1917–2017
100th Anniversary
Shanghai University of Finance and Economics
上海财经大学100周年校庆

机器学习基本原理及其在因果识别中的潜在应用

郭峰
上海财经大学公共经济与管理学院
2023年10月

目录

一、引言

- 在经济学、社会学和政治学等各类社会科学研究中，特别是最近二十年的社会科学研究中，**识别因果关系**（Causal Relationship）已经成为重中之重（Athey，2017；Imbens and Wooldridge，2009；Athey and Imbens，2017；Abadie and Cattaneo，2018）。

一、引言

- 社会科学研究中，数据来源越来越丰富，文本、图像、音频、视频、遥感等大数据（Big Data）都成为社会科学研究者的重要数据来源。

- 大数据的特征

  - 高维且稀疏

  - 非结构化

  - 非线性

- 大数据广泛应用，为因果识别带来很多挑战

一、引言

- 擅长于处理这种非结构化大数据的机器学习方法也成为社会科学家工具箱的重要组成（Varian，2014；Grimmer，2015；Mullainathan and Spiess，2017；Athey，2018；Athey and Imbens，2019；洪永淼和汪寿阳，2021）。

一、引言

- 机器学习方法的主要优势就在于对包括非结构化数据在内的大数据进行**降维、分类和预测**等（Ghoddusi *et al.*，2019）。

- **机器学习方法和因果识别存在隔阂：**预测中，仅仅需要知道变量之间存在相关关系（Kleinberg *et al.*，2015），因此很多机器学习算法也就忽略了变量间的因果关系，而只关心结果变量和特征变量之间是否存在相关关系。

- 不过，机器学习方法与因果关系识别之间并不全然是冲突的，凭借其在处理高维数据、非线性关系等上的优势，以及在进行变量预测等方面所取得的成功，机器学习方法对因果关系识别也有非常重要的价值。

一、引言

- 机器学习代表性综述：Varian（2014）、Grimmer（2015）、Mullainathan and Spiess（2017）、Athey（2018）、Athey and Imbens（2019）、Ghoddusi *et al.*（2019）、Storm *et al.*（2019）、黄乃静和于明哲（2018）、王芳等（2020）。

- 机器学习代表性教科书：Hastie *et al.*（2017）、James *et al.*（2013）、Burkov（2019）

- 面向机器学习研究者，介绍因果关系文献不少，例如：Schölkopf（2019）、Kreif and DiazOrdaz（2019）、Guo *et al.*（2020）、Varian（2016）。

- 但面向熟悉因果识别，但不熟悉机器学习的社会科学工作者介绍机器学习及其在因果识别中的意义的文献不多。我们进行了这个尝试，撰写了这篇综述文章。

二、机器学习的原理与应用

- 计量经济学重点：参数估计、因果推断

- 相对于计量经济学，机器学习不仅仅是使用了不同的方法（很多方法其实是重叠的），更重要的是关注点不同。

- 譬如，传统社会科学实证更关心无偏性。为了实现无偏估计，不知道也无法获得数据的真实分布，最好的策略是建立一个非常复杂的模型，以尽可能实现一致估计。但这种情形下，模型通常会“过度拟合”样本数据，从而导致在样本以外的数据无效（Yarkoni and Westfall，2017）。

- 相对于计量经济学，机器学习更加关注模型的预测能力：结论是否可以外推（范化，generalize）

二、机器学习的原理与应用

- 线性回归： $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$
- 参数估计： $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \mathrm{L} + \hat{\beta}_{ki} X_{ki}$
- 基于以下公式的最小化： $\sum (Y_i - \hat{Y}_i)^2$
- 估计好参数后，对于全新的X， 就是其预测 $\hat{Y}_i$
- 计量经济学关注的是参数估计$\hat{\beta}$，机器学习关注的是 $\hat{y}$

二、机器学习的原理与应用

- 为了检验泛化能力，机器学习将数据分成两组:训练集和测试集

将数据对象进行特征（feature）化表示

给定一个数据样本集，从中学习出规律（模型）

目标：该规律不仅适用于训练数据，也适用于未知数据（称为泛化能力）

对于一个新的数据样本，利用学到的模型进行预测

二、机器学习的原理与应用

- 从计量经济学视角看机器学习

- 给定一个未知函数$f(x_i,\beta)$，机器学习的目标是通过训练数据$(y_i,x_i)_{i=1,\dots,n}$来学习未知函数$f(x_i,\beta)$

- 其中，β为未知参数（甚至不存在，非参估计）

- 具体而言，根据训练数据，找到一个函数$\hat{y_i}=f(\hat{x_i},\beta)$

- 使得所做的预测$\hat{y_i}$与实际的y“差距”最小。“均方误差”是这个“差距”的常见形式，但不唯一

- 然后再检验测试集中，上述函数表现如何

二、机器学习的原理与应用

- **机器学习与计量经济学的术语差异**
  - “自变量”或“解释变量”，但机器学习则称为“表征”或“特征”（Features）。
  - “因变量”或“被解释变量”，而机器学习则称为“响应”（Response）。
  - 数据为“观测值”（Observation），而机器学习则直接称为“案例”（Example）
  - 对于我们计量经济学的“模型”，他们一律称之为为“算法”。

二、机器学习的原理与应用

- **过拟合与欠拟合**
  - 在监督学习中，我们的目的是在训练数据中构建模型，然后能够对没见过的新数据做出准确预测。
  - 如果一个模型能够对没见过的数据做出准确预测，我们就说它能够从训练集范化（generalize)到测试集。
  - 过拟合（over-fitting）：机器学习算法在训练样本中表现得过于优越，导致在测试数据集中表现不佳。
  - 欠拟合（under-fitting）：如果模型过于简单，无法捕抓数据的全部内容以及数据的变化，在训练集就表现很差，更别提测试数据。

二、机器学习的原理与应用

- **过拟合与欠拟合：直观举例**

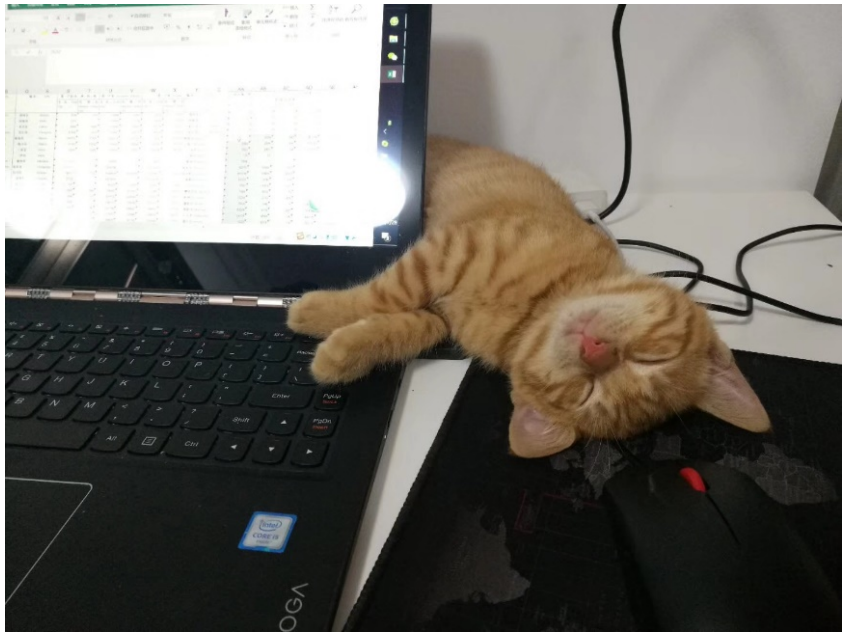
  - 假如进行图形识别的机器学习，从很多很多图片中，挑选出猫的图片

  - 把这些图片调整进行提取：腿长、鼻子大小、耳朵大小等等（这就是计量中的解释变量，机器学习中的特征）

二、机器学习的原理与应用

- 假如训练样本中的所有训练图片都是上述猫品种，那么经过多次迭代训练之后，模型训练好了，并且在训练集中表现得很好。基本上该猫身上的所有特点都涵括进去，甚至猫的颜色都囊括了。

- 但如果测试样本是如下一个小猫。那么很有可能模型最后输出的结果就是该猫猫不是一条猫

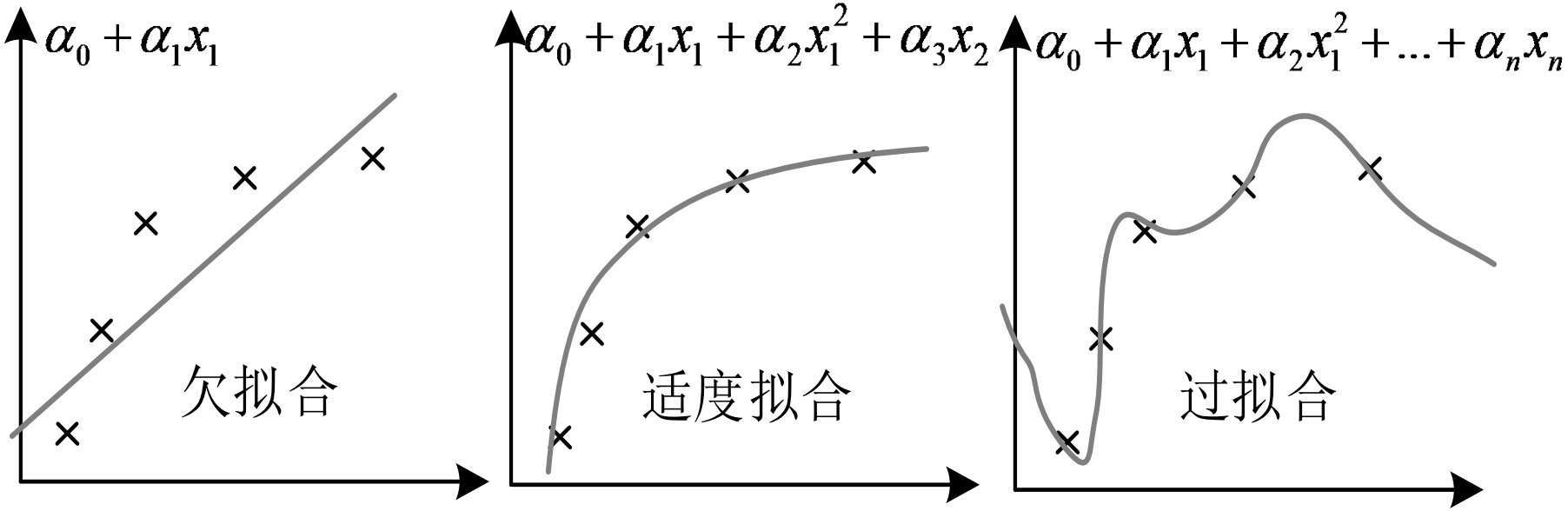


二、机器学习的原理与应用

- 猫科动物：其体型中、大，躯体均匀，四肢中长，趾行性。头大而圆，吻部较短。

二、机器学习的原理与应用

- 过拟合与欠拟合：回归



二、机器学习的原理与应用

- **重要的事情说三遍**
- 当某个模型过度的学习训练数据中的细节和噪音，以至于模型在新的数据上表现很差，我们称发生了过拟合。
- 这意味着训练数据中的噪音或者随机波动也被当做概念被模型学习了。而问题就在于这些概念不适用于新的数据，从而导致模型泛化性能的变差。
- 过拟合更可能在无参数非线性模型中发生，因为学习目标函数的过程是易变的具有弹性的。同样的，许多的无参数器学习算法也包括限制约束模型学习概念多少的参数或者技巧。
- 造成过拟合的原因有可以归结为：模型过于复杂，参数过多。

二、机器学习的原理与应用

- **从OLS回归到正则化**
  - 为了克服最小二乘法的过度拟合，泛化能力差的问题，可以对OLS的回归参数进行一个惩罚，这个思路就叫做正则化（Regularization)。

  - Ridge回归和Lasso回归都是这个思路当中的一种。具体而言，在目标函数(损失函数）上加上如下的惩罚。其中，参数可以通过交叉验证等方式获取

岭回归：
$$\sum_i (y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso回归：
$$\sum_i (y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_i|$$

  - 牺牲无偏性，提高泛化能力：可以证明存在参数 λ ，使得上述回归泛化能力强于OLS回归

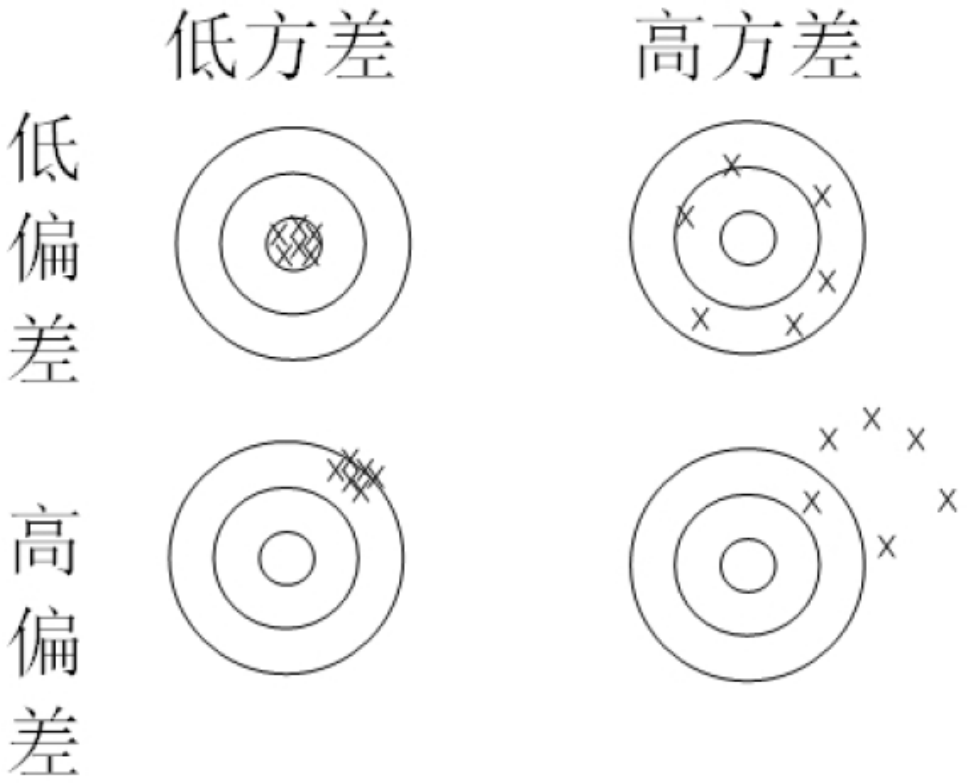
二、机器学习的原理与应用

左上角的低偏差、低方差情形为最为理想的模型，其估计值总在*真实值*附近。

右上角的模型虽然平均而言系统偏差很小，但方差很大，故经常偏离靶心，存在“过拟合”(overfit)。

左下角的模型则正好相反，虽然方差很小，几乎总打在相同的地方，但遗憾的是此地并非靶心，故偏差较大，存在“欠拟合”(underfit)。

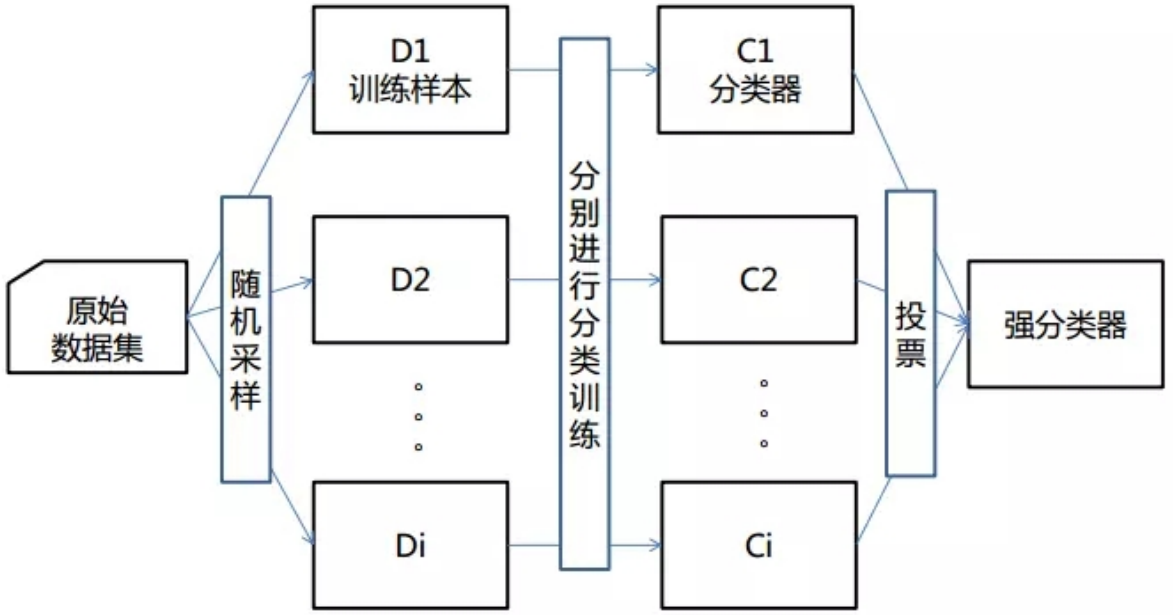
右下角的模型则偏差与方差都较大，不仅存在较大系统偏差，而且波动幅度大，故是最糟糕的模型。



二、机器学习的原理与应用

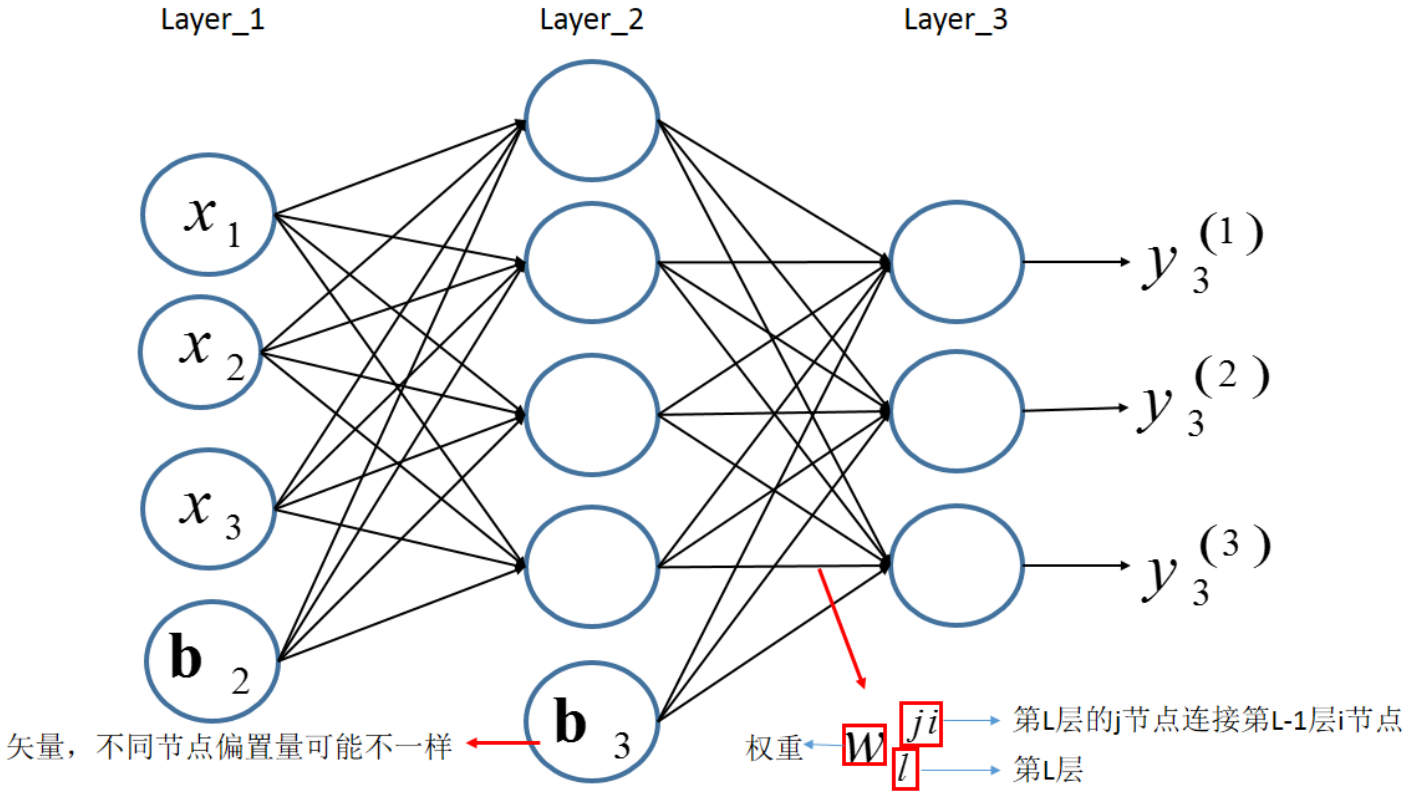
- 为了更好预测，机器学习什么事都干得出来，甚至会放弃可解释性

- 比如，不用全样本回归（分类），而是对原始样本、特征进行随机抽样，N次抽样得到回归（分类）结果，求平均（投票），算作这个集成算法的结果。（随机森林）



二、机器学习的原理与应用

- 为了提高泛化能力，机器学习算法还可以对算法进行多层嵌套（深度学习）



二、机器学习的原理与应用

- **机器学习的优势：**

  •处理高维数据，筛选变量（识别混淆变量，挑选工具变量）

  •处理非结构化数据

  •处理非线性关系

  •注重检验泛化性能（预测准确性）

  •异质性估计

  •数据驱动（模型设定）

二、机器学习的原理与应用

- **我们能用机器学习做什么？**
  - 预测（公司破产概率、经济是否衰退）
  - 数据生成（文本中的情绪、从姓名推测性别）
  - 因果识别（构造反事实结果、异质性检验）

三、机器学习对因果识别的意义

- **机器学习助力更好地控制混淆因素：可观测变量**

  - 选择控制变量常规方法：依据理论分析或理论直觉

  - 存在的问题：（1）控制变量会存在人为操纵，以获得统计上的显著性（Fafchamps and Labonne，2016）；（2）大数据时代依据理论分析或理论直觉控制变量的选取变得非常困难，因为非结构化大数据的一个非常显著的特征就是高维稀疏：潜在的控制变量可能成百上千个，而最终能被用上的可能只有数个。

  - **机器学习的应对：**Belloni et al.（2014；2019）等提出一种称为“Post Double Selection”的数据驱动的策略：首先通过Lasso等附带正则项的机器学习算法，经过交叉验证等方法，识别出一组对结果变量有解释力的变量，进而重新将结果变量对这些挑选出的特征变量进行普通的线性回归。

三、机器学习对因果识别的意义

- **机器学习助力更好地控制混淆因素：工具变量**

  - 我们追求X对y的因果影响，但IV与x的关系，则不需要非要是因果关系

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

$$x_1 = \delta_0 + \delta_1 Z + \delta_2 x_2 + ... + \delta_k x_k + \nu$$

  - 工具变量的寻找本质上是一个预测问题（什么外生变量对X影响最大）。另外，如果潜在的工具变量太多，可能会产生弱工具变量问题，而工具变量数量比内生变量多一至两个时才是最优的（Bollen，2012）

  - **工作步骤：**使用LASSO等方法，进行一阶段的回归，挑选出对x有影响的工具变量IV，然后重新进行普通的两阶段最小二乘法。

三、机器学习对因果识别的意义

- **使用机器学习算法来进行工具变量挑选**

  - Belloni *et al.*（2012）推荐使用Lasso方法，在一些潜在的工具变量池中，挑选与内生变量最为相关的变量作为工具变量，然后重新进行普通的两阶段最小二乘法回归，并证明了这一方法满足相关的统计学的假设。

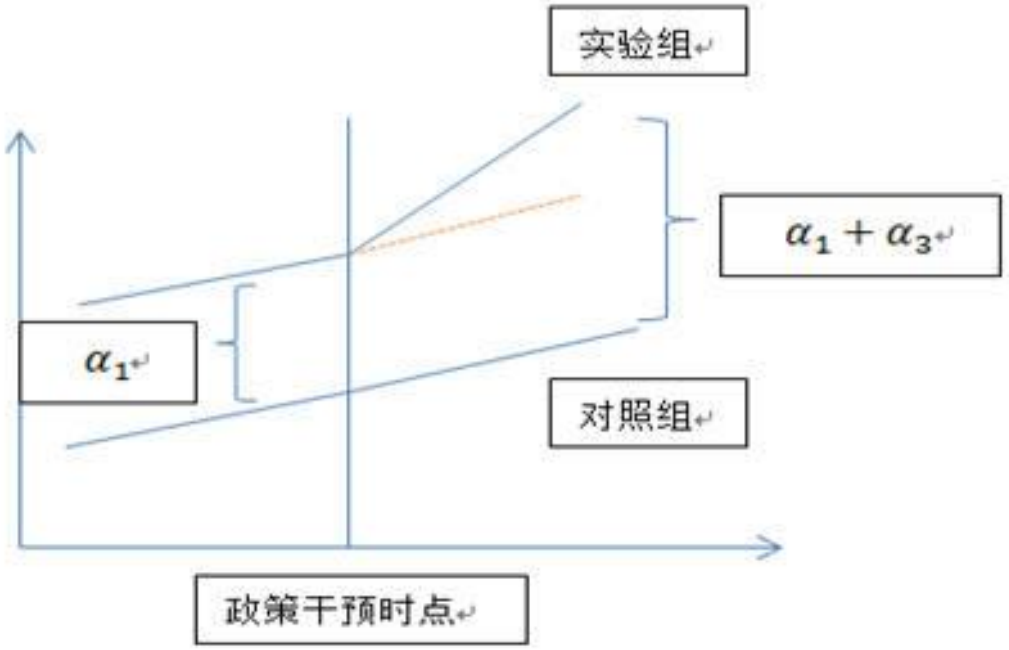
  - Qiu et al.（2020）通过Cluster-Lasso方法在一组天气等变量池中，寻找各地新冠肺炎病例的最佳工具变量

  - Gilchrist and Sands（2016）、方娴和金刚（2020）利用Lasso方法，选择最优的天气与空气污染变量作为电影首映周非预期票房的工具变量，进而考察电影首映周票房对随后几周电影票房的影响。

三、机器学习对因果识别的意义

- 更好地构建对照组：双重差分法

- **双重差分法核心思想：**如果没有政策发生，**处理组和对照组增长趋势保持一致**，那么就可以使用对照组在政策前后的增长趋势，以及处理组在处理前的特征和趋势，来构建如果没有政策发生，处理组的“反事实”结果，进而获得政策的因果效应。



三、机器学习对因果识别的意义

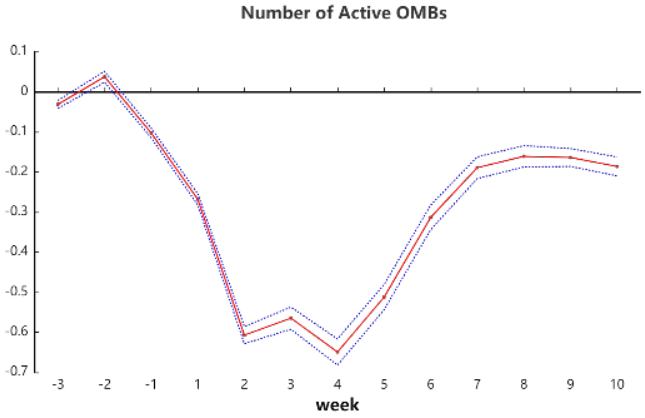
- 在某些情况下，处理组和对照组在处理政策前**可能并非同样的线性趋势。**
  - 在估计Covid-19疫情对中国 线下微型商户冲击中，我们（Guo *et al.*，2020）利用双重差分法框架，但将线性回归改成了机器学习中的梯度提升树方法，利用2020年疫情前数据、2019和2018年“疫情”前后数据，成功预测了如果没有疫情发生，2020年应该具有的反事实结果，从而获得了疫情对线下微型商户冲击的科学估计。

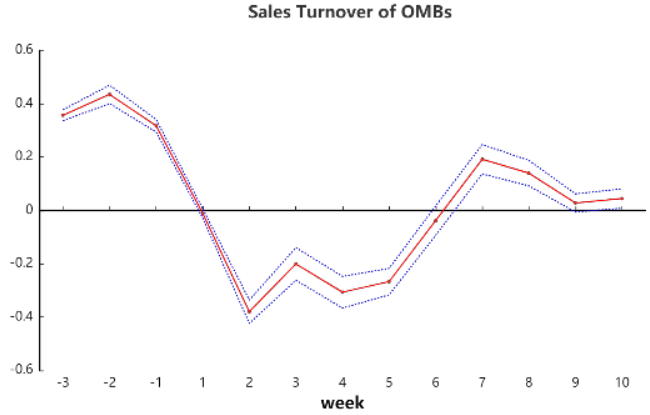
$$OMB_{i,2019+k} = F(OMB_{i,2018+k}, OMB_{i,2018+(k-1)}, OMB_{i,2018-h}, OMB_{i,2019-h}, X_{i,2019+k}, Z_i) \quad (2)$$

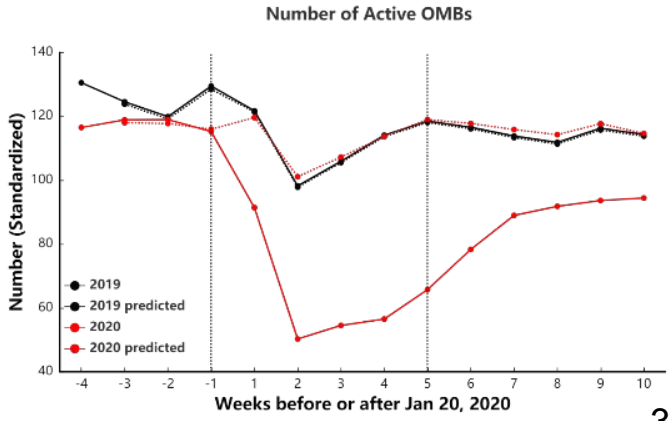
$$\widehat{OMB_{i,2020+k}} = F(OMB_{i,2019+k}, OMB_{i,2019+(k-1)}, OMB_{i,2019-h}, OMB_{i,2020-h}, X_{i,2020+k}, Z_i) \quad (1)$$

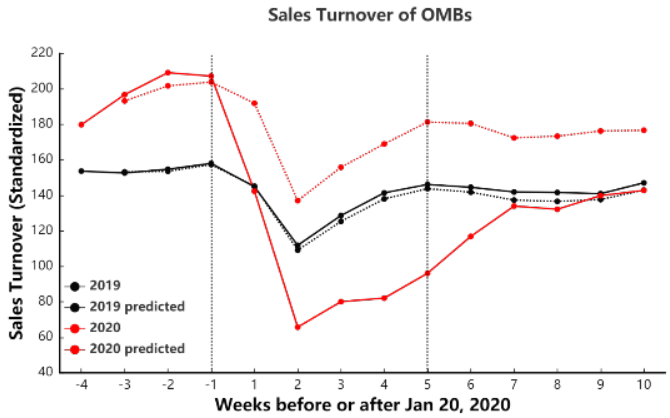
三、机器学习对因果识别的意义

- 上一行是传统线性DID回归，不满足平行线趋势
- 下一行为DID框架下的机器学习方法，2019年和2020年疫情前的预测值与实际值，高度重叠









Feng Guo, Yiping Huang, Jingyi Wang, Xue Wang, The Informal Economy at Times of COVID-19 Pandemic, Working Paper, 2020.

三、机器学习对因果识别的意义

- 更好地构建对照组：匹配法

变量		吸烟组(n=233)	不吸烟组(n=949)	P 值
年龄(岁)	<45	151(64.8)	553(58.3)	0.131
	45-	44(18.9)	249(26.2)	
	55-	26(11.2)	105(11.1)	
	65-	12(5.2)	42(4.4)	
性别	男	183(78.5)	135(14.2)	<0.001
	女	50(21.5)	814(85.8)	
高血压家族史	是	123(52.8)	402(42.4)	0.004
	否	110(47.2)	547(57.6)	
饮酒	是	171(73.4)	160(16.9)	<0.001
	否	62(26.6)	789(83.1)	
BMI(kg/m^2)	<24	86(36.9)	438(46.2)	0.036
	24-27.9	96(36.9)	308(32.5)	
	28-	61(26.2)	203(21.4)	
血脂异常	是	80(34.3)	196(20.7)	<0.001
	否	153(65.7)	753(79.3)	

三、机器学习对因果识别的意义

- **更好地构建对照组：匹配法**

- 匹配的过程是一个典型的分类问题：Y为是否处理组（处理组为1，对照组为0），X为协变量等

- 传统方法：广义线性方法Probit，logit等

- 机器学习优势：处理非线性关系；协变量很多；多种方法比较；

三、机器学习对因果识别的意义

- **高维数据匹配：**在传统的倾向得分匹配当中，协变量一般不能太多，比如最好是5-50个。但是，即便我们在极少的协变量情况下满足了协变量平衡和共同支撑假设，我们也无法置信十几个特征，甚至几个特征相似就让我们认为处理组和对照组可比。

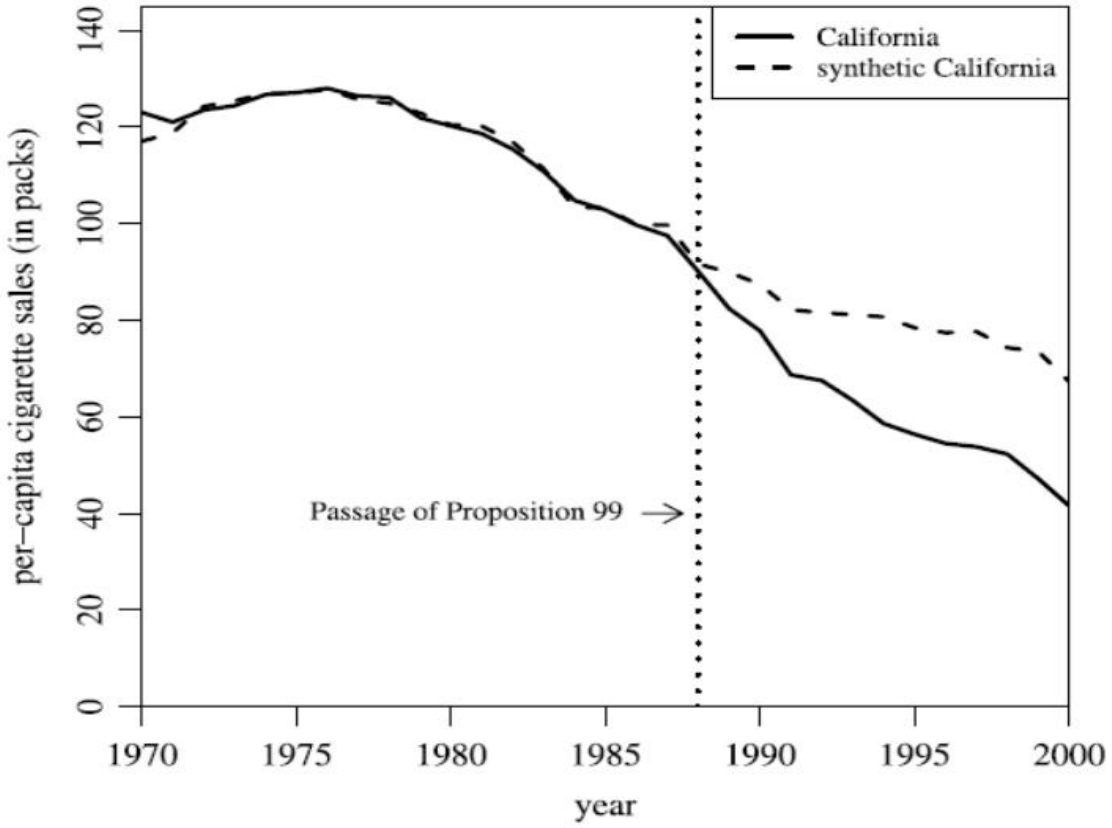
- 也就是说，在能够满足共同支撑假设和协变量平衡假设的前提下，协变量的数量应该是越多越好。Knaus et al.（2020）就利用Post-Lasso的方法在1268个协变量中进行特征筛选然后匹配，最终获得是否参与工作培训项目对失业人员就业时间的因果效应。

- **文本数据匹配：** Robertsy et al.（2020）在研究中国社交媒体中，有被审查经历是否会增加其再次被审查的概率。为此，他们设计了一个机器学习方法，来解决文本数据的匹配问题，进而应用在上述问题的因果识别当中。具体而言，他们使用一个合适的主题模型来表征文本，然后再用倾向得分来匹配文本主题。

三、机器学习对因果识别的意义

- **更好地构建对照组：合成控制法**
  - 在处理组非常独特，难以寻找到或匹配到合适的对照组的情况下，也可以考虑利用众多对照组“合成”一个合适的对照组。

  - 逻辑：政策前，**处理组与其他众多对照组之间的拟合关系**，能够在处理政策后（如果没有处理政策发生）仍然保持不变。



三、机器学习对因果识别的意义

- 如果在处理政策之前，处理组与对照组之间拟合得较好，但这种拟合关系到了处理政策后，可能就不存在了，那么用机器学习的术语，这就是出现了过拟合的问题。

$$\hat{w}_i = \arg\min_{w_i} \sum_{j \in pre} (Y_T(j) - \sum_{j \in pre} w_i Y_{C,i}(j))^2, \quad s.t. \sum_{i=1}^{N} w_i = 1; w_i \geq 0$$

- 机器学习：引入惩罚项，提高泛化能力；允许权重为负；高度非线性。

- Guo and Zhang（2019）在研究襄樊市更名为襄阳市对经济增长的影响时，使用了机器学习算法（Lasso和Elastic net）进行控制个体的筛选，以便为襄樊市合成出一个更好的对照组。

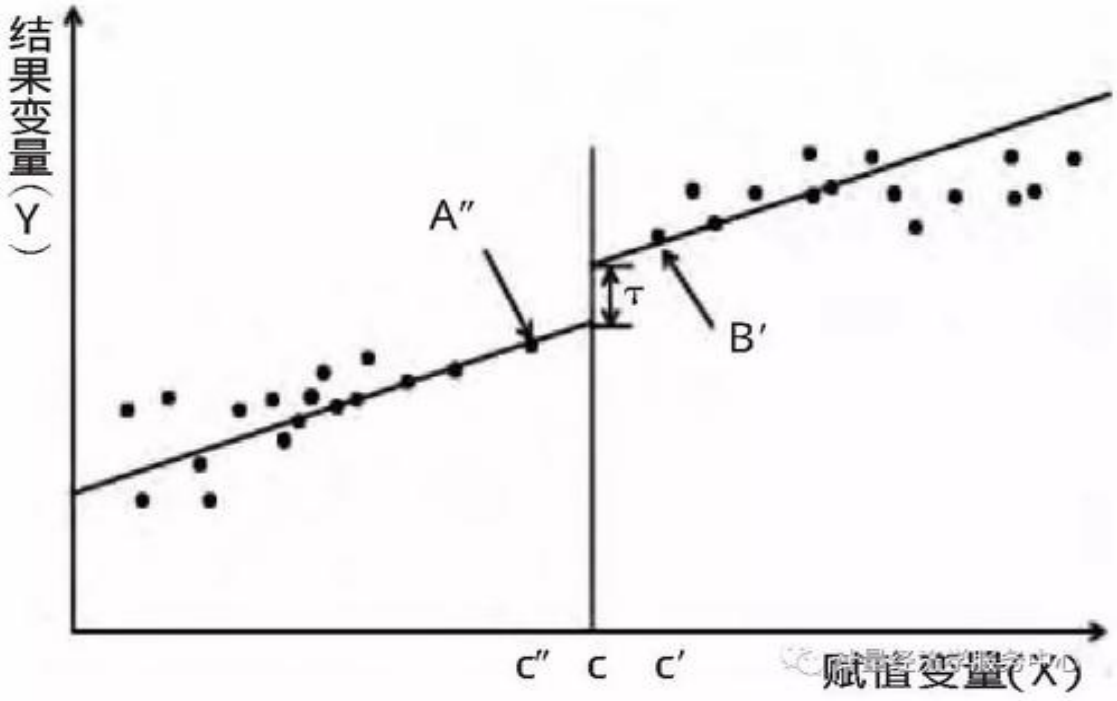
- Mühlbach（2020）提出了一种基于树（随机森林）的合成控制方法，并证明在考察美国驻以色列大使馆迁址的后果时，该方法比其他方法可以取得更好的效果。

- 王立勇等（2021）利用这一方法，考察了中国加入WTO对财政政策波动性的因果效应。

三、机器学习对因果识别的意义

- **更好地构建对照组：断点回归**

  - 利用断点左侧的数据，对结果变量与配置变量（以及其他变量）之间的关系进行建模，进而将建模参数泛化到断点右侧，从而估计断点右侧如果没有处理政策的法应该具有的“反事实结果”，而这正是机器学习算法的优势所在（Imbens and Wager，2019；王芳等，2020）



三、机器学习对因果识别的意义

- **更好地构建对照组：断点回归**

  - 在传统的断点回归设计中，配置变量一般都是一维的，研究者可以通过对研究问题背景了解，明确断点的位置。**但如果考虑多维的配置变量**，则断点的具体位置就变得不直观了，甚至无法通过人工观察而确定。

  - 此时可以使用机器学习的方法来自动判别断点的具体位置（Herlands et al.，2018）。

三、机器学习对因果识别的意义

- **更好地识别异质性因果效应**
  - 相比传统社会科学实证分析关注某一政策的平均因果效应，异质性因果效应的评估能够回答如下问题
  - 怎样的一种催票组合方式能够带来最有效的选票（Imai et al.，2013）
  - 哪些人最应该得到失业人员工作求职培训（Knaus et al.，2020）
  - 上大学更能提高哪些人的工资收入（Xie et al.，2012）
  - 以及地铁开通使周边哪一类型的住房增值最大（Seungwoo et al.，2018）。

三、机器学习对因果识别的意义

- **传统异质性因果识别方法：交互项或分组回归**

- 可能存在问题

  - 其一，存在异质性的变量可能很多，但是在给定数据的情况下，**我们不可能无限制地在模型中添加大量交互项**，尤其是存在高维数据的时候。分组回归中也是如此，我们也无法获知哪一个变量是具有异质性特征的。即便我们知道，对于一个连续变量如何切分也是一个难题。因而，传统方法中，交互项和分组的设置有一定的主观性。

  - 其二，**不能处理多重异质性的问题**，也即是交互或分组变量具体形式很可能是二次、三次甚至更多变量的交互等非线性形式，而这种设定往往是研究者主观设置的，未必符合数据生成过程的特征。

- 这两个问题实际上可以进一步转化并归纳为四个问题：**高维数据变量筛选、变量切分、复杂和非线性的数据建模、模型设定错误。**

三、机器学习对因果识别的意义

- 如何在成百上千的协变量中筛选变量并且保证计算的可行，便成为机器学习正则化算法的优势所在。

- Knaus et al.（2020）用**Post-Lasso**算法在1268个协变量中进行筛选，对几乎所有可能的样本分组进行异质性分析，系统地考察了不同失业人员在接受求职培训后的因果效应差异，并依据异质性因果效应的结果，提供了针对不同类失业人员参加求职培训的分配规则，进而达到改善求职培训效果的目的。

- Imai et al.（2013）也使用**Lasso来**筛选变量，并结合支持向量机算法分析了包括登门拜访、电话留言、发送0-3封邮件、靠公民义务、邻里呼吁等193种处理组合对美国选举拉票效果的影响，发现亲自登门拜访是获得选票最有效的拉票方式，发三封邮件并附上公民义务的消息是除上门访谈最为有效的方式，而通过邻里呼吁或电话留言则会降低公众投票率。

三、机器学习对因果识别的意义

- Seungwoo et al.（2018）借助机器学习中的回归**树算法**，在双重差分法的框架下，讨论了首尔某条地铁开通对周边房地产市场的**条件平均因果效应**：在地铁周边住房的142个特征变量中，有89个特征会带来住房价格增值，53个特征会使得房价减值。

Appendix C. Conditional Average Treatment Effects

	CATE	S.E	N	Size (Category)	Room (No.)	Bath (No.)	Old (Category)	Other Transits Within 1km	District (Name)
1	1.0573	0.0509	201	Top 25%	3	2	< 5 yrs	Yes	Seocho
2	0.9023	0.0993	360	Bottom 25%	1	1	> 10yrs	Yes	Yeongdeungpo
3	0.8203	0.0968	96	Top 25%	5	2	5 ~ 10yrs	No	Yangcheon
4	0.6403	0.0780	208	Bottom 25% - 50%	3	1	5 ~ 10yrs	Yes	Seocho
5	0.5959	0.0442	1179	Bottom 25% - 50%	3	2	< 5 yrs	Yes	Seocho
6	0.5524	0.0630	536	Top 25% - 50%	4	2	< 5 yrs	Yes	Seocho
7	0.5453	0.0551	787	Top 25% - 50%	4	2	5 ~ 10yrs	Yes	Gangnam
8	0.5011	0.0726	1530	Bottom 25% - 50%	3	2	5 ~ 10yrs	No	Gangseo
9	0.4972	0.0941	628	Top 25%	4	2	5 ~ 10yrs	No	Gangseo
10	0.4637	0.0470	532	Bottom 25%	2	1	5 ~ 10yrs	Yes	Gangseo
11	0.4396	0.0911	1586	Top 25% - 50%	3	2	5 ~ 10yrs	No	Gangseo
12	0.4328	0.1324	228	Top 25%	3	2	5 ~ 10yrs	Yes	Seocho
13	0.4118	0.0283	757	Bottom 25%	3	1	5 ~ 10yrs	Yes	Gangseo

三、机器学习对因果识别的意义

- **传统异质性因果识别方法：以倾向值为导向**

  - Xie et al.（2012）和Zhou and Xie（2019）提出了三种以倾向值为导向的异质性处理效应分析方法。分别是：分层法，匹配—平滑法，平滑—差分法。其核心思想都是看处理效应如何随着倾向值变化而变化。

  - 可能的问题

    - **第一**，因为采用哪些变量估计倾向值仍然不确定，这使得倾向值模型可能设定错误（胡安宁，2017）。
    - **第二**，将变量降维为倾向值，但是让我们损失了很多信息，即我们仍然不知道究竟哪个变量有异质性特征。
    - **第三**，这种方法更无法让我们获得每一个人的个体处理效应。

三、机器学习对因果识别的意义

- **传统异质性因果识别方法：以倾向值为导向**

- 对于上述三个问题，机器学习方法却有用武之地。**一些集成算法**在最优模型设定、高维数据处理以及个体处理效应估计方面已经相当成熟。比如，因果森林在个体处理效应估计方面表现尤为出色。

  - 在实际应用中，Knittel et al.（2019）将因果森林算法应用于一个大规模行为干预实验的评估，讨论发送家庭能源报告是否有助于推动家庭节能，在其中，他们依据因果森林获得的异质性因果效应的结果，再次针对性地发送家庭能源报告，发现社会效益会额外提升12-120%。

三、机器学习对因果识别的意义

- **更好地检验因果关系的外部有效性**

  - 在传统因果推断的社会科学实证研究中，一般都缺乏结论是否能够外推的考察，很少去强调模型的验证问题，似乎默认存在一个根据理论推导而得来的一个“正确”的实证模型。

  - 实证结论可能缺乏外部有效性这一问题之所以会经常出现是因为其跟目前社会科学领域的研究和学术发表惯例相关：P值敲打。

  - 解决办法：训练集和测试集

四、大数据时代因果识别的挑战

- **因果关系在某些情形下变得不再重要**

  - 首先，在很多政策问题当中，并不是非要知道因果关系，有时候只需要预测到一个结果就行。

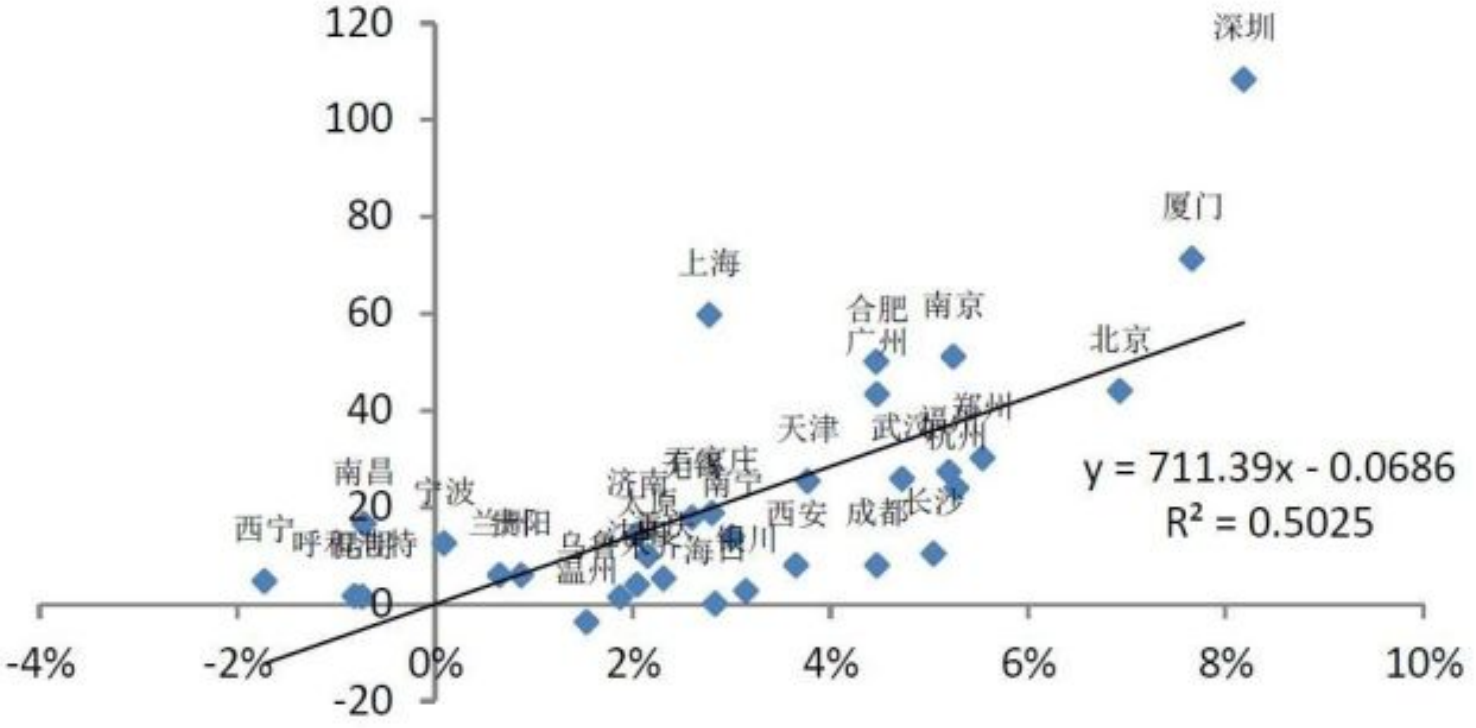
  - 其次，当要研究某些重要的经济社会问题，而又缺乏直接的数据时，有时候放弃对因果效应的执着，改为只关注经济社会变量的相关关系，会为我们开辟一个新的“脑洞”。

  - 例如使用手机使用习惯预测信贷违约，不能认为手机使用习惯“导致”了该个体的信用状况，但又不能否认这种大数据征信的巧妙之处。

四、大数据时代因果识别的挑战

- 小学生在校人数与房价涨幅正相关

图3：小学在校生人数增速和房价涨幅，%



数据来源：国家统计局统计年鉴，安信证券

四、大数据时代因果识别的挑战

- **大数据和机器学习让某些情形下因果关系识别更困难**

  - 机器学习生成的变量（政治态度、投资者情绪等），准确率不高，产生测量误差，影响因果识别

  - 机器学习实操中，可能会穷尽训练集数据中的各种细节和噪音，从而在训练集得到的处理变量和结果变量之间的关系，无法泛化到一般化的情形当中，从而产生虚假的因果关系。

四、大数据时代因果识别的挑战

- **部分机器学习算法缺乏可解释性**
  - 很多机器学习算法还是一个黑箱，讲不清楚x如何影响y，还怎么讨论因果关系。
  - 很多机器学习算法甚至无法对标准误和置信区间等进行估计（Athey，2017）

五、结论与展望

- **机器学习有没有用**
  - 在很多方面拓展了传统因果关系识别的适用条件
  - 而且，机器学习方法应用在因果识别中，还可以帮助社会科学家发现新问题。

- **不能过度追求机器学习**
  - 过分地追求机器学习方法，也可能会与社会科学目标相悖。
  - 我们首先是一个社会科学家，其次才是一个数据分析人员，我们只是在利用大数据和机器学习的工具，来帮助我们更好的理解这个社会。

重点推荐文献

- **Lasso与工具变量**

- Gilchrist, D. S., and E. G. Sands 2016, “Something to Talk About: Social Spillovers in Movie Consumption”, Journal of Political Economy 124(5).

- Qiu, Y., X. Chen, and W. Shi 2020, “Impacts of Social and Economic Factors on the Transmission of Coronavirus Disease 2019 (COVID-19) in China”, Journal of Population Economics 33.

- 方娴、金刚，2020，《社会学习与消费升级：来自中国电影市场的经验证据》，《中国工业经济》,第1期。

重点推荐文献

- 机器学习与DID

- Guo, F., Y., Huang, J. Wang, X. Wang 2021a, “The Informal Economy at Times of COVID-19 Pandemic”, Working Paper.

- Cicala, S. 2017, “Imperfect Markets versus Imperfect Regulation in U.S. Electricity Generation”, National Bureau of Economic Research Working Paper, No.23053.

重点推荐文献

- 机器学习与匹配法

- Knaus, C. M., M. Lechner, and A. Strittmatter 2020, “Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach”, Journal of Human Resources, 2020.

- Robertsy, M., B. Stewartz, and R. Nielsen 2020, “Adjusting for Confounding with Text Matching”, American Journal of Political Science64(4).

重点推荐文献

- 机器学习与合成控制法

- Guo, J., and Z. Zhang 2019,“Does Renaming Promote Economic Development? New Evidence from a City-renaming Reform Experiment in China”, China Economic Review 57.

- Mühlbach, N. N. 2020, “Tree-based Synthetic Control Methods: Consequences of moving the US Embassy”, Institut for Økonomi, Aarhus Universitet, CREATES Research Papers, No. 2020-04.

重点推荐文献

- **机器学习与异质性因果**

- 胡安宁，吴晓刚，陈云松，2021，《处理效应异质性分析——机器学习方法带来的机遇与挑战》，《社会学研究》第1期。

- Seungwoo, C., M. E. Kahn, and M. H. Roger 2018, “Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach”, Real Estate Economics 48(3).

1917—2017
100th Anniversary
Shanghai University of Finance and Economics
上海财经大学 100周年校庆

Thank You!

公众号：经济数据勘探小分队