doi:10. 16055/j. issn. 1672-058X. 2022. 0006. 009

基于因子分析的因果推断研究

付举望, 孔新兵
(南京审计大学 统计与数据科学学院,南京 211815)

摘　要:针对反事实框架下的因果推断问题,在因子分析视角下,从优化角度提出利用 L_2 因子分析方法估计反事实值,并引入 L_1 损失函数优化 L_1 风险;结合因果推断与正交因子模型,将面板数据中需要估计的反事实值视作缺失值,从而把因果推断反事实值估计转变为带有缺失值的潜在因子模型估计;舍弃面板数据中的缺失值,通过优化一步得到潜在结果与平均处理效应,避免了信息丢失问题;采用 L_1 因子分析代替 L_2 因子分析来估计模型,做出稳健性上的改进,并获得中位数处理效应;介绍了一种交替凸优化算法解决 L_1、L_2 因子分析中的目标函数最小化问题,并给出其具体实现步骤;对于加利福尼亚州限制烟草政策案例做了实证研究,将 L_1、L_2 因子分析与已有因果推断方法进行比较分析,结果表明:因子模型的 L_1、L_2 估计量同样适用于宏观经济变量预测;最后通过设置伪实验组与伪介入的假设,验证了 L_1 因子分析较其他方法具有更稳健的预测效果。

关键词:反事实估计;因果推断;L_1 因子分析
中图分类号:O212. 1　　**文献标志码**:A　　**文章编号**:1672-058X(2022)06-0071-08

0　引　言

统计学中,相关关系能够通过相关系数进行度量,而因果关系却很难有一个明确的度量指标,由此衍生出的因果推断(Causal Inference)问题成为统计学者们关注的焦点。Rubin [1]在反事实理论基础上构建了潜在结果模型(Rubin Causal Model),其核心是比较同一个研究对象在接受干预和不接受干预时的结果差异,即该干预的因果效应(Causal Effects)。在该反事实框架下,因果推断问题转变为反事实值的估计问题。目前,关于受数据限制较小,能够服务于高维数据,且估计效果更好的因果推断方法,还有待进一步研究。

Rubin 关注单个协变量情形下平均处理效应的估计问题,通过处理组与控制组的分离与再匹配估计反事实值; Heckman 等[2-3]使用双重差分法(Difference-in-Difference)估计反事实值,并将其应用于社会公共政策评估;Abadie 等[4]提出合成控制法(Synthetic Control Method),将控制组个体加权,合成一个与处理组相似的虚拟组,通过比较干预前后的处理组和虚拟组的变量变化,得出平均处理效应, 改进了双重差分法的内生性问题; Zheng 等[5]又在此基础上,使用机器学习二次规划法来确定控制组的权重并重构虚拟组,以预测反事实值; Doudchenko 等[6]比较了双重差分法、合成控制法、约束回归法、最优子集选择法对于估计参数的不同限制条件,并使用弹性网法(Elastic Net)设置惩罚项,以此筛选控制变量构造模型,获得反事实值预测。这些方法都是基于观测到的面板数据特点进行模型假设,导致其受限于所得观测数据,在面对具有不同特征的数据时稳健性较弱。

收稿日期:2021-03-05;**修回日期**:2021-05-18.
基金项目:国家自然科学基金项目资助(71971118).
作者简介:付举望(1997—),男,四川成都人,硕士,从事统计因果推断研究.

因子分析作为常见的宏观经济变量预测方法,面对高维数据表现优异且能应用于不同特征的数据。因此在反事实框架下,学者们通过非随机观测数据,在因子分析的视角下进行因果推断,提出了新的反事实值估计方法。Xiong 等[7]提出带有缺失值情况下的潜在因子模型,其通过行列调整后的协方差矩阵估计获得公共因子与因子载荷,由此得到反事实值估计;Bai 等[8]提出 TW 算法(Tall-Wide algorithm),其假设面板数据具有强因子结构,并将数据划分为 bal、tall、wide、miss 4 块(block),分别利用 tall-block 估计公共因子、wide-block 估计旋转后的因子载荷,从而得到平均处理效应的估计。这些方法都是在因子分析前对面板数据进行调整,将缺失值所在行、列或是周围一整块数据丢弃,导致已有观测数据信息无法完全利用。

本文从优化的角度提出 L_2 因子分析方法,获得平均处理效应估计,且避免了信息丢失问题。与 Xiong 等不同,该优化方法的显著优势在于勿需调整面板数据的行和列,也勿需为了构造满足奇异值分解的子矩阵而整行或整列地丢弃数据;而与 Bai 等的区别在于:该优化方法将除缺失值以外的所有数据作为整体进行因子分析,而非将整个面板数据划分后根据特定数据块分别进行潜在因子与因子载荷估计,勿需舍弃未使用数据块的信息。该方法仅需舍弃面板中的缺失值,即可通过优化一步得到潜在数据生成效应,即潜在结果,从而提高估计效率。另外,与上述文献关心平均处理效应不同,本文还通过引入 L_1 损失函数并优化 L_1 风险,得到中位数处理效应的估计。

1 处理效应模型

本文以政策评估为例,简述反事实框架与随机对照实验。假设某个城市(i)在某个时刻(t)被政策介入,在该地区对感兴趣的某项指标 $y_{I,t}$ 进行研究。考虑拥有关于该城市以及其他未被政策介入的城市的面板数据 $\boldsymbol{Y}\in\boldsymbol{R}^{N\times T}$,其中在政策实施前的数据一共有 a 年,实施后的数据一共有 b 年,共计 T 年($T=a+b$)。

借鉴已有因果模型变量设置方法,记城市 $i=1,2,\cdots,n_1$ 为被政策介入的城市,作为实验组(treat unit);城市 $j=1,2,\cdots,n_2$ 为未被政策介入的城市,作为控制组(control unit),共计 N 个城市($N=n_1+n_2$)。时间 $t=T_1,T_2,\cdots,T_a,T_{a+1},\cdots,T_{a+b}$,其中 $t=T_1,T_2,\cdots,T_a$ 为介入前,$t=T_{a+1},T_{a+2},\cdots,T_{a+b}$ 为介入后,将面板数据 $\boldsymbol{Y}_{N\times T}$ 划分为 4 个部分,具体变量设置如表 1 所示。

表 1 变量设置

Table 1 Setting variables

	介入前	介入后
实验组	$y_{i,T_1}^{\text{obs}},\cdots,y_{i,T_a}^{\text{obs}}$	$y_{i,T_{a+1}}^{\text{obs}},\cdots,y_{i,T_{a+b}}^{\text{obs}}$
控制组	$y_{j,T_1}^{\text{obs}},\cdots,y_{j,T_a}^{\text{obs}}$	$y_{j,T_{a+1}}^{\text{obs}},\cdots,y_{j,T_{a+b}}^{\text{obs}}$

在 t 时刻($t=T_{a+1},T_{a+2},\cdots,T_{a+b}$),对于任意一个实验组城市 i,所关心的政策效应不能简单地由 $E(y_{j,t}^{\text{obs}}-y_{i,t}^{\text{obs}})$ 表示,原因在于该指标不仅受政策的影响,而且城市自身的影响也十分显著。一个自然的想法是,如果能够获得该指标在不受该政策影响下(即没有政策介入情况下)的反事实值 $\hat{y}_{i,t}^{\text{predict}}$,即可消除城市自身的影响效应。该政策效应为总体期望:

$$\hat{\tau}=E(y_{i,t}^{\text{obs}}-\hat{y}_{i,t}^{\text{predict}}),t=T_{a+1},T_{a+2},\cdots,T_{a+b} \tag{1}$$

由式(1)可以看出:只需要多次进行实验估计反事实值,将其与真实观测值的差求期望即可得到政策的因果效应。Rubin(1973)提出了一般化的估计方法框架:

$$\hat{y}_{i,t}^{\text{predict}}=\boldsymbol{\mu}+\sum_{j=1}^{n_2}\omega_j y_{j,t}^{\text{obs}},t=T_{a+1},\cdots,T_{a+b} \tag{2}$$

即实验组的值可以表示为控制组值的线性组合。由此一来,就可以通过对处理前实验组和控制组进行回归,得到参数 $\boldsymbol{\omega}=(\omega_1,\omega_2,\cdots,\omega_{n_2})$ 和 μ,再使用估计出的参数和处理后控制组的值估计处理后实验组的反事实预测值。这样一来,问题就由求反事实估计值变为求回归估计参数 $\theta=(\mu,\boldsymbol{\omega})$。一个简单的想法是利用简单最小二乘法最小化以下目标函数 Q_0,来得到参数 $\boldsymbol{\omega}$ 和 μ 的估计:

$$Q_0(\mu,\boldsymbol{\omega}\mid y_{i,t}^{\text{obs}})=\sum_{j=1}^{n_2}\sum_{t=T_1}^{T_a}\|y_{i,t}^{\text{obs}}-\mu-\boldsymbol{\omega}^{\mathrm{T}}y_{j,t}^{\text{obs}}\|_2^2$$

但当 $\theta=(\mu,\boldsymbol{\omega})$ 的维数较高时,最小二乘法将导致较大的预测误差,使得政策效应的估计变得非常不稳定。因此,在这基础上还需要一些假设条件对参数 $\theta=(\mu,\boldsymbol{\omega})$ 加以限制。而目前的各种方法,均是基于已有面板数据设立模型假设,在该假设下对参数 $\theta=(\mu,\boldsymbol{\omega})$ 施加条件,以获得更好的估计效果。

2 基于因子分析的反事实值估计

2.1 因子模型及其估计

近年来,因子模型的理论和应用已经得到了很

大的完善和发展,动态因子模型常被应用于宏观经济政策的评估、经济指数的构建和经济指标的预测。统计学中,学者们关注因子个数的确定研究以及公共因子与因子载荷的估计方法研究。Bai[9]提出确定静态因子个数的信息准则,该准则在保证因子模型的拟合优度前提下,通过面板数据结构特征得到因子个数的无偏估计;Fan 等[10]讨论具有条件稀疏结构的高维协方差估计问题,通过引入 POET(Principal Orthogonal compl Ement Thresholding)方法来探索这种近似因子结构;Kong[11]提出一种局部主成分分析方法,用高频数据确定具有时变因子载荷的连续时间因子模型的公共因子数,该模型采用离散时间因子模型在收缩块上进行局部近似。

假设面板数据 $\boldsymbol{Y}\in\boldsymbol{R}^{N\times T}$ 具有因子结构,记 $\boldsymbol{f}_t\in\boldsymbol{R}^r$ 为潜在因子;$\boldsymbol{\lambda}_s\in\boldsymbol{R}^r$ 为因子载荷;$\boldsymbol{C}_{st}=\boldsymbol{\lambda}_s^{\mathrm{T}}\boldsymbol{f}_t$ 为公共因子部分;$\boldsymbol{e}_{\mathrm{st}}$ 为特殊因子部分,即有

$$\begin{aligned}&\boldsymbol{Y}_{st}=\boldsymbol{\lambda}_s^{\mathrm{T}}\boldsymbol{f}_t+\boldsymbol{e}_{st}\\&s=1,\cdots,n_1,n_1+1,\cdots,N\\&t=1,\cdots,a,a+1,\cdots,T\end{aligned}\tag{3}$$

在反事实估计中,由于面板数据 $\boldsymbol{Y}_{st}$ 满足式(3)的线性关系,因此可以通过估计潜在因子 $\widetilde{\boldsymbol{f}_t}$ 以及相关因子载荷 $\widetilde{\boldsymbol{\lambda}_s}$,得到 $\boldsymbol{Y}_{st}$ 的估计 $\widehat{\boldsymbol{Y}}_{st}=\widetilde{\boldsymbol{\lambda}_s}^{\mathrm{T}}\widetilde{\boldsymbol{f}_t}$。主成分方法(PCA)就是一种十分流行的估计方法。

通常,一个数据集总是由若干随机变量的若干观测组成。因子分析的目标就是将原始数据集进行降维,将这些观测投射到一个低维因子空间中。这样的投射有无数种,主成分方法希望找到这样一种投射,可以使得数据在低维空间的投影拥有最大的方差。而 L_2因子分析问题则可以表述为目标函数 Q_1 的最小化问题:

$$Q_1(\boldsymbol{\Lambda},\boldsymbol{F})=\|\boldsymbol{Y}-\boldsymbol{\Lambda F}\|_{L_2}=\sum_{s=1}^{N}\sum_{t=1}^{T}(\boldsymbol{y}_{st}-\Lambda F)^2$$

其中:$\Lambda=(\lambda_1,\cdots,\lambda_{n_1},\lambda_{n_1+1},\cdots,\lambda_N)^{\mathrm{T}}$,$\boldsymbol{F}=(f_1,\cdots,f_a,f_{a+1},\cdots,f_T)$。对于目标函数 Q_1 的最小化问题,常用方法为对协方差矩阵 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 进行奇异值分解(SVD),寻找最大的 r 个特征值,再用其对应的 r 个特征向量构成的矩阵做低维投影降维。

2.2　L_2 范数因子分析

通过估计潜在因子 $\widetilde{\boldsymbol{f}_t}$ 以及相关因子载荷 $\widetilde{\boldsymbol{\lambda}_s}$,即可得到反事实估计值 $\widehat{\boldsymbol{Y}}^{\mathrm{predict}}=\widetilde{\boldsymbol{\lambda}_s}^{\mathrm{T}}\widetilde{\boldsymbol{f}_t}$。这样一来,问题就由求反事实估计值变为求潜在因子 $\widetilde{\boldsymbol{f}_t}$ 与相关因子载荷 $\widetilde{\boldsymbol{\lambda}_s}$。为了进行反事实估计,将表 1 中所有受到政策干预的变量($y_{i,T_{a+1}}^{\mathrm{obs}},\cdots,y_{i,T_{a+b}}^{\mathrm{obs}}$)视作缺失值,再进行因子分析。此时用于因子分析的面板数据 $\boldsymbol{Y}'$ 是由余下的控制组城市数据以及实验组城市未被介入前的数据构成,最终估计出的潜在因子 $\widetilde{\boldsymbol{f}_t}$ 以及相关因子载荷 $\widetilde{\boldsymbol{\lambda}_s}$ 也能通过余下的变量表示。其思想与式(2)十分相似,都是通过未被介入的变量估计反事实值。

在对带有大量缺失值的面板数据进行因子分析时,基于最小二乘法的经典方法会遇到稳健性问题,难以得到满意的公共因子与因子载荷。对于缺失值处理,Xiong 等提出带有缺失值情况下的潜在因子模型,在协方差矩阵估计时,对所有个体重加权,删除缺失值所在的行和列,从而将面板数据调整为满足奇异值分解的子矩阵,再根据所得矩阵的特征向量进行潜在因子与因子载荷的估计; Bai 等将面板数据划分为 4 块(block),分别利用 tall-block 估计公共因子、wide-block 估计旋转后的因子载荷;而本文则从优化的角度直接舍弃目标函数中对应缺失的累加项,此时的目标函数 Q_2 改写为

$$Q_2(\boldsymbol{\Lambda},\boldsymbol{F})=\sum_{s=1}^{n_1}\sum_{t=1}^{a}(\boldsymbol{y}_{st}-\boldsymbol{\lambda}_s^{\mathrm{T}}\boldsymbol{f}_t)^2+\sum_{s=n_1+1}^{N}\sum_{t=1}^{T}(\boldsymbol{y}_{st}-\boldsymbol{\lambda}_s^{\mathrm{T}}\boldsymbol{f}_t)^2$$

s. t. $\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}/\boldsymbol{N}=\boldsymbol{I}_N$,　$\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}/T$ 为对角阵。

使用该方法的优点是勿需对面板数据的行列进行拼凑和删减操作,且使用了所有未缺失数据的信息。

2.3　L_1 范数因子分析

L_2 因子分析的缺陷是对离群值十分敏感,为了解决该问题,本文使用更具有稳健性的 L_1 因子分析进行研究。

将目标函数 Q_2 更换为使用 L_1 范数,即可得目标函数 Q_3:

$$Q_3(\boldsymbol{\Lambda},\boldsymbol{F})=\sum_{s=1}^{n_1}\sum_{t=1}^{a}|\boldsymbol{y}_{st}-\boldsymbol{\lambda}_s^{\mathrm{T}}\boldsymbol{f}_t|+\sum_{s=n_1+1}^{N}\sum_{t=1}^{T}|\boldsymbol{y}_{st}-\boldsymbol{\lambda}_s^{\mathrm{T}}\boldsymbol{f}_t|$$

s. t. $\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}/N=\boldsymbol{I}_N$, $\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}/T$ 为对角阵。

相较于 L_2 范数,L_1 范数在统计学中主要作为稳健手段。在优化问题中,L_2 范数对于离群值具有放大作用,而 L_1 范数则更具稳健性。另外一个区别是 L_1 范数意味着 $\boldsymbol{y}_{st}$ 的中位数被建模为 $\boldsymbol{\lambda}_s^{\mathrm{T}}\boldsymbol{f}_t$,即 $m(\boldsymbol{y}_{st}|\boldsymbol{\lambda}_s,\boldsymbol{f}_t)=\boldsymbol{\lambda}_s^{\mathrm{T}}\boldsymbol{f}_t$。而 L_2 范数意味着 $\boldsymbol{y}_{st}$ 的均值被建模为 $\lambda_s^{\mathrm{T}}\boldsymbol{f}_t$,即 $M(\boldsymbol{y}_{st}|\boldsymbol{\lambda}_s,\boldsymbol{f}_t)=\boldsymbol{\lambda}_s^{\mathrm{T}}\boldsymbol{f}_t$。与均值建模的近似因子模型(Approximate Factor Model)不同,分位数建模的分位数因子模型(Quantile Factor Model)估计出的共同因子对可观测变量整体分布的影响更为全面;现有文献多为利用主成分分析方法估计近似因子模型,而基于分位数因子模型的估计方法不仅可以提取出更多的潜在因子,而且对极端值、缺失值和厚尾分布具有较强的稳健性[12]。

2.4 交替凸优化算法

目标函数 Q_1、Q_2 与 Q_3 的最小化问题实质为无约束最优化问题。其中 Q_1 为均值插补后的因子分析,可使用奇异值分解进行直接计算;Q_2 与 Q_3 则需要使用优化算法解决。

一种经典的方法是使用交替凸优化算法求解该问题。该算法的思想是:当面临一个两维变量的优化问题,而该问题不是凸优化问题因此无法求其最优解时,可以采用交替迭代的方法,每一步将其中一维未知变量的值看作是常数(使用该变量上次迭代的取值),来求解另一维未知量。由目标函数 Q_2 与 Q_3 给出式(4)、式(5):

$$\boldsymbol{F}^{(m)}=\arg\min_{F} Q_k(\boldsymbol{\Lambda}^{(m-1)},\boldsymbol{F}),k=2,3 \tag{4}$$

$$\boldsymbol{\Lambda}^{(m)}=\arg\min_{\boldsymbol{\Lambda}} Q_k(\boldsymbol{\Lambda},\boldsymbol{F}^{(m)}),k=2,3 \tag{5}$$

改写目标函数 Q_2 与 Q_3,有

$$Q_2(\boldsymbol{\Lambda},\boldsymbol{F})=\|\boldsymbol{Y}-\boldsymbol{\Lambda}^{(m-1)}\boldsymbol{F}\|_{L_2}=\sum_{t=1}^{a}(\boldsymbol{y}_t-\boldsymbol{\Lambda}^{(m-1)}\boldsymbol{f}_t)^2+\sum_{t=1}^{T}(\boldsymbol{y}_t-\boldsymbol{\Lambda}^{(m-1)}\boldsymbol{f}_t)^2$$

$$Q_3(\boldsymbol{\Lambda},\boldsymbol{F})=\|\boldsymbol{Y}-\boldsymbol{\Lambda}^{(m-1)}\boldsymbol{F}\|_{L_1}=\sum_{t=1}^{a}|\boldsymbol{y}_t-\boldsymbol{\Lambda}^{(m-1)}\boldsymbol{f}_t|+\sum_{t=1}^{T}|\boldsymbol{y}_t-\boldsymbol{\Lambda}^{(m-1)}\boldsymbol{f}_t|$$

其中:$\boldsymbol{\Lambda}^{(m-1)}=(\lambda_1^{(m-1)},\cdots,\lambda_{n_1}^{(m-1)},\lambda_{n_1+1}^{(m-1)},\cdots,\lambda_N^{(m-1)})^{\mathrm{T}}$,$\boldsymbol{y}_t$ 是矩阵 $\boldsymbol{Y}$ 的第 t 列,$\boldsymbol{f}_t$ 是 F 的第 t 列。于是式(4)中问题可以分解为 T 个独立的子优化问题,求解相应的 $\boldsymbol{f}_t$:

$$\boldsymbol{f}_{t,L_2}=\operatorname{argmin}_{\theta}\sum_{s=1}^{n_1}(\boldsymbol{y}_{st}-\boldsymbol{\lambda}_s^{(m-1)\mathrm{T}}\boldsymbol{\theta})^2+\sum_{s=n_1+1}^{N}(\boldsymbol{y}_{st}-\boldsymbol{\lambda}_s^{(m-1)\mathrm{T}}\boldsymbol{\theta})^2 \tag{6}$$

$$\boldsymbol{f}_{t,L_1}=\operatorname{argmin}_{\theta}\sum_{s=1}^{n_1}|\boldsymbol{y}_{st}-\boldsymbol{\lambda}_s^{(m-1)\mathrm{T}}\boldsymbol{\theta}|+\sum_{s=n_1+1}^{N}|\boldsymbol{y}_{st}-\boldsymbol{\lambda}_s^{(m-1)\mathrm{T}}\boldsymbol{\theta}| \tag{7}$$

同理将式(5)转化为下面的 N 个独立的子优化问题:

$$\lambda_{s,L_2}=\operatorname{argmin}_{\theta}\sum_{t=1}^{a}(\boldsymbol{y}_{it}-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{f}_t)^2+\sum_{t=1}^{T}(\boldsymbol{y}_{it}-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{f}_t)^2 \tag{8}$$

$$\lambda_{s,L_1}=\arg\min_{\theta}\sum_{t=1}^{a}|\boldsymbol{y}_{it}-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{f}_t|+\sum_{t=1}^{T}|\boldsymbol{y}_{it}-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{f}_t| \tag{9}$$

其中:λ_s 是 $\boldsymbol{\Lambda}$ 的第 s 行。式(6)—式(9)都可以采用线性规划问题求解。在带有缺失值的情况下,直接舍弃目标函数中的对应累加项,在上述算法中对应的做法就是直接删除一个约束条件。

需要注意的是,交替凸优化算法只能保证在每一步求得当前最优解,并不能保证最后得到全局最优解。具体算法步骤如下(以 L_2 因子分析为例):

Step1 初始化参数:给出 $\boldsymbol{\Lambda}$、$\boldsymbol{\Sigma}$ 的初始值 $\boldsymbol{\Lambda}^{(0)}$、$\boldsymbol{\Sigma}^{(0)}$。

Step2 交替凸优化:对于迭代次数 $m(m=1,2,\cdots,M)$,有

$$\boldsymbol{F}^{(m)}=(f_1^{(m)},\cdots,f_T^{(m)})=\operatorname{argmin}_{F}\sum_{t=1}^{a}\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}^{(m-1)}\sum^{(m-1)}\boldsymbol{f}_t\right)^2+\sum_{t=1}^{T}\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}^{(m-1)}\sum^{(m-1)}\boldsymbol{f}_t\right)^2$$

$$\boldsymbol{\Lambda}^{(m)}=(\lambda_1^{(m)},\cdots,\lambda_N^{(m)})^{\mathrm{T}}=\operatorname{argmin}_{\Lambda}\sum_{s=1}^{n_1}\left(\boldsymbol{y}_t-\boldsymbol{\lambda}_s^{\mathrm{T}}\sum^{(m-1)}\boldsymbol{F}^{(m)}\right)^2+\sum_{s=n_1+1}^{N}\left(\boldsymbol{y}_t-\boldsymbol{\lambda}_s^{\mathrm{T}}\sum^{(m-1)}\boldsymbol{F}^{(m)}\right)^2$$

Step3 输出结果:$\boldsymbol{\Lambda}=\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{\frac{1}{2}}$,对 $\boldsymbol{\Lambda}$ 进行 QR 分解取正交矩阵得到 $\widetilde{\boldsymbol{\Lambda}}$,$\widetilde{\boldsymbol{F}}=\widetilde{\boldsymbol{\Lambda}}^{Y}$。

将算法中 L_2 范数改为 L_1 范数即为交替凸优化求解 L_1 因子分析。关于初始值的选取,$\boldsymbol{\Sigma}$ 为对角矩阵,使用单位矩阵作为初始值 $\boldsymbol{\Sigma}^{(0)}$;而 $\boldsymbol{\Lambda}$ 可以采用简单随机数进行初始化,这里为了加快收敛速度,使用均值插补后通过奇异值分解(SVD)算法得到的因子载荷矩阵作为初始值 $\boldsymbol{\Lambda}^{(0)}$。

2.5　预测效果评价

由于反事实值永远无法得到真实的观测值,所以在随机对照实验中无法获得预测误差,从而无法比较各个方法的预测效果。常用的解决方法是 Abadie 提出的安慰剂检验法,其在因果研究中用于检验反事实估计量是否具有稳健性。安慰剂(placebo)源于医学上的随机实验。研究者为检验药物是否有效,将实验人群随机分为服用真药的实验组与服用安慰剂的控制组,通过不让实验者知道自己服用的是真药还是安慰剂,来避免主观心理作用的影响。以此为基础的安慰剂检验,其核心思想在于从控制组中选取伪实验组,并用相同的方法估计虚拟的"反事实值",这样能同时获得真实的观测值与估计的预测值,即可对预测结果进行评价。由此得到的安慰剂效应(即真实值与虚拟反事实值之差)越趋于 0,说明其与该政策的因果效应差距越大,也就说明估计方法越稳健。在实证研究中,安慰剂检验受到了广泛使用。Abadie 通过假定控制组城市受到政策影响估计安慰剂效应,并作图比较,说明合成控制法的稳健性;Athey 等[13]则分别假设虚拟实验组与虚拟政策实施时间,对双重差分、带惩罚项的横向递减法(horizontal regression)、矩阵补全方法(Matrix Completion Methods)等多种方法进行了比较。

本文具体考虑以下两种情况:

(1) 随机选取控制组城市 j^*,假设其在 $t=T_{a+1},T_{a+2},\cdots,T_{a+b}$ 时间段中受到政策介入影响,成为伪实验组(pseudo-treat unit)。其余控制组城市($j^{(\cdot)}$)仍然为控制组,政策介入时间不变。具体变量设置由表 2 所示。

表 2　伪实验组假设

Table 2　Pseudo-treat assumption

	介入前	介入后
伪实验组	$Y^{\text{obs}}_{j^*,T_1},\cdots,Y^{\text{obs}}_{j^*,T_a}$	$Y^{\text{obs}}_{j^*,T_{a+1}},\cdots,Y^{\text{obs}}_{j^*,T_{a+b}}$
控制组	$Y^{\text{obs}}_{j^{(\cdot)},T_1},\cdots,Y^{\text{obs}}_{j^{(\cdot)},T_a}$	$Y^{\text{obs}}_{j^{(\cdot)},T_{a+1}},\cdots,Y^{\text{obs}}_{j^{(\cdot)},T_{a+b}}$

(2) 随机选取时刻 T_{a+c} 与 T_{a+d}(其中 T_{a+c} 满足时间顺序 $T_{a+1},T_{a+2},\cdots,T_{a+c},\cdots,T_{a+d};1<c<d$),假设实验组城市在 T_{a+c} 时刻再次受到政策介入影响。该情况下实验组与控制组城市不变,具体变量设置由表 3 所示。

表 3　伪时间假设

Table 3　Pseudo-time assumption

	伪介入前	伪介入后
实验组	$Y^{\text{obs}}_{i,T_{a+1}},\cdots,Y^{\text{obs}}_{i,T_{a+c}}$	$Y^{\text{obs}}_{i,T_{a+c+1}},\cdots,Y^{\text{obs}}_{i,T_{a+d}}$
控制组	$Y^{\text{obs}}_{j,T_{a+1}},\cdots,Y^{\text{obs}}_{j,T_{a+c}}$	$Y^{\text{obs}}_{j,T_{a+c+1}},\cdots,Y^{\text{obs}}_{j,T_{a+d}}$

在以上两种情形中,同样地将"介入后"的"实验组"数据视作缺失值,用上节的方法进行因子分析,并将得到的 L_1 因子和 L_2 因子以及对应因子载荷建立预测模型,预测"反事实值"。在安慰剂检验中,能同时获得真实的观测值与估计的预测值,对预测结果进行评价。下面选取 3 种指标:F_{MSE}、F_{MAE} 和 F_{MPAE},并由这 3 种指标比较各方法的预测精度。计算方法如式(10)所示(e_{it} 为实验组城市 i 在 t 时刻的预测误差):

$$\begin{cases} F_{\text{MSE}}=\dfrac{1}{N}\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\sum_{n=1}^{N}e_{it}^2 \\ F_{\text{MAE}}=\dfrac{1}{N}\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\sum_{n=1}^{N}\left|e_{it}\right| \\ F_{\text{MPAE}}=\dfrac{1}{N}\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\sum_{n=1}^{N}\left|\dfrac{e_{it}}{y_{it}}\right| \end{cases} \tag{10}$$

3　实证分析

3.1　数据选取与变量设置

本文选取因果推断研究中部分学者使用的关于加利福尼亚州限制吸烟政策的数据(Abadie 等,2010; Doudchenko 等,2016; Athey 等,2021)。使用该数据的优势在于:可以通过本文使用方法得出的结果与以往已有估计方法得到的结果进行比较,对因果效应是否存在进行基本验证。在该数据中,加利福尼亚州于 1988 年被限制吸烟政策介入,作为实验组;选取另外 38 个未被任何吸烟管控政策介入的州作为控制组。同时选取了 1970—2000 年共计 31a 间各州的烟草销量数据,并设定 1988 年为政策介入时刻(T_a),该政策于 1989 年(T_{a+1})起对烟草销量产生因果效应。

3.2　模型设置

实证研究中,面板数据 $\boldsymbol{Y}$ 为 39 行、31 列的矩阵。在具体因子分析模型中,$N=39,n_1=1,n_2=38$;$T=31,a=16,b=15$。采用 Bai 等提出的信息准则确定静态因子个数。在该信息准则下,因子个数需要

最小化:

$$Q(r)=\ln\left(\frac{V(\boldsymbol{\Lambda},\boldsymbol{F})}{NT}\right)+rG(N,T)$$

其中,$V(\boldsymbol{\Lambda},\boldsymbol{F})$ 为因子残差平方和,$G(N,T)$ 为惩罚函数,其使得在 $N,T\to\infty$ 时,$G(N,T)\to 0$,且 $\min(N,T)G(N,T)\to\infty$。参考 Bai&Ng(2002)建议,本文在实证研究中选择式(11)作为惩罚函数:

$$G(N,T)=\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right) \tag{11}$$

之后对目标函数 Q_2 与 Q_3 利用交替凸优化算法进行迭代,分别得到对应的潜在因子 $\boldsymbol{F}_t$ 与因子载荷 $\boldsymbol{\Lambda}_s$;再通过 $\boldsymbol{F}_t$ 与 $\boldsymbol{\Lambda}_s$ 进行估计,获得加州在没有被政策介入情况下的反事实预测值;最后由式(1)得到加州限制吸烟政策在不同方法下计算出的政策效应。

3.3 预测结果分析

实证研究中,除了本文因子分析相关的 3 种方法外,使用因果推断中常见的双重差分法(Difference-in-Difference, DID)以及 Doudchenko(2016)使用的弹性网络法(Elastic Networks)作为比较,由式(4)计算所得的政策效应(表 4)。

表 4 政策效应估计

Table 4 Policy effect estimation

估计方法	简单因子分析	L_2 因子分析	L_1 因子分析	双重差分	弹性网络
政策效应 τ	38.85	40.08	41.39	8.60	18.73

表 4 给出了用不同方法得到的限制吸烟政策对于烟草价格的因果效应。可以看出政策效应均为正,且本文所使用的 3 种方法效应更显著。具体每年的预测值如图 1 所示。

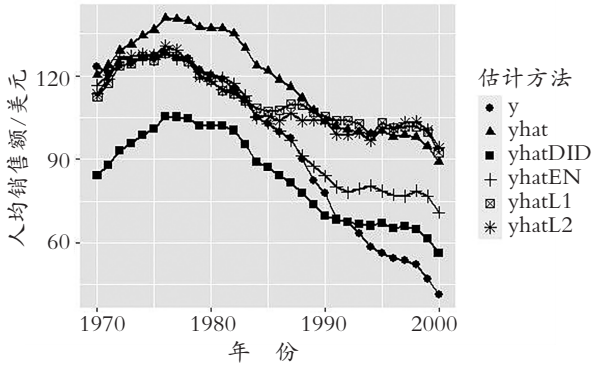

图 1 各方法每年预测值

Fig. 1 Estimation of each method

其中,线 y 为加州烟草销售额的实际变化。由图 1 可以看出,双重差分法(线 yhatDID)全时间段估计效果均较差,不能说明政策效应是否显著。经过均值插补后的简单因子分析(线 yhat)虽然在 1989 年政策介入后的估计与实际值有显著差异,能够说明政策的因果效应,但在政策介入前的时间段(1970—1988)与实际值相差较大。而 L_1 因子分析(线 yhatL1)、L_2 因子分析(线 yhatL2)以及弹性网络方法(线 yhatEN)均能在政策介入前与实际值保持一致,且在介入后与实际值表现出显著差距。

3.4 预测误差分析

伪实验组假设中,随机抽取 5 个原控制组州作为伪实验组,其余 33 个州仍为控制组。政策介入时刻、介入前后总时间不变,由式(10)计算所得误差如表 5 所示。

表 5 伪实验组假设预测误差

Table 5 Estimation error of pseudo-treat assumption

	简单因子分析	L_2 因子分析	L_1 因子分析
F_{MSE}	523.831	2.411	1.537
F_{MAE}	19.726	1.246	0.888
F_{MPAE}	0.194	0.013	0.010

由表 5 可以看出:通过均值插补后的简单因子分析得出的结果误差较大,而 L_1、L_2 因子分析的效果均不错,且 L_1 范数较 L_2 范数具有更小的平均预测误差。具体到每个伪实验组的平均预测误差,L_1、L_2 因子分析所得的箱线图如图 2 所示。

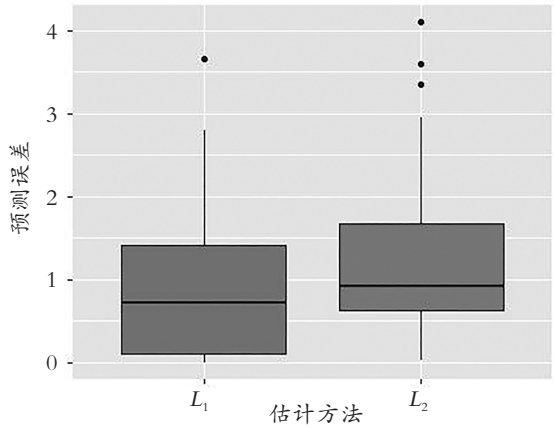

图 2 伪实验组假设预测误差

Fig. 2 Estimation error of pseudo-treat assumption

由图 2 可以看出:L_1 因子分析离群值点更少,最大值也更小,总体而言预测误差也小于 L_2 因子分析,可以认为该情况下,L_1 范数预测效果略好于 L_2 范数。

伪时间假设情况中,将 1999 年作为伪政策伪介入时刻(T_{a+c}),该政策于 2000 年(T_{a+c+1})起对烟草销

量产生因果效应。控制组与实验组不变,选取1989—2019 年共计 31a 间各州的烟草销量数据,由式(10)计算所得误差如表 6 所示。

表 6　伪时间假设预测误差

Table 6　Estimation error of pseudo-time assumption

	简单因子分析	L_2 因子分析	L_1 因子分析
F_{MSE}	903.833	2.325	1.971
F_{MAE}	29.862	1.087	0.733
F_{MPAE}	0.563	0.051	0.036

由表 6 可以看出:通过均值插补后的简单因子分析得出的结果误差较大,而 L_1、L_2 因子分析的效果均不错,且 L_1 范数与伪实验组假设一样具有更小的预测误差。具体到实验组每年的预测误差,L_1、L_2 因子分析所得的箱线图如图 3 所示。

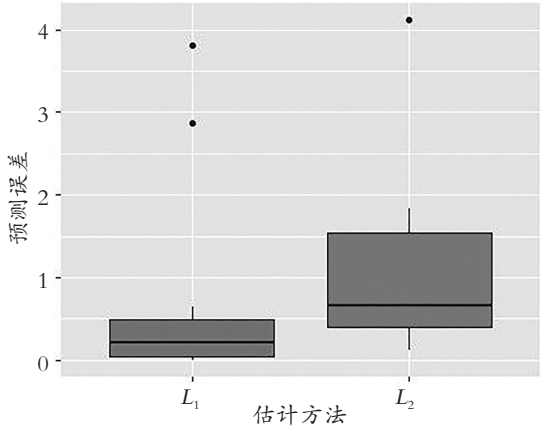

图 3　伪时间假设预测误差

Fig. 3　Estimation error of pseudo-time assumption

由图 3 可以看出:L_1 因子分析离群值点情况与 L_2 因子分析相差无几,而总体预测误差明显小于 L_2 因子分析,可以认为该情况下 L_1 范数预测效果同样好于 L_2 范数。综上所述,经均值插补后的简单因子分析表现较差,各个误差显著大于另外两种方法。而 L_1、L_2 因子分析的预测效果在不同情况下表现优秀,且 L_1 因子分析在两种假设下效果均略好于 L_2 因子分析。通过比较,可以认为 L_2 因子分析在进行因果推断时,虽然具有一定稳健性,但较之于 L_1 因子分析仍有所欠缺。

4　总　结

首先简要介绍了因果推断的提出与发展历程。通过已有文献,概述了因果推断的相关模型,确定其最终目的是反事实值估计,并介绍了已有的估计反事实值的方法与理论模型;之后引入因子模型的基础理论,包括基本概念和模型参数的估计方法:正交因子模型主要通过主成分分析法来估计,而主成分估计得出的因子得分可以很好地应用于宏观经济变量预测;由此,结合因果推断与正交因子模型,将因果推断反事实值估计转变为带有缺失值的潜在因子模型估计。

基于以上理论依据,舍弃面板数据中的缺失值,通过优化一步得到潜在结果与平均处理效应;并采用 L_1 因子分析代替 L_2 因子分析来估计模型,做出稳健性上的改进,获得中位数处理效应;通过对 L_1 因子分析的问题进行表述,介绍了一种交替凸优化算法并给出其实现步骤。

最后,为了验证 L_1 因子分析能否代替 L_2 因子分析作为因子模型的估计,基于加利福尼亚州限制烟草政策案例作了实证研究,将两种不同的主成分估计应用在反事实值的预测上。实证研究的结果表明:因子模型的 L_1 估计量同样适用于宏观经济变量预测。后又通过伪实验组与伪介入的假设,分析比较了 L_1、L_2 因子分析的预测效果,结果表明 L_1 因子分析较之其他方法具有更稳健的预测效果。

参考文献(References):

[1] RUBIN D B. Estimating causal effects of treatments in randomized and nonrandomized studies [J]. Journal of Educational Psychology, 1974, 66: 688—701.

[2] HECKMAN J J, ROBB J R R. Alternative methods for evaluating the impact of interventions: an overview [J]. Journal of Econometrics, 1985, 30(1-2): 239—267.

[3] HECKMAN J J, ROBB J R R. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes: in drawing inferences from self-selected samples[M]. Springer: New York, NY, 1986.

[4] ABADIE A, DIAMOND A, HAINMUELLE J. Synthetic control methods for comparative case studies: estimating the effect of California' s tobacco control program [J]. Journal of the American Statistical Association, 2010, 105 (490): 493—505.

[5] ZHENG Y, ZHENG H, YE X. Using machine learning in environmental tax reform assessment for sustainable development: a case study of Hubei province, China [J]. Sustainability, 2016, 8(11): 1124.

[6] DOUDCHENKO N, IMBENS G. Balancing, regression, difference-in-differences and synthetic control methods: a synthesis [R]. Massachusetts: National Bureau of

Economic Research, 2016.

[7] XIONG R X, PELGER M. Large dimensional latent factor modeling with missing observations and applications to causal inference[J]. Journal of Econometrics, 2022, 208(1): 23—43.

[8] BAI J S, NG S. Matrix completion, counterfactuals, and factor analysis of missing data [J]. Journal of the American Statistical Association, 2021, 116(536): 1746—1763.

[9] BAI J S, NG S. Determining the number of factors in approximate factor models[J]. Econometrica, 2002, 70(1): 191—221.

[10] FAN J Q, Liao Y, MINCHEVA M. Large covariance estimation by thresholding principal orthogonal complements[J]. Journal of the Royal Statistical Society Series B, Royal Statistical Society, 2013, 75(4): 603—680.

[11] KONG X B. On the number of common factors underlying large panel high-frequency data[J]. Biometrika, 2017, 104: 397—410.

[12] CHEN L, DOLADO J J, GONZALO J. Quantile Factor Models[J]. Econometrica, 2021, 89(2): 875—910

[13] ATHEY S, BAYATI M, DOUDCHENKON, et al. Matrix completion methods for causal panel data models[J]. Journal of the American Statistical Association, 2021, 116(536): 1716—1730.

Research on Causal Inference Based on Factor Analysis

FU Ju-wang, KONG Xin-bing

(School of Statistics and Data Science, Nanjing Audit University, Nanjing 211815, China)

Abstract: According to the causal inference problem under the counter-factual framework, an L_2 factor analysis method from the perspective of factor analysis and optimization is proposed to estimate counter-factual value, and L_1 loss function is introduced to optimize L_1 risk. Combining causal inference with orthogonal factor model, the counter-factual value, which is supposed to estimate, is regarded as missing value, transforming the counter-factual value estimation of causal inference into latent factor model estimation with missing value. Discarding the missing value in the panel data, the latent results and mean treatment effect are derived directly by optimization, and therefore avoid the loss of information; using L_1 factor analysis instead of L_2 factor analysis to estimate the model, making robustness improvements and obtaining the median treatment effect. Alternate convex programming is introduced to minimize the objective function in the L_1 and L_2 factor analysis and its implementation steps are given. Empirical study based on the case of tobacco policy in California is made to compare L_1, L_2 factor analysis and other causal inference methods. The results show that the L_1 and L_2 factor estimator is also applicable to the prediction of macroeconomic variables. Finally, by setting up pseudo-treat unit assumption and pseudo-time assumption respectively, it has been verified that L_1 has more robust prediction effect than other methods.

Key words: counter-factual estimation; causal inference; L_1 factor analysis

责任编辑:李翠薇

引用本文/Cite this paper:

付举望, 孔新兵. 基于因子分析的因果推断研究[J]. 重庆工商大学学报(自然科学版),2022,39(6):71—78.

FU Ju-wang, KONG Xin-bing. Research on causal inference based on factor analysis [J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(6):71—78.