



**专项设计报告**

**变分自编码器的演变由来及其**

**在因果推断中的应用**

学 院 计算机学院

专 业 软件工程

年级班别 2018级（5）班

学 号 3118005164

学生姓名 杨泽勤

指导教师 刘鹏

2021年 10 月

**变分自编码器的演变由来及其在因果推断中的应用**

**杨泽勤**

**计算机学院**

# 摘 要

在机器学习领域，密度估计是一类无监督学习问题，其任务是根据数据集来估计概率密度函数。然而有的时候直接对概率密度函数建模是非常困难的，因此我们通常通过引入隐变量来简化模型。如果要建模含有隐变量的分布，传统方法是使用EM算法，当E步的后验概率无法计算的时候，通常引入变分推断，得到变分EM算法，再利用神经网络对分布进行建模，就可以得到变分自编码器。本文的第一部分工作就是基于变分推断的视角，从EM算法出发，一步步导出变分自编码器。

近年来，基于观察数据估计因果效应已成为一个更具有吸引力的研究课题。潜在结果框架是目前应用最为广泛的因果推断框架之一，该框架要回答的核心问题是一个反事实估计的问题。然而，现有的大部分在潜在结果框架下进行因果推断都是在满足SUTVA、ignorability和positivity三个假设的前提下提出的。在实践中，第二个假设往往很难成立。本文的第二部分工作就是以变分自编码器为基础，当存在未观察到的混杂因素的时候，利用代理变量恢复混杂并估计因果效应。

**关键词：**EM算法，变分推断，变分自编码器，因果效应估计

# Abstract

In the field of machine learning, density estimation is a type of unsupervised learning problem. Its task is to estimate the probability density function based on the dataset. However, sometimes it is very difficult to directly model the probability density function , so we usually simplify the model by introducing a hidden variable . If we want to model the distribution with hidden variables, the traditional method is to use the EM algorithm. When the posterior probability of E step cannot be calculated, the variational inference is usually introduced to obtain the variational EM algorithm, and when the neural network is used to model the distribution. we can get the variational autoencoder. The first part of the work is based on the perspective of variational inference, starting from the EM algorithm, step by step derivation of the variational autoencoder.

In recent years, estimating causal effects based on observational data has become a more attractive research topic. The potential outcome framework is currently one of the most widely used causal inference frameworks. The core question that the framework needs to answer is a counterfactual estimation problem. However, most of the existing causal inferences under the framework of potential outcomes are proposed under the premise of satisfying the three assumptions:SUTVA, ignorability and positivity. In practice, the second assumption is often difficult to establish. The second part of this paper is based on the variational autoencoder, use proxy variables to restore confounding and estimate causal effects when there exists unobserved confounding factors.

**Key words:** EM algorithm, variational inference, variational autoencoder, causal effect estimation

**目 录**

[1设计任务 1](#_Toc86680623)

[1.1从EM到变分自编码器的演变 1](#_Toc86680624)

[1.2变分自编码器在因果推断中的应用 1](#_Toc86680625)

[2专项设计的时间计划安排 3](#_Toc86680626)

[3任务一：从EM到变分自编码器的演变 4](#_Toc86680627)

[3.1 EM算法 4](#_Toc86680628)

[3.2变分推断 5](#_Toc86680629)

[3.3 变分EM 7](#_Toc86680630)

[3.4 变分自编码器 7](#_Toc86680631)

[4 任务二：变分自编码器在因果推断中的应用 9](#_Toc86680632)

[4.1 潜在结果框架简介 9](#_Toc86680633)

[4.2 发展现状与设计任务二 9](#_Toc86680634)

[4.3 CEVAE-因果效应变分自编码器 11](#_Toc86680635)

[5 总结与展望 15](#_Toc86680636)

[6 参考资料 16](#_Toc86680637)

[7 致 谢 17](#_Toc86680638)

# 1设计任务

## 1.1从EM到变分自编码器的演变

在机器学习中，密度估计是一类无监督学习问题：给定一组数据，假设它们都是独立地从相同的概率密度函数中产生的。密度估计的任务就是根据数据集来估计其概率密度函数。举个例子，在手写体数字图像的密度估计中，我们将图像表示为一个随机变量，其中每一维表示一个像素值。假设手写体数字图像服从一个未知分布，希望通过一些观测样本来估计该分布。然而手写体数字图像的不同像素之间存在复杂的依赖关系，很难用一个图模型来描述其依赖关系，所以直接对建模是非常困难的。因此我们通常通过引入隐变量来简化模型，如此密度估计问题可以转换为估计变量的两个局部条件概率和。一般为了简化模型，假设隐变量的先验分布是标准高斯分布且隐变量的每一维之间是独立的，在这个假设下先验分布不存在参数，因此密度估计的重点是估计条件分布。

如果要建模含有隐变量的分布，一般利用EM算法来进行密度估计。在EM算法中，需要估计条件分布和后验分布。问题是，当后验分布无法计算的时候，如何利用变分推断求取近似后验分布？当这两个分布比较复杂的时候，如何利用神经网络来进行建模？对于第一个问题，答案是变分EM算法；对于第二个问题，答案是变分自编码器。因此，此专项设计的第一个任务就是理清楚从EM到变分EM，再到变分自编码器的演变过程，专项设计主要是从理论上进行学习。

## 1.2变分自编码器在因果推断中的应用

因果推断是一个跨越许多领域的关键研究课题，在例如统计学、计算机科学、教育、公共政策和经济学等学科中发挥着重要的作用。通常来说，因果推断最有效的方法是进行随机对照试验，将参与者随机分配到治疗组或对照组，随机对照实验对治疗策略的分配是完全随机的，从而能够保证治疗组和对照组之间的唯一区别只有治疗变量。然而在现实中，随机对照试验存在以下问题：一是费时费力，代价昂贵；二是只关注参与者的平均表现，无法解释针对于个体的因果机制；三是可能涉及一些伦理道德问题。

因此，相比于传统通过进行随机对照试验来估计因果效应，如何基于观察数据估计因果效应已成为一个更具有吸引力的研究课题。潜在结果框架是目前应用最为广泛的因果推断框架之一，该框架要回答的核心问题是一个反事实估计的问题，例如我们想回答“如果这个病人当初接受了另外一种药物治疗，他会有不同的结果吗？” 回答这些反事实的问题是具有挑战性的，有两个原因：一是我们只观察到了事实结果，而不是如果他们选择了不同的治疗方案，可能发生的反事实结果；二是在观察数据中，治疗策略通常不是随机分配的，这可能导致治疗组的协变量分布相较于对照组发生了分布偏移，我们称这种情况为存在混杂。

得益于机器学习、深度学习领域的快速发展，各种基于观测数据的因果效应估计方法已经涌现出来。首先，针对于反事实估计问题，借助诸如决策树、集成方法、深度神经网络等强大的模型方法，我们可以更准确地估计反事实结果；其次，机器学习方法也为处理混杂提供了一个新的视角：受益于深度表征学习方法的发展，我们可以通过学习所有协变量的平衡表征来调整混杂变量，使得在学习到的平衡表征的条件下，治疗分配策略独立于混杂。

然而，现有的大部分基于观察数据的因果效应估计方法都是在满足SUTVA、ignorability和positivity三个假设的前提下提出的。然而在实践中，针对于第二个假设，我们往往是无法保证的。那么当存在未观察到的混杂因素的时候，如何基于变分自编码器，利用代理变量恢复混杂并估计因果效应？这就是专项设计的第二个任务，这份报告主要是从理论上进行学习。

报告的结构如下：第三节解决的是第一个任务，依次由EM算法、变分推断、变分EM和变分自编码器四小节组成；第四节解决的是第一个任务，4.1节简单介绍了潜在结果框架，4.2节对现阶段，基于潜在结果框架的因果效应估计方法做了总结，并明确了第二个任务，4.3节介绍CEVAE-因果效应变分自编码器；第五节是对此专项设计的一些总结与展望。

# 2专项设计的时间计划安排

**表2.1 时间计划安排表**

|  |  |
| --- | --- |
| 时间 | 任务 |
| 9月23号~9月24号 | 理解EM算法 |
| 9月25号~9月29号 | 理解变分推断 |
| 9月30号 | 理解变分EM |
| 10月1号~10月2号 | 理解变分自编码器 |
| 10月3号~10月5号 | 学习潜在结果框架的基本知识 |
| 10月6号~10月7号 | 总结现阶段基于潜在结果框架  的因果效应估计方法 |
| 10月8号 | 学习因果效应变分自编码器 |
| 10月9号 | 撰写报告 |

# 3任务一：从EM到变分自编码器的演变

## 3.1 EM算法

EM算法用于求解带有隐变量的参数估计问题，其流程就是反复迭代图3.1中的(b)和(c)，分别表示E步和M步[1]：

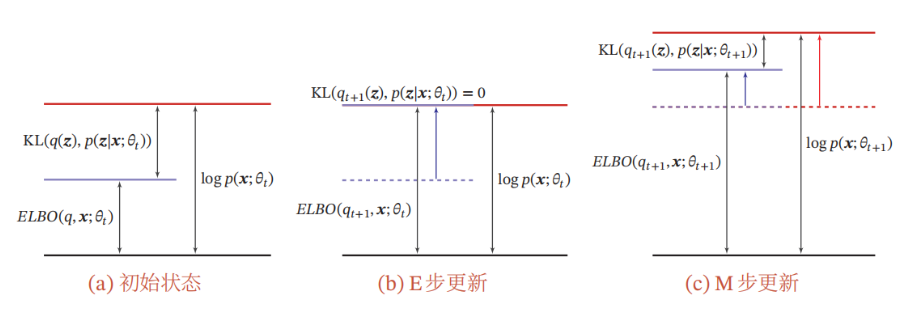


图3.1 EM算法流程图

具体地，对数边际似然可以分解为：

显然，可以按照图3.1所示，不断重复E、M两步使得逐渐变大：

(a)图：第t次迭代的初始状态，此时参数为，并且通常有；

(b)图：E步。固定参数，找到分布尽可能接近后验，使得尽可能小，从而使得尽可能接近。在传统EM中，考虑的是理想情况，即令使得，这时；

(c)图：M步。固定分布，寻找参数，使得最大。

## 3.2变分推断

变分推断的核心思想是：我们感兴趣的目标是推断出后验分布。但是当我们的目标后验概率不可计算时，变分推断的核心思想是引入参数化的变分分布族，将inference问题转化为optimization问题。如图3.2所示[2]：

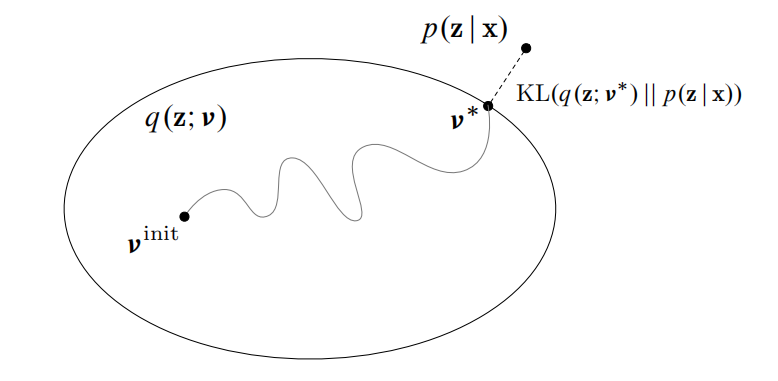


图3.2 变分推断核心思想图

由于后验无法直接推断，故引入变分分布族，参数为（图3.2中的椭圆为其参数空间），再利用KL散度作为两者之间距离的度量。如此，inference问题转化为如下的泛函optimization问题：

公式第一行转化到第二行的原因是：在KL散度中，仍然是不可计算的，所以根据式（1）可以将其转化到第二行的形式。

在变分推断中，变分分布族的复杂性决定了优化问题的复杂性。一个最简单的假设就是平均场假设，即可以拆解成多组相互独立的变量，即：

如此，可以通过坐标上升法迭代地优化每个：

进一步地，我们可以假设Conditionally conjugate models，这样式（4）具有闭式解。

如果式（4）没有闭式解或者平均场假设太强不适应怎么办呢？回退到式（2），我们可以采用黑盒变分推断求解，具体地，可以分为Score gradients和Reparameterization gradients两类方法：

### Score gradients

首先将关于参数的梯度写成期望的形式：

之后，采用MCMC的思路：先从变分分布中采样，再依次计算得到（取多个样本取平均值），最后进行梯度更新。

### Reparameterization gradients

假设变分分布可以由变换得到，且变分参数是变换的一部分，具体地：

举个高斯分布的例子：

类似Score gradients，Reparameterization gradients也可以将ELBO关于参数的梯度写成期望的形式，不过不再是关于的期望，而是关于的期望：

之后的思路和Score gradients一致，不过换成了从中采样。

## 3.3 变分EM

变分EM其实非常简单。前面已经提到，传统EM的E步的目的是最小化，且其假设的是最理想的情况：后验可计算。那么当后验并非可计算的时候怎么办呢？巧了，变分推断解决的不就是这个问题吗？也就是说，将传统EM的E步用变分推断的思路去做就可以了。而这就是变分EM。

如此，EM的E步和M步的目标都是最大化ELBO，不过优化的参数不同：E步是；M步是。

## 3.4 变分自编码器

我们可以发现，其实变分自编码器可以看作是变分EM的神经网络版本。只不过还需要补充下面几点：

(1)E步的变分推断具体采用的是Reparameterization gradients；

(2)变分EM的E步中，每个隐变量对应一个参数，是通过梯度下降不断优化得到的，注意，梯度下降优化的对象是；而变分自编码器将E步当作了推断网络，网络的输出是参数，反向传播优化的是神经网络的权重，也就是说将隐变量对应的参数又进行了一次参数化，通过一个共享的神经网络来输出所有样本的变分参数，这也叫做Amortized Inference；

(3)不仅把E步当作推断网络，还把M步当作生成网络，如此一趟正向-反向传播下来，变分参数和模型参数同时得到了优化。

因此，把变分EM的E步和M步合起来，也就得到了变分自编码器的目标函数（分别表示推断网络和生成网络的权重参数）：

式(8)最后两行是Reparameterization的过程，。

最后，变分自编码器的网络结构如图3.3所示：

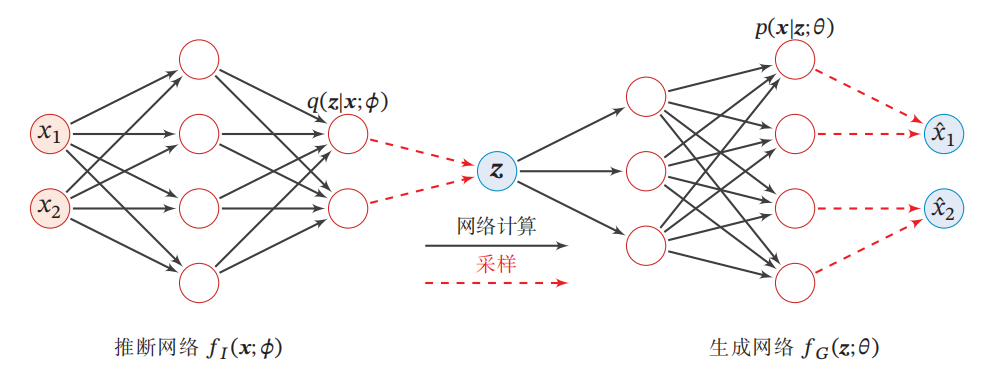


图3.3 变分自编码器网络结构图

# 4 任务二：变分自编码器在因果推断中的应用

## 4.1 潜在结果框架简介

潜在结果框架下的一个重要任务是因果效应的计算，即在给定观察数据的情况下，基于观察数据计算因果效应。因果效应可以在不同的尺度上计算，例如总人群、治疗组、子人群和个体，因此对应的，有以下四种常见的因果效应：

(1)平均治疗效应（Average Treatment Effect，ATE）：

(2)治疗组平均因果效应（Average Treatment effect on the Treated group，ATT）：

(3)条件平均治疗效应（Conditional Average Treatment Effect，CATE）：

(4)个体治疗效应（Individual Treatment Effect，ITE）：

上述公式中，分别表示接受治疗和不接受治疗的潜在结果。

为了估计因果效应，现有的大部分工作都做了以下三个假设：

(1)稳定个体治疗数值假设（Stable Unit Treatment Value Assumption，SUTVA）：任何个体的潜在结果都不因分配给其他个体的治疗策略变化而变化，而且，对于每个个体，每个治疗策略都没有不同的形式或版本，从而导致不同的潜在结果；

(2)可忽略性假设（Ignorability）：给定协变量，治疗策略独立于潜在结果；

(3)积极性假设（Positivity）：对于协变量的任意一种取值，治疗策略都不是确定性的，即。

然而SUTVA、ignorability和positivity这三个假设在实践中往往并不成立，如社交媒体分析，它往往涉及到特殊的数据结构（网络拓扑）、特殊的数据类型（时间序列数据）或特定的条件（未观察到的混杂因素的存在）。

## 4.2 发展现状与设计任务二

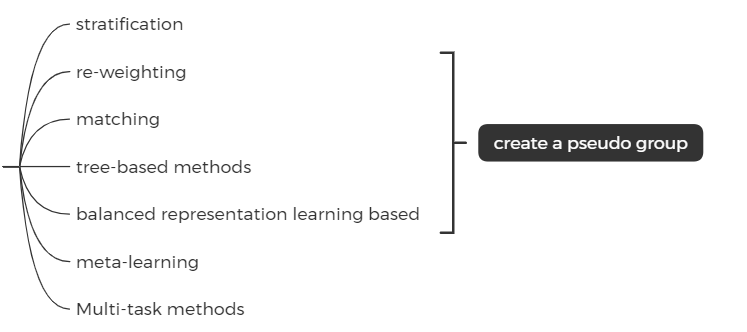


图4.1 现有的因果效应估计方法

现有的因果效应估计方法大多是需要满足SUTVA、ignorability和positivity三个假设，具体可分为如图4.1所示的七类：

(1)重加权方法（Re-weighting Methods）：通过给观察数据中的每个个体分配适当的权重，从而创建一个伪种群，在该伪种群中治疗组和对照组的分布是相似的，不存在分布偏移；

(2)分层方法（Stratification Methods）：通过将整个人群划分为同构的子群（块），来调整由治疗组和对照组之间的差异而产生的偏差，本质上其实是后门准则；

(3)匹配方法（Matching Methods）：基于某种距离度量，对于治疗组的个体，在处理组中寻找与其最接近的个体作为其反事实结果；对于处理组的个体，在治疗组中寻找与其最接近的个体作为其反事实结果；

(4)基于树模型的方法（Tree-based Methods）：基于决策树和集成方法提出，著名的算法有贝叶斯加性回归树（ Bayesian Additive Regression Trees，BART）和因果森林（Causal Forest）；

(5)表征学习方法（Representation Learning Methods）：反事实分布通常会不同于事实分布。因此，通过从事实数据中学习来预测反事实的结果，本质上是将因果推理问题转化为领域适应问题；

(6)多任务学习方法（ Multitask Learning Methods）：治疗组和对照组除其特殊特征外，总是有一些共同的特征。因此，因果推理可以被定义为一个多任务学习问题，治疗组和对照组先有一组共享层，再分别有一组特定层；

(7)元学习方法（ Meta-Learning Methods）：包括T-learner，S-learner，X-learner，U-learner，R-learner。

然而，在实际应用中，上述三个假设往往是不成立的。因此，本次专项设计的第二个任务聚焦于基于变分自编码器，如何打破上述的ignorability假设进行因果效应估计。

可忽略性假设（Ignorability）也叫做无混杂假设（unconfoundedness），表示给定协变量的情况下，治疗策略独立于潜在结果。也就是说，对于具有协变量的个体，他们的治疗分配策略可以被看作是随机的。现有的工作大多依赖于Ignorability假设，即假设所有的混杂因素都可以被测量，然而这一假设在实践中可能是无法满足的，例如”经济社会地位“这一治疗是无法清晰定义和测量的，因此叫做隐藏的混杂。

## 4.3 CEVAE-因果效应变分自编码器

混杂变量的定义是既影响了治疗策略又影响了治疗结果，因果效应估计很关键的一步就是去除混杂。当混杂变量可观察且能被测量到的时候，最直接的去除混杂的方式就是通过协变量调整、重加权等方法去控制它；但是当混杂变量无法观察和测量的时候，一般情况下很难直接估计出治疗策略对结果的因果效应。例如在药物实验中，经济社会地位既影响了病人所选择的药物（治疗策略），又影响了病人的死亡率（结果），因此经济社会地位扮演了混杂变量的角色，但是由于这一指标很难量化测量，所以它其实是一个隐藏的混杂变量，所以在该情况下ignorability假设被打破。

针对这种情况，最常见的一种做法就是寻找“代理变量”，例如在上面的例子中，我们无法测量经济社会地位这一指标，因此可以测量工作类型、年收入、受教育情况等指标作为经济社会地位的“代理变量”。文献[7]和文献[8]给出了利用“代理变量”的方法，能够保证在“代理变量”被观测到的情况下正确恢复的因果效应，但是这些方法依赖于一定的假设，例如要求隐藏的混杂变量是类别变量，要求模型是线性高斯等等。

文献[5]提出了因果效应变分自编码器CEVAE，基于变分自编码器，由“代理变量”恢复出隐藏的混杂变量，该模型的好处是对数据生成过程和隐藏的混杂变量的结构做了很弱化的假设，不过由于利用了变分自编码器，很显然缺点是无法找到正确的模型，陷入局部最优，这是深度学习模型的通病。

如图4.2所示是CEVAE对应的图模型，表示治疗策略，例如药物，表示结果，例如死亡率，表示隐藏的混杂变量，例如经济社会地位，表示的代理变量，例如年收入和居住地。显然在这个图模型中，利用后门准则就可以保证模型的可识别性。

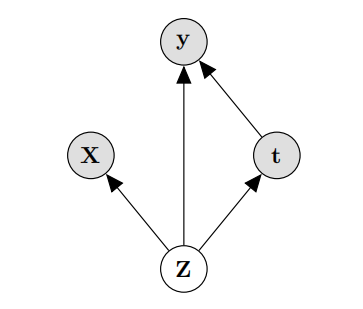


图4.2 CEVAE对应的图模型

CEVAE的网络结构和VAE非常类似，都由一个推断网络和生成网络组成，具体地，CEVAE的网络结构如下图所示。其中白色结点表示普通的神经网络变换，灰色结点表示需要从指定分布中进行采样，白色圆圈表示需要根据治疗策略选择对应的路径。

在推断网络中，为了使得模型能够应用到新个体，需要在推断的后验分布之前知道治疗策略和结果，因此网络结果如图4.3中的（a）所示。假设为服从伯努利分布的离散变量，均为服从高斯分布的连续变量，则有：

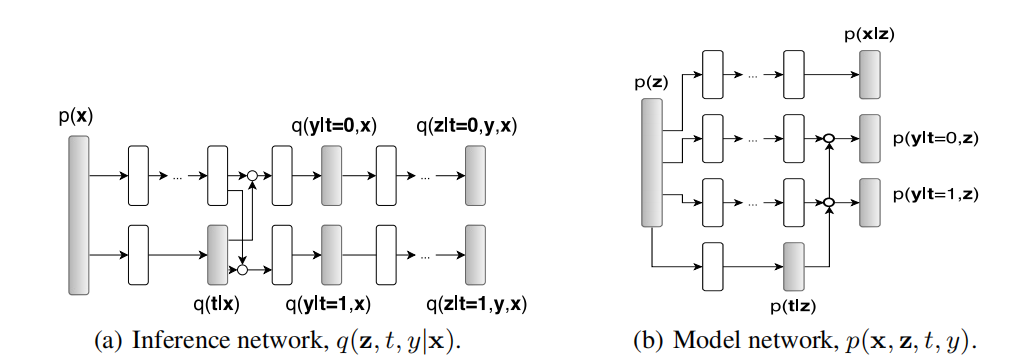


图4.3 CEVAE的网络结构图

在生成网络中，使用TARnet[9]的架构，如图4.3中的（b）所示，则有：

CEVAE的损失函数为：

其中是训练阶段的观测变量，分别表示结果，代理变量和治疗策略。

# 5 总结与展望

总的来说，此次专项设计主要分为两部分：第一部分主要是理清基于变分推断，从EM到变分EM，再到变分自编码器的演变过程，在学习过程中，我对相关内容搭建了完善的知识框架，掌握了变分自编码器这个基础工具，为后续的研究打下坚实的基础；第二部分主要是基于变分自编码器，研究如何在打破ignorability假设的情况下估计因果效应，在学习过程中，我学习了潜在结果框架并调研了基于此框架的各类方法的发展现状，受益匪浅。

但仔细学习了CEVAE[5]后可以发现，其实这个模型是在已有模型的基础上稍作修改，并不复杂。但是这个工作的贡献在于作者针对于鲜有人关注的问题和难点，对问题进行形式化定义之后，仅在现有模型的基础上稍作调整，就能够很好地解决了难题。这给我的启示是：与其对现有模型盲目地进行缝缝补补，不如脚踏实地的做研究。先好好发现一个问题，接着对该问题进行形式化地定义，将其抽象成一个本质上的数学问题，最后根据实际面临的问题对现有工具进行修改去解决问题。

后续的工作方向是针对于因果效应估计领域，先进行理论上的突破，再应用到实际问题上。例如打破了假设的因果效应估计，此次专项设计涉及到的是变分自编码这个工具，但是针对具体的数据类型，能够有更加匹配的模型可以更好地解决这个问题？以打破SUTVA假设为例，本质上就是个体之间加入了“干扰”：如果是在网络数据，要解决的难点其实就是如何聚合一阶邻居、二阶邻居甚至更高阶的邻居的“干扰”信息，而这种网络中信息聚合的方式，能够基于图神经网络来完成？因为它们的目标是一致的；如果是在时序数据，循环神经网络、状态空间模型等模型能够更好地聚合“干扰”信息？这些都是待探索的事情。而一旦理论上有了进展，再落地迁移到诸如推荐系统等实际应用中就是非常容易的事情了。

# 6 参考文献

[1] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, Huan Liu. A Survey of Learning Causality with Data: Problems and Methods. ACM Computing Surveys（2020）.

[2] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, Aidong Zhang. A Survey on Causal Inference. arXiv preprint arXiv:2002.02770, 2020.

[3] David M. Blei, Alp Kucukelbir, Jon D. McAuliffe. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association（2017）.

[4] D. P. Kingma and M. Welling. Auto-encoding variational bayes. International Conference on Learning Representations. In ICLR'14.

[5] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In NIPS’17.

[6] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, Huan Liu. Linked Causal Variational Autoencoder for Inferring Paired Spillover Effects. In CIKM’18.

[7] M. Kuroki and J. Pearl. Measurement bias and effect restoration in causal inference. Technical report, DTIC Document, 2011.

[8] J. Pearl. On measurement bias in causal inference. arXiv preprint arXiv:1203.3504, 2012.

[9] U.Shalit, F.Johansson, and D.Sontag. Estimating individual treatment effect: generalization bounds and algorithms. ArXiv e-prints, June 2016.

# 7 致 谢

本设计是在我的指导教师刘鹏老师的亲切关怀和悉心指导下完成的。他严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。从题目的选择到最终完成，刘鹏老师都始终给予我细心的指导和不懈的支持。

感谢刘鹏老师在课题研究初期给予的帮助，在刘鹏老师的帮助下课题才得以很好的展开，有了一个很好的开端！在本设计完成过程中，本人得到了刘鹏老师和许多师兄的热心帮助，本人向他们表示深深的谢意！