]. Journal of Business & Economic Statistics, 1999, 17(1):74-90.

[116] Brookhart M A, Schneeweiss S, Rothman K J, et al. Variable selection for propensity score models[J]. American journal of epidemiology, 2006, 163(12):1149-1156.

[117] VanderWeele T J, Shpitser I. A new criterion for confounder selection[J]. Biometrics, 2011, 67 (4):1406-1413.

[118] Sauer B C, Brookhart M A, Roy J, et al. A review of covariate selection for non-experimental comparative effectiveness research[J]. Pharmacoepidemiology and drug safety, 2013, 22(11): 1139-1145.

[119] Lee B K, Lessler J, Stuart E A. Improving propensity score weighting using machine learning [J]. Statistics in medicine, 2010, 29(3):337-346.

[120] Su X, Kang J, Fan J, et al. Facilitating score and causal inference trees for large observational studies[J]. JMLR, 2012, 13(1):2955-2994.

[121] Rosenbaum P R. Model-based direct adjustment[J]. JASA, 1987, 82.

[122] Basu A, Polsky D, Manning W G. Use of propensity scores in non-linear response models: the case for health care expenditures[R]. : National Bureau of Economic Research, 2008.

[123] Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects[J]. PNAS, 2016, 113(27):7353-7360.

[124] Athey S, Imbens G W. Machine learning methods for estimating heterogeneous causal effects [J]. stat, 2015, 1050:5.

[125] Parikh N, Boyd S. Proximal algorithms[J]. Foundations and Trends in optimization, 2013, 1 (3):123-231.

[126] MacKinnon D P, Lockwood C M, Hoffman J M, et al. A comparison of methods to test mediation and other intervening variable effects.[J]. Psychological methods, 2002, 7(1):83.

[127] Hainmueller J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies[J]. Political Analysis, 2012, 20(1):25-46.

[128] Chernozhukov V, Chetverikov D, Demirer M, et al. Double machine learning for treatment and causal parameters[J]. arXiv preprint arXiv:1608.00060, 2016.

[129] Farrell M H. Robust inference on average treatment effects with possibly more covariates than observations[J]. Journal of Econometrics, 2015, 189(1):1-23.

[130] McCaffrey D F, Ridgeway G, Morral A R. Propensity score estimation with boosted regression for evaluating causal effects in observational studies.[J]. Psychological methods, 2004, 9(4): 403.

[131] Westreich D, Lessler J, Funk M J. Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression[J]. Journal of clinical epidemiology, 2010, 63(8):826-833.

[132] Parikh N, Boyd S P, et al. Proximal algorithms.[J]. Foundations and Trends in optimization, 2014, 1(3):127-239.

[133] Rolling C A, Yang Y. Model selection for estimating treatment effects[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 76(4):749-769.

[134] Kuang K, Cui P, Li B, et al. Treatment effect estimation with data-driven variable decomposition. [C]//AAAI. 2017: 140-146.

[135] Diamond A, Sekhon J S. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies[J]. Review of Economics and Statistics, 2013, 95(3):932-945.

[136] LaLonde R J. Evaluating the econometric evaluations of training programs with experimental data[J]. The American economic review, 1986:604-620.

[137] Torkkola K. Feature extraction by non-parametric mutual information maximization[J]. Journal of machine learning research, 2003, 3(Mar):1415-1438.

[138] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on pattern analysis and machine intelligence, 2005, 27(8):1226-1238.

[139] Kuang K, Cui P, Li B, et al. Estimating treatment effect in the wild via differentiated confounder balancing[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 265-274.

[140] Kuang K, Jiang M, Cui P, et al. Effective promotional strategies selection in social media: A data-driven approach[J]. IEEE Transactions on Big Data, 2017.

[141] Yu B, et al. Stability[J]. Bernoulli, 2013, 19(4):1484-1500.

[142] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks[C]// Advances in neural information processing systems. 2007: 153-160.

[143] Johnson W B, Lindenstrauss J. Extensions of lipschitz mappings into a hilbert space[J]. Contemporary mathematics, 1984, 26(189-206):1.

[144] Menard S. Applied logistic regression analysis: volume 106[M]. : Sage, 2002

[145] Chen Y, Lin Z, Zhao X, et al. Deep learning-based classification of hyperspectral data[J]. IEEE Journal of Selected topics in applied earth observations and remote sensing, 2014, 7(6): 2094-2107.

[146] Thomee B, Shamma D A, Friedland G, et al. Yfcc100m: The new data in multimedia research [J]. Communications of the ACM, 2016, 59(2):64-73.

[147] Csurka G, Dance C, Fan L, et al. Visual categorization with bags of keypoints[C]//Workshop on statistical learning in computer vision, ECCV: volume 1. 2004: 1-2.

致 谢

衷心感谢我的导师杨士强教授和崔鹏副教授对本人的精心指导与帮助。您们不仅是良师，在科研上对我悉心指导，教会我很多科学思考方式和科研方法。记得博士刚入学时，我一方面非常渴望在科研的海洋遨游，实现自己的学术追求，一方面又觉得自己还没具备科研的基本素质，缺乏科研上的自信。是您们，以渊博的学识，丰富的科研经验教会我如何做科研，让我在科研的世界里找到自信，并顺利地走上科研的道路。也是您们多年来的言传身教，让我明白什么是好的研究，也意识到科研品味的重要性。亦是益友，在生活上对我关心照顾，每当我遇到困难或挫折时，总能伸出援助之手提供帮助。记得出国交换初期，我非常不适应国外生活，是您们的关心和帮助，让我愉快并顺利地完成了为期一年的学术交流。也非常感谢您们在我毕业求职期间给予的大力支持和帮助。

感谢实验室朱文武教授对我科研的关心和指导。您对学术的追求，科研的态度在令我佩服的同时，也为我树立了标杆，指引我不断前进。您是我终身学习的榜样。感谢黎波副教授在我读博期间对我科研上的悉心指导和帮助。每当我科研上遇到问题，您总会及时给予我指导和建议，帮助我一起思考解决办法，攻克科研难关，也非常感谢您平时对我生活的关心和帮助。感谢王飞副教授对我科研工作的指导和帮助。

在斯坦福大学访学期间，十分感谢我国外导师 Susan Athey 教授对我学习和科研上的悉心指导和帮助。您的敦敦教诲，让我对科研方向有了更清楚的认识，对科研追求有了更崇高的目标。感谢合作者熊若轩（Ruoxuan Xiong）对我科研上的关心和帮助，是你屡次帮助我攻克科研难题。同时也感谢在斯坦福遇到的各位小伙伴对我学习生活的关心与帮助。

感谢蒋朦师兄带我走上科研之路，手把手教我做科研，逐字逐句帮我改论文。感谢余林韵师兄、张天扬师兄和臧承熙师兄跟我一起讨论学术问题，虽然你们跟我科研方向十分不同，但是你们总是不厌其烦地跟我一起探讨我的科研方向，给予我指导和帮助。也非常感谢欧明栋师兄、刘少伟师兄、张文鹏师兄、王岱鑫师兄、王啸师兄和王鑫师兄等对我在学习和生活上的关心和帮助，让我博士生活过得非常愉快。感谢科研小组成员沈哲言同学、邹昊同学、何玥同学和周琳均同学在每周组会上跟我一起讨论、解决科研问题，让我感受到科研的快乐。另外，也感谢实验室与我多年相处的朱定元、涂珂、陈许旻、路云飞、张子威、李国豪等小伙伴的陪伴，让我博士生活丰富多彩。

感谢我的父亲况旭明，母亲黎顺秀，妻子舒琼和其他亲朋好友对我的科研和生活给予的支持、鼓励和关怀。你们的支持，让我没有后顾之忧；你们的鼓励，让我勇往无前，不惧艰险；你们的关怀，让我感受到无尽的幸福。谢谢你们!

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1992 年 11 月 18 日出生于江西省高安市。

2010 年 9 月考入北京理工大学计算机系计算机科学与技术专业，2014 年 7 月本科毕业并获得工学学士学位。

2014 年 9 月免试进入清华大学计算机科学与技术系攻读博士学位至今。

攻读博士学位期间的获奖情况

[1] 国家奖学金，2017

[2] 综合二等奖学金，清华大学，2015

发表的学术论文

[1] **Kun Kuang**, Meng Jiang, Peng Cui, Shiqiang Yang. Steering social media promotions with effective strategies[C]//2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016: 985-990.

[2] **Kun Kuang**, Meng Jiang, Peng Cui, Shiqiang Yang. Effective Promotional Strategies Selection in Social Media: A Data-Driven Approach[J]. IEEE Transactions on Big Data, 2018, 4(4): 487-501.

[3] **Kun Kuang**, Peng Cui, Bo Li, Meng Jiang, Fei Wang, Shiqiang Yang. Treatment effect estimation with data-driven variable decomposition[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[4] **Kun Kuang**, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang. Estimating treatment effect in the wild via differentiated confounder balancing[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 265-274.

[5] **Kun Kuang**, Peng Cui, Susan Athey, Ruoxuan Xiong, Bo Li. Stable Prediction across Unknown Environments[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018: 1617-1626.

[6] Zheyan Shen, Peng Cui, **Kun Kuang***, Bo Li, Peixuan Chen. Causally regularized

learning with agnostic data selection bias[C]//2018 ACM Multimedia Conference on Multimedia Conference. ACM, 2018: 411-419. (⋆ Corresponding author)

[7] Hao Zou, **Kun Kuang**⋆, Peng Cui. Focused Context Balancing for Robust Offline Policy Evaluation[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019.(⋆ Corresponding author)

[8] Jianxin Ma, Peng Cui, **Kun Kuang**, Xin Wang, Wenwu Zhu. Disentangled Graph Convolutional Networks[C]//Proceedings of the 36th International Conference on Machine Learning (ICML), 2019.

[9] **Kun Kuang**, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang. Treatment Effect Estimation via Differentiated Confounder Balancing and Regression. (Submitted to ACM Transactions on Knowledge Discovery from Data (TKDD))

[10] **Kun Kuang**, Ruoxuan Xiong, Peng Cui, Susan Athey, Bo Li. Stable Prediction across Unknown Environments. (Submitted to The Journal of Machine Learning Research (JMLR))

[11] **Kun Kuang**, Ruoxuan Xiong, Peng Cui, Susan Athey, Bo Li. Stable Prediction with Model Misspecification. (Submitted to NeurIPS 2019)

[12] Zheyan Shen, Peng Cui, Tong Zhang, **Kun Kuang**. Stable Learning of Linear Models via Sample Reweighting. (Submitted to NeurIPS 2019)