合成控制法

合成控制法（SCM）被广泛用于经济学和社会科学的实证研究，常用于政策评估。合成控制法适用于面板数据，其原理是通过选择权重，加权后构造一个合成的控制组，使得其结果与处理组干预前结果相匹配，从而通过反事实估计得到处理效应。合成控制法被称为“过去15年政策评估文献中最重要的创新”（Athey & Imbens，2017）。

合成控制法的基本设定

基本原理通过对控制组个体进行加权构造出一个合成的控制组，使得合成控制组与处理组干预前的特征十分相似，从而可以作为处理组的反事实估计。

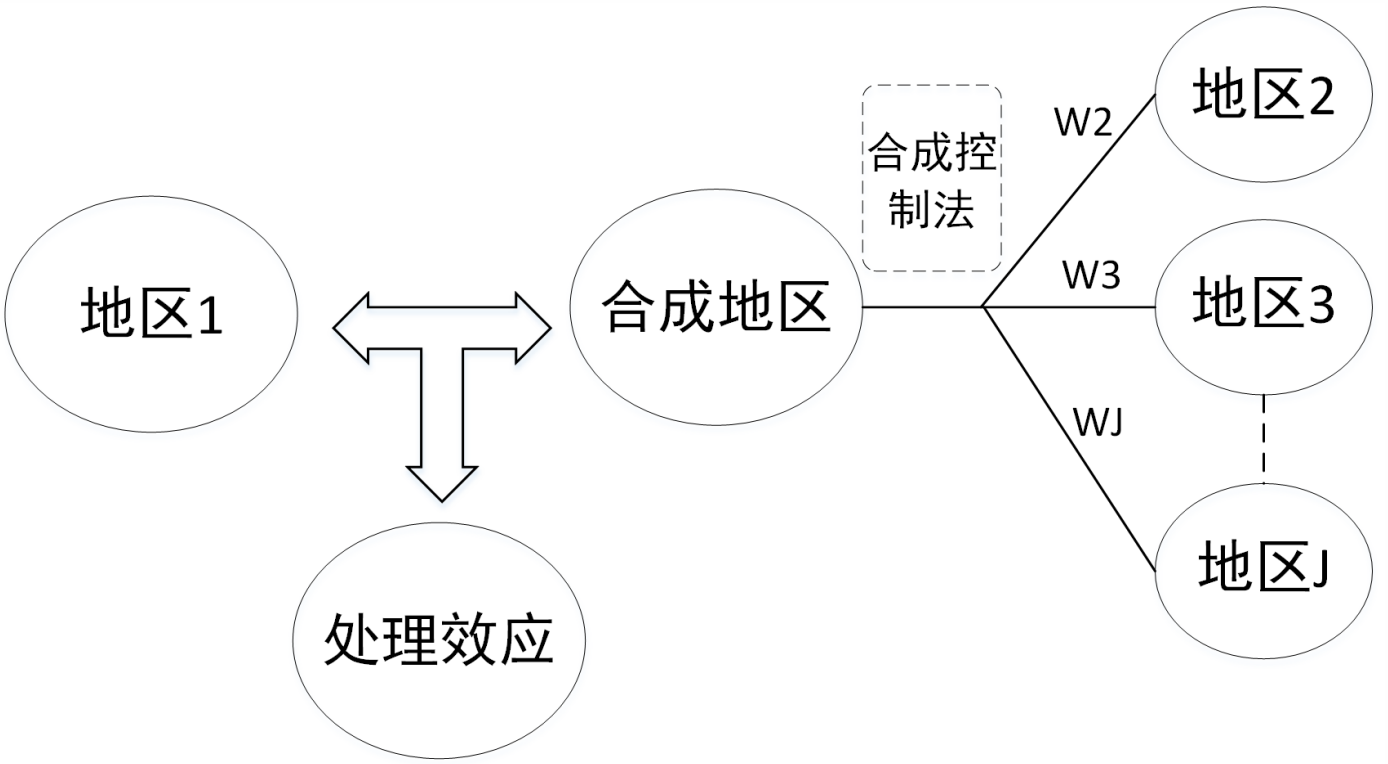

图1 合成控制法原理示意图

模型设定：假设有$J+1$个地区，地区1在T_0期后受到政策干预，而其他J个地区没有受到政策干预。Y_{it}^N表示个体i在t期没有受到政策干预时的潜在结果，Y_{1t}^I为实际干预后的观测结果。其中 $i=1,\ldots,J+1, t=1,\ldots,T$。令$T_0$ 为干预前期数，T_0后处理组受到政策干预。

$$D_{it}=\begin{cases}1 & \text{if } i=1 \text{ and } t>T_0\\ 0 & \text{otherwise}\end{cases} \tag{1-1}$$

所以个体i在t期的观测结果为：

$$Y_{it}=D_{it}Y_{1t}^I+(1-D_{it})Y_{1t}^N=Y_{1t}^N+\tau_{it}D_{it} \tag{1-2}$$

因此，对于$t>T_0$，政策效应为：

$$\tau_{1t}=Y_{1t}-Y_{1t}^N \tag{1-3}$$

个体1在$t > T_0$期受到政策干预，我们可以直接观察到受到干预时的结果Y_{1t}，而无法观测到若它没有受到政策干预情况时的潜在结果Y_{it}^N。因此政策评估的关键是估计出潜在的反事实结果Y_{it}^N。类似于DID，假设Y_{it}^N用以下因子模型表示：

$$Y_{it}^N = \delta_t + \boldsymbol{\theta}_t \mathbf{Z}_i + \lambda_t \boldsymbol{\mu}_i + \varepsilon_{it} \tag{1-4}$$

其中，δ_t为时间趋势，所有个体都具有相同的时间趋势。Z_i是 $(r \times 1)$ 维（不受政策影响的）可观测协变量向量，$\boldsymbol{\theta}_t$ 是$(1 \times r)$ 维未知系数向量，λ_t 是 $(1 \times F)$ 维未观测公共因子（混杂因素）， μ_i 是$(F \times 1)$维系数向量，误差项 ε_{it} 表示未观测的暂时性冲击，假设在地区层面满足零均值。

从结构上看，（1-4）式对传统的双重差分模型（DID）进行了扩展。DID允许存在不能观察到的混杂因素，但是限制这些混杂因素的影响为时间上的常数，因此可以通过取时间差分来消除。而合成控制法的因子模型设定λ_t不为常数，允许未观察到的混淆因素的影响随时间变化，因此取时间差并不能消除混杂因素，因此合成控制法提供了更广泛的有效估计。

现考虑$(J \times 1)$的权重向量$\mathbf{W} = (w_2, \ldots, w_{J+1})'$ ，满足 $w_j \geq 0$ ， $j = 2, \ldots, N+1$ 并且$w_2 + \cdots + w_{J+1} = 1$。这里将权重限制为非负，相当于用控制组个体的凸组合来合成控制组，是为了避免外推造成的可能偏差。因此合成的模型表达式为：

$$\sum_{j=2}^{J+1} w_j Y_{jt} = \delta_t + \boldsymbol{\theta}_t \sum_{j=2}^{J+1} w_j \mathbf{Z}_j + \lambda_t \sum_{j=2}^{J+1} w_j \boldsymbol{\mu}_j + \sum_{j=2}^{J+1} w_j \varepsilon_{jt} \tag{1-5}$$

假设存在权重向量 $\left(w_2^*, \ldots, w_{J+1}^*\right)$满足：

$$\begin{aligned} &\sum_{j=2}^{J+1} w_j^* Y_{j1} = Y_{11}, \quad \sum_{j=2}^{J+1} w_j^* Y_{j2} = Y_{12}, \quad \ldots \\ &\sum_{j=2}^{J+1} w_j^* Y_{jT_0} = Y_{1T_0}, \quad \text{and} \quad \sum_{j=2}^{J+1} w_j^* \mathbf{Z}_j = \mathbf{Z}_1 \end{aligned} \tag{1-6}$$

Abadie et al.(2010)在附录中证明，如果$\sum_{t=1}^{T_0} \lambda_t' \lambda_t$ 是非奇异的，则有式子（1-7），并且证明当干预之前的时期足够长（$T_0 \to \infty$）,式子（1-7）趋近于0。

$$\begin{aligned} Y_{1t}^N - \sum_{j=2}^{J+1} w_j^* Y_{jt} = &\sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' \left(\varepsilon_{js} - \varepsilon_{1s} \right) \\ &- \sum_{j=2}^{J+1} w_j^* \left(\varepsilon_{jt} - \varepsilon_{1t} \right) \end{aligned} \tag{1-7}$$

从而干预组个体1的反事实结果近似可以用合成控制组组来进行表示，即：

$$\widehat{Y}_{1t}^N = \sum_{j=2}^{J+1} \omega_j^* Y_{jt} \tag{1-8}$$

因而，处理组（干预组）个体1的政策处理效应可以表示为：

$$\widehat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}, \ t \in \{T_0 + 1, \ldots, T\} \tag{1-9}$$

条件（1-6）式是关键，如果存在权重向量W^*，使得干预前各合成控制组的观测结果与处理组观测结果相等，即所有可观测因素相同，从而事前合成控制组的未观测因素也会与干预组未观测因素相同，即$\sum_{j=2}^{J+1} w_j^* \boldsymbol{\mu}_j = \boldsymbol{\mu}_1$。这意味着合成控制组与干预组将非常相似，从而可以将合成控制组的结果作为干预组个体反事实结果的估计。

估计变量的无偏性证明

现在来证明：在满足合成控制权重条件（1-6）下得到的反事实估计量在T_0期后是Y_{1t}^N的无偏估计。Abadie et al. (2010)在附录中的证明过程如下：

估计偏差表达式为：

$$\begin{aligned} Y_{1t}^N-\sum_{j=2}^{J+1}w_jY_{jt}^N=\boldsymbol{\theta}_t\left(\mathbf{Z}_1-\sum_{j=2}^{J+1}w_j\mathbf{Z}_j\right) \\ +\lambda_t\left(\boldsymbol{\mu}_1-\sum_{j=2}^{J+1}w_j\boldsymbol{\mu}_j\right)+\sum_{j=2}^{J+1}w_j\left(\varepsilon_{1t}-\varepsilon_{jt}\right). \end{aligned} \tag{1-10}$$

令$\mathbf{Y}_i^P$ 为 $T_0\times1$ 向量，且第t个元素为 Y_{it}。同样的， ε_i^P 为 $(T_0\times1)$ 向量，且第t个元素为 ε_{it}。$\boldsymbol{\theta}^P$ 和λ^P 分别为 $(T_0\times r)$ 矩阵和$(T_0\times F)$ 矩阵。根据前T_0期数据，我们有：

$$\begin{aligned} \mathbf{Y}_1^P-\sum_{j=2|}^{J+1}w_j\mathbf{Y}_j^P=\boldsymbol{\theta}^P\left(\mathbf{z}_1-\sum_{j=2}^{J+1}w_j\mathbf{Z}_j\right) \\ +\lambda^P\left(\boldsymbol{\mu}_1-\sum_{j=2}^{J+1}w_j\boldsymbol{\mu}_j\right)+\sum_{j=2}^{J+1}w_j\left(\boldsymbol{\varepsilon}_1^P-\boldsymbol{\varepsilon}_j^P\right). \end{aligned} \tag{1-11}$$

令 $\xi(M)$ 为$\frac{1}{M}\sum_{t=T_0-M+1}^{T_0}\lambda_t'\lambda_t$的最小特征值。假设$\xi(M)\geq\xi>0$。 假设存在常数$\bar{\lambda}$， $|\lambda_{tf}|\leq\bar{\lambda}$ ，$t=1,\ldots,T,f=1,\ldots,F$。 因为 $\lambda^{P'}\lambda^P$ 是非奇异的，则有：

$$\begin{aligned} Y_{1t}^N-\sum_{j=2}^{J+1}w_jY_{jt}^N=&\lambda_t\left(\lambda^{P'}\lambda^P\right)^{-1}\lambda^{P'}\left(\mathbf{Y}_1^P-\sum_{j=2}^{J+1}w_j\mathbf{Y}_j^P\right) \\ &+\left(\boldsymbol{\theta}_t-\lambda_t\left(\lambda^{P'}\lambda^P\right)^{-1}\lambda^{P'}\boldsymbol{\theta}^P\right)\left(\mathbf{Z}_1-\sum_{j=2}^{J+1}w_j\mathbf{Z}_j\right) \\ &-\lambda_t\left(\lambda^{P\prime}\lambda^P\right)^{-1}\lambda^{P'}\left(\varepsilon_1^P-\sum_{j=2}^{J+1}w_j\varepsilon_j^P\right) \\ &+\sum_{j=2}^{J+1}w_j\left(\varepsilon_{1t}-\varepsilon_{jt}\right) \end{aligned} \tag{1-12}$$

假定存在 $\left\{w_2^*,\ldots,w_{J+1}^*\right\}$ 满足条件（1-6）式的权重条件，则有

$$Y_{1t}^N-\sum_{j=2}^{J+1}w_j^*Y_{jt}^N=R_{1t}+R_{2t}+R_{3t} \tag{1-13}$$

其中

$$\begin{aligned} R_{1t}&=\lambda_t\left(\lambda^{P'}\lambda^P\right)^{-1}\lambda^{P'}\sum_{j=2}^{J+1}w_j^*\varepsilon_j^P \\ R_{2t}&=-\lambda_t\left(\lambda^{P\prime}\lambda^P\right)^{-1}\lambda^{P\prime}\varepsilon_1^P \\ R_{3t}&=\sum_{j=2}^{J+1}w_j^*\left(\varepsilon_{jt}-\varepsilon_{1t}\right) \end{aligned} \tag{1-14}$$

考虑 $t>T_0$ 的情况， R_{2t} 和 R_{3t} 有零均值。现在主要是证明 R_{1t} 趋近于0。Abadie et al.(2010)在附录中得到R_{1t}式如下：

$$E\left|R_{1t}\right| \leq C(p)^{1/p}\left(\frac{\bar{\lambda}^2 F}{\xi}\right) J^{1/p} \max\left\{\frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{\sigma}}{T_0^{1/2}}\right\} \tag{1-15}$$

其中 $C(p)$表示-1加上参数为1的泊松随机变量的p阶矩。令$\sigma_{jt}^2=E|\varepsilon_{jt}|^2$，　$\sigma_j^2=(1/T_0)\sum_{t=1}^{T_0}\sigma_{jt}^2$，　$\bar{\sigma}^2=\max_{j=2,\ldots,J+1}\sigma_j^2$, 以及 $\bar{\sigma}=\sqrt{\bar{\sigma}^2}$。同样的, 令$m_{p,jt}=E|\varepsilon_{jt}|^p, m_{p,j}=(1/T_0)\sum_{t=1}^{T_0}m_{p,jt}$, 以及 $\bar{m}_p=\max_{j=2,\ldots,J+1}m_{p,j}$。方程（15）表明，随着处理前周期数到$T_0$的增加，$E|R_{1t}|$趋于0，因此证明了估计量的偏差趋近于0。

合成控制法的估计过程

权重的确认

合成控制法实施的关键是找到满足权重条件（1-6）的权重向量。令 $\mathbf{W}$ 为$(J\times 1)$ 向量，即 $\mathbf{W}=(w_2,\ldots,w_{J+1})'$，其中$w_j\geq 0$、$w_2+\cdots+w_{J+1}=1$。$X_1$为干预组个体的事前的特征，包括可观测协变量$Z_1$和事前结果的若干线性组合，为$(k\times 1)$向量。具体的，定义$(T_0\times 1)$ 向量 $\mathbf{K}=(k_1,\ldots,k_{T_0})'$，干预前结果的一个线性组合可表示为: $\bar{Y}_i^{\mathrm{K}}=\sum_{s=1}^{T_0}k_sY_{is}$。例如, 当$k_1=k_2=\cdots=k_{T_0-1}=0$，$k_{T_0}=1$, 则 $\bar{Y}^{\mathrm{K}}=Y_{iT_0}$；若 $k_1=k_2=\cdots=k_{T_0}=1/T_0$，则$\bar{Y}_i^{\mathrm{K}}=T_0^{-1}\sum_{s=1}^{T_0}Y_{is}$。考虑有$M$ 个这样的线性组合 $\mathbf{K}_1,\ldots,\mathbf{K}_M$，则干预组个体的事前特征的协变量可表示为 $\mathbf{X}_1=\left(\mathbf{Z}_1',\bar{Y}_1^{\mathbf{K}_1},\ldots,\bar{Y}_1^{\mathbf{K}_M}\right)'$，为 $(k\times 1)$ 向量，$k=r+M$。同样的，X_0为控制组的事前特征，$\mathbf{X}_0=\left(\mathbf{Z}_j',\bar{Y}_j^{\mathbf{K}_1},\ldots,\bar{Y}_j^{\mathbf{K}_M}\right)'$是 $(k\times J)$ 向量。

合成控制法的权重$\mathbf{W}^*=\left(w_2^*,\ldots,w_{J+1}^*\right)'$ 为最小化下面的距离（Abadie and Gardeazabal（2003 ）；Abadie et al. (2010)）：

$$\left\|\mathbf{X}_1-\mathbf{X}_0\mathbf{W}\right\|_{\mathbf{V}}=\sqrt{(\mathbf{X}_1-\mathbf{X}_0\mathbf{W})'\mathbf{V}\left(\mathbf{X}_1-\mathbf{X}_0\mathbf{W}\right)} \tag{1-16}$$

其中 $\mathbf{V}$ 是一个 $(k\times k)$ 的对称正定矩阵，通常为对角阵，对角元素为$v_1,\ldots,v_k$ 。即可写成:

$$\left\|\mathbf{X}_1-\mathbf{X}_0\boldsymbol{W}\right\|=\left(\sum_{h=1}^{k}v_h\left(X_{h1}-w_2X_{h2}-\cdots\quad -w_{J+1}X_{hJ+1}\right)^2\right)^{1/2} \tag{1-17}$$

正常数 $v_1,\ldots,v_k$反映合成控制中特征变量的相对重要性。 对于给定的一组权重 $v_1,\ldots,v_k$，会得到相应的最优化的 $\mathbf{W}^*=\left(w_2^*,\ldots,w_{J+1}^*\right)'$。V的选择很重要。

$\mathbf{V}$的确认，有不同的方法。可以是研究者对各协变量的主观评价，也可以通过回归分析哪些协变量预测能力更强。一个较好的办法是根据 Abadie et al.（2010 ）的方法选择$\mathbf{V}$，使用事前均方预测误差（ mean squared prediction error (MSPE)）最小的矩阵V，即：

$$\sum_{t=1}^{T_0}(Y_{jt}-\sum_{i=2}^{J+1}\omega_j^*(V)Y_{jt})^2 \tag{1-18}$$

如果事前时期足够长，可以使用Hainmueller(2015)提出的样本外验证方法，将事前样本分成训练期和验证期。步骤为：

1、将事前样本分成训练期和验证期。

2、给定任意的矩阵V，利用训练期数据计算最优权重矩阵$W^*(V)$ 。

3、然后再利用验证期数据，选择V^*使得均方预测误差最小化。

合成控制法的假设检验

在利用合成控制法进行研究时，一般个体数不会太多，因而基于大样本的假设检验方法往往不合适。Abadie et al. (2010) 提出了一种基于置换检验（permutation test）的推断方法。

检验方法：安慰剂检验（placebo test）。

原假设是H0：政策效应不显著。即假设政策干预对个体没有因果影响。

检验原理：从控制组个体中随机抽出一个个体作为伪干预组，同样地利用合成控制法去估计政策干预效应。因此对应于J个控制组个体，得到J个相应的政策效应估计，然后将处理组和控制组的政策效应合并构建排列分布（permutation distribution）。当处理组的干预效果处在排列分布的极端时，比如处于尾部的5%，则说明干预组的政策效应是显著的。但如果发现处理组的政策效应处在分布的中间位置，则说明任取一个控制组个体也能得到相应的因果效应，说明因果效应不一定是由于政策实施造成的。

相应的**统计量表示**：参考Abadie et al. (2010)、Abadie（2021）有：

$$R_j\left(t_1, t_2\right)=\left(\frac{1}{t_2-t_1+1} \sum_{t=t_1}^{t_2}\left(Y_{jt}-\hat{Y}_{jt}^N\right)^2\right)^{1/2} \tag{1-19}$$

$$r_j=\frac{R_j\left(T_0+1, T\right)}{R_j\left(1, T_0\right)} \tag{1-20}$$

r_j表示个体j 在政策干预前后结果的差异比率，基于 r_j的置换分布的假设检验的 p 值表示如下：

$$p=\frac{1}{J+1} \sum_{j=1}^{J+1} I_{+}\left(r_j-r_1\right) \tag{1-21}$$

其中 $I_{+}(\cdot)$ 是一个指示函数，对于非负参数返回 1，否则返回 0。这里的 p 值类似于统计推断中的显著性水平。例如，在对样本总体$J+1$的个体进行安慰剂检验中，如果处理个体1的r_1都大于其余J个控制组的r_j，则$p=\frac{1}{J+1}$，若此p<0.05,则可表明在 5%的水平上是显著的。

实证论文中，一般有两种可报告/可视化的安慰剂检验：

1、安慰剂差距的排列分布图：$Y_{jt}-\hat{Y}_{jt}^N$。基于RMSPE的排列分布。这种形式更常用到。

2、r_j的排序分布（permutation distribution）。

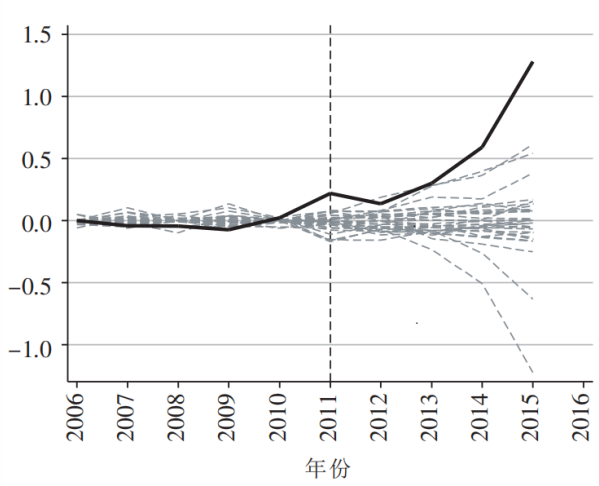

图 3　各城市工业相对产值差值分布

注：实线表示重庆，虚线表示 RMSPE 值比重庆 1.5 倍低的城市。

图2（a）安慰剂差距的排列分布图

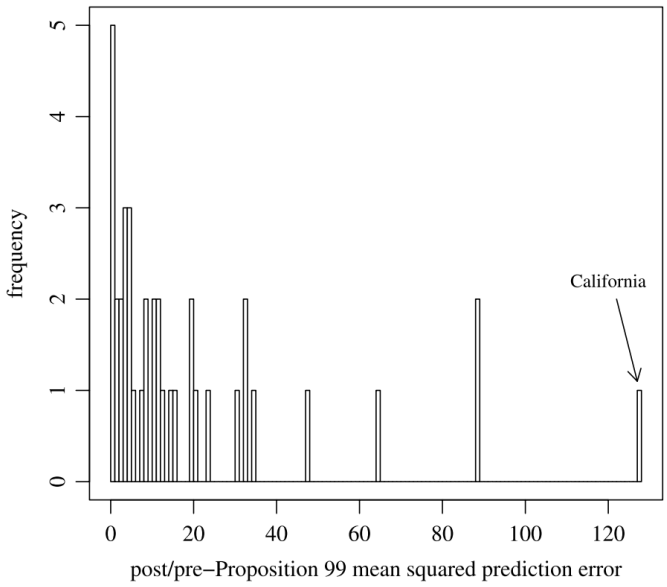

图2（b）r_j的排序分布

注意：当个体的合成控制对象在政策实施前的拟合效果不好时，就不再考虑将这个个体加入到排序检验分布中。理由是：控制组个体的处理效果差异可能是因为拟合效果不好导致的，与政策干预无关。

SCM实施的注意点

1）Abadie et al.（2015）探讨了使用合成控制法的前提与注意事项，指出在构建潜在控制地区时，应该将控制组中受到政策影响的地区除去；个体特征变量必须是干预前的变量或者不受政策干预影响的变量；在样本期间受到很大特殊冲击的地区应该排除在控制组以外；为了避免“内插偏差（interpolation bias），应该将控制组选择限定在处理组具有相似特征的单元。

2）过度拟合风险。对于任何给定的 T_0，较大的J会使得控制组更容易拟合处理组处理前结果，即使处理组个体和合成控制之间的因子载荷存在显著差异。因此，控制组个体过多可能会导致过度拟合风险。

3）并不是T_0越大偏差越小，偏差也与拟合程度有关。由于Abadie et al.(2010) 中的偏差界限是在 $X1 = X_0W^*$ 下导出的，其估计效果也取决于合成控制再现处理个体结果轨迹的能力。 当$T_0 \to \infty$时，相当大的偏差可能会持续存在，除非拟合的质量 $X1 = X_0W^*$ 很好。 也就是说，若控制组拟合效果不好，合成控制法就不一定适用。

合成控制法的优点

合成控制方法（Synthetic Control Method）的优点如下：

1）该方法选出的控制组相对客观，透明度高，比较可信。通过数据驱动来选择线性组合的最优权重，避免了研究者主观选择控制组的随意性。

2）反事实的透明度。 合成控制明确了每个控制单元对处理组的反事实的贡献，通过对多个控制对象加权来模拟目标对象政策实施前的情况，不仅可以清晰地反映每个控制对象对"反事实"事件的贡献，同时也避免了过分外推。

与双重差分（DID）和倾向得分匹配（PSM）方法相比，优点如下：

1）双重差分法对参照组的选择具有主观性和随意性，缺乏说服力。 而相比较而言，合成控制法是通过数据驱动构造控制组。

2）倾向得分匹配法是基于条件独立的假设建立参照组模拟随机试验来分析政策的影响。 虽然倾向得分匹配法与合成控制法都是利用参照组的信息构造人工对照组，**但倾向得分匹配法将面板数据作为个体形成的混合数据**，不能分析个体的具体情况，个体与年份的交错将会导致结果的偏差（苏治和胡迪，2015）。

合成控制法的局限性

合成控制法的局限性如下：

1）合成控制组法对需要的面板数据比较苛刻。要求政策实施前的时期T_0较长，因为合成控制法的可信度取决于合成控制组能否在干预前相当一段时期内很好地追踪处理地区的特征与结果变量。如果干预前时期太短，则不建议使用合成控制法。此外，由于政策冲击的效应需要一段时间才会显现，也要求干预后期数足够大（Abadie et al.，2015）。

2）合成控制法要求合成控制权重必须为正，如果干预组观测特征远远大于或小于控制组个体特征，将无法找到合适的权重拟合干预组个体，此时将无法使用合成控制法。

合成控制法的案例实证

案例：房产税对产业转移的影响：来自重庆和上海的经验证据。

刘友金和曾小明（2018）利用合成控制法分析了 2011 年以来重庆和上海实施房产税这一政策干预对产业转移的影响。作者使用的数据是 2006—2015 年 35 个大中城市的平衡面板数据，来实证分析房产税征收对重庆和上海产业转移的影响，数据来源于历年《中国城市统计年鉴》和国家统计局网站。 这篇文章的被解释变量为产业转移变量，该指标通过使用相对产值和相对就业率来衡量；控制变量包括相对工资、人均 GDP、财政支出占GDP 比重、人口密度、年末金融机构存款余额、医院卫生院床位数、国际互联网用户数。

以分析房产税征收对重庆产业转移影响为例。stata的代码为：

```
// 图1(a) 真实重庆与合成重庆的相对工业总产值
preserve
drop if 城市=="上海"
synth 工业相对产值 工业相对产值(2006(1)2010) 相对工资 ln人均GDP 财政支出占GDP比重 ln人口密度人平方公里  ln年末金融机构存款余额万元 ln医院卫生院床位数张 ln国际互联网用户数户  工业相对产值(2006) 工业相对产值(2008) 工业相对产值(2010), trunit(26) trperiod(2011) nested fig
restore

//图1(b) 真实重庆与合成重庆的工业相对就业率
preserve
drop if 城市=="上海"
synth 工业相对就业率 工业相对就业率(2006(1)2010) 相对工资 ln人均GDP 财政支出占GDP比重 ln人口密度人平方公里  ln年末金融机构存款余额万元 ln医院卫生院床位数张 ln国际互联网用户数户  工业相对就业率(2006) 工业相对就业率(2008) 工业相对就业率(2010), trunit(26) trperiod(2011) nested fig
```

```
restore

//图1（c） 真实重庆与合成重庆的第三产业相对产值
preserve
drop if 城市=="上海"
synth 第三产业相对产值 第三产业相对产值(2006(1)2010) 相对工资 ln人均GDP 财政支出占GDP比重 ln人口密度人平方公里 ln年末金融机构存款余额万元 ln医院卫生院床位数张 ln国际互联网用户数户 第三产业相对产值(2006) 第三产业相对产值(2008) 第三产业相对产值(2010), trunit(26) trperiod(2011) nested fig
restore

//图1（d）真实重庆与合成重庆的第三产业相对就业率
preserve
drop if 城市=="上海"
synth 第三产业相对就业率 第三产业相对就业率(2006(1)2010) 相对工资 ln人均GDP 财政支出占GDP比重 ln人口密度人平方公里 ln年末金融机构存款余额万元 ln医院卫生院床位数张 ln国际互联网用户数户 第三产业相对就业率(2006) 第三产业相对就业率(2008) 第三产业相对就业率(2010), trunit(26) trperiod(2011) nested fig
restore
****
```

重庆和对应的合成控制城市在 2006—2015 年间的工业相对产值和工业相对就业率如图 3（a）和图 3（b）所示，其中垂直虚线所在的位置表示房产税改革实施的年份。可以发现，开征房产税能够促使重庆的工业相对产值和工业相对就业率大幅度增加，这种“促增作用”表明房产税政策显著促进了工业向重庆转移。

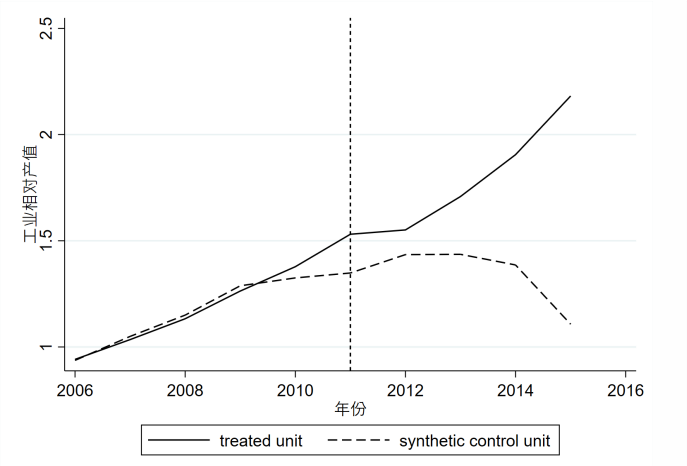

图3（a） 重庆与合成重庆的工业相对产值

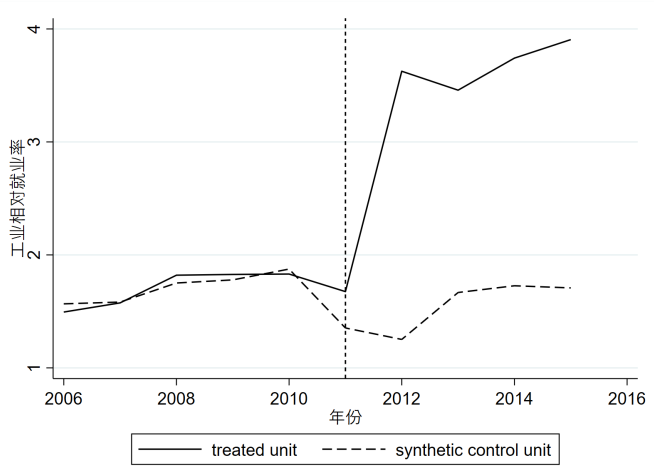

图3（b） 重庆与合成重庆的工业相对就业率

重庆结果的稳健性检验：现对结果进行安慰剂检验，这里以重庆房产税对工业相对产值影响的稳健性检验为例。

稳健性检验代码：

```
****稳健性检验*****************************
//有效性检验（仅展示重庆房产税对工业相对产值影响的稳健性检验程序）
//图3 各城市工业相对产值差值分布
************************************稳健性检验一(工业相对产值为目标变量)************************************
//政策实施前均方预测误差的平方根
tempname resmat
        forvalues i = 1/35 {
```

```
        synth 工业相对产值 相对工资 ln人均GDP 财政支出占GDP比重 ln人口密度人平方公里  ln年末金融机构存款余额万元 ln医院卫生院床位数张 ln国际互联网用户数户  工业相对产值(2006) 工业相对产值(2008) 工业相对产值(2010), trunit(`i') trperiod(2011) xperiod(2006(1)2010) mspeperiod
        matrix `resmat' = nullmat(`resmat') \ e(RMSPE)
        local names `"`names' `"`i'"'"'
        }
        mat colnames `resmat' = "RMSPE"
        mat rownames `resmat' = `names'
        matlist `resmat' , row("Treated Unit")
        ** loop through units

//各城市预测误差分布图
forval i=1/35{
qui synth 工业相对产值 相对工资 ln人均GDP 财政支出占GDP比重 ln人口密度人平方公里  ln年末金融机构存款余额万元 ln医院卫生院床位数张 ln国际互联网用户数户  工业相对产值(2006) 工业相对产值(2008) 工业相对产值(2010), xperiod(2006(1)2010) trunit(`i') trperiod(2011) keep(synth_`i', replace)
}

forval i=1/35{
use synth_`i', clear
rename _time years
gen tr_effect_`i' = _Y_treated - _Y_synthetic
keep years tr_effect_`i'
drop if missing(years)
save synth_`i', replace
}
**

use synth_1, clear
forval i=2/35{
qui merge 1:1 years using synth_`i', nogenerate
}

**
**删除拟合不好的城市及上海市（干预组）
drop tr_effect_2     //删除天津
drop tr_effect_20    //删除武汉
drop tr_effect_35    //删除上海


local lp1
forval i=1/1 {
   local lp1 `lp1' line tr_effect_`i' years, lpattern(dash) lcolor(gs8) ||
}
**
local lp2
forval i=3/19 {
   local lp2 `lp2' line tr_effect_`i' years, lpattern(dash) lcolor(gs8) ||
}

 local lp3
forval i=21/34 {
```

```
   local lp3 `lp3' line tr_effect_`i' years, lpattern(dash) lcolor(gs8) ||
}

 **create plot
twoway `lp1' `lp2' `lp3'  || line tr_effect_26 years, ///
lcolor(black) legend(off) xline(2011, lpattern(dash))
```

合成控制法的stata原代码：

```
synth depvar predictorvars,trunit(#) trperiod(#)[counit(numlist) xperiod(numlist)
mspeperiod() resultsperiod() nested allopt unitnames(varname) figure keep(file)
customV(numlist) optsettings]
```

- “ **depvar** ”为被解释变量（outcome variable）。
- **predictorvars**为预测变量，即选择的协变量。
- 必选项“ **trunit(#)** ”用于指定处理地区。
- 必选项“ **trperiod(#)** ”用于指定政策干预开始的时期。
- 选择项“ **counit(numlist)** ”用于指定潜在的控制地区，即对照组，默认为数据集中的除处理地区以外的所有地区。
- 选择项“ **xperiod(numlist)** ”用于指定将预测变量（predictors）进行平均的期间，默认为政策干预开始之前的所有时期。
- 选择项“ **mspeperiod()** ”用于指定最小化均方预测误差（MSPE）的时期，默认为政策干预开始之前的所有时期。
- 选择项“ **figure** ”表示将处理地区与合成控制的结果变量画时间趋势图，而选择项“resultsperiod()”用于指定此图的时间范围（默认为整个样本期间）。
- 选择项“ **nested** ”表示使用嵌套的数值方法寻找最优的合成控制（推荐使用此选项），这比默认方法更费时间，但可能更精确。在使用选择项“nested”时，如果再加上选择项“ **allopt** ”（即“ **nested allopt** ”），则比单独使用“nested”还要费时间，但精确度可能更高。
- 选择项“ **keep(filename)** ”将估计结果（比如，合成控制的权重、结果变量）存为另一Stata数据集（filename.dta），以便进行后续计算。

回归合成控制法及扩展

回归合成方法

Hsiso et al.(2012)提出了“回归合成方法”。 其基本思想是：利用截面个体之间的相关性估计干预组个体事后的反事实结果，他们将这种相关性归因于驱动截面个体的公共因子。与合成控制法的区别是，合成控制法要求权重为非负数，不允许外推，而回归合成控制允许权重为负数，允许常数项的存在以修正合成控制组何干预组之间的差异。

基本设定

假设有N+1个个体，第1个个体在t>T_0期受到政策干扰，其他N个个体为潜在的控制组，没有受到政策干预。用D_{it}来表示个体i在t期的干预状态；用Y_{it}表示观测结果。我们关心的是干预组个体1在政策干预之后的政策效应，即

$$\tau_{1t} = Y_{1it} - Y_{0it} = Y_{it} - Y_{0it}, t = T_0 + 1, \cdots, T \tag{1-1}$$

假设所有个体的基线潜在结果服从下列共同因子模型：

$$Y_{0it} = \mu_i + b_i' f_t + \varepsilon_{it} \quad i = 1, \cdots, N+1, \quad t = 1, \cdots, T \tag{1-2}$$

其中μ_i为个体固定效应，f_t 为K×1维的未观测的时变共同因子，b_i为不随时间变化但可能随个体变化的常数，ε_{it}为误差项，满足$E(\varepsilon_{it} = 0)$。这个模型假设个体结果是由两部分构成的，影响所有个体结果的时变共同因子f_t和个体固定效应μ_i及个体扰动项$\varepsilon_i t$。

将模型（1-2）写成矩阵形式：

$$Y_{0t} = \mu + Bf + \varepsilon_t \tag{1-3}$$

其中，$Y_{0t} = (Y_{0t}, \cdots, Y_{0N+1t})', \varepsilon = (\varepsilon_{1t}, \cdots, \varepsilon_{N+1t})', B = (b_1, \cdots, b_{N+1})'$是$(N+1) \times K$的共同因子系数矩阵，引入下列假设：

假设1 对于所有个体i，有$||b_i|| = c < \infty$。

假设2 ε_t~I(0), 并且$E[\varepsilon_t] = 0, E[\varepsilon_t \varepsilon_t'] = V$,V为对角常数矩阵。

假设3 $E[\varepsilon_t f_t'] = 0$。

假设4 Rank(B)=K。

假设5 $E[\varepsilon_{js} | D_{it}] = 0, j \neq i$。

在假设1-5下，如果能够识别μ_1、b_1和f_t,则可以利用$\hat{Y}_{01t} = \mu_1 + b_1' f_t, t = T_0 + 1, \cdots, T$来预测政策实施后干预组的反事实结果$Y_{01t}$。但是，这里个体固定效应、共同因子都是不可观测的，因而无法直接估计干预组个体事后的政策效应。根据上文的假设，所有个体都受到时变共同因子的影响，事后共同因子的影响将体现在控制组的观测结果中，因而，可以从事后控制组的观测结果中反推出时变因子，并利用事前干预组观测结果与控制组事前观测结果之间受共同因子影响而造成的相关关系，估计出事后如果干预组个体没有受到政策干预的反事实结果。实际上，回归合成方法直接利用控制组观测结果作为干预组个体观测结果的预测变量，其基本逻辑是由于所有个体均受到共同因子的影响，从而造成截面个体之间的相关性，根据事前截面个体之间的相关性，预测干预组和控制组个体之间的关系，如果事后干预组个体没有受到政策干预，那么，各截面个体将维持相似的依赖关系，从而利用这种事前的依赖关系，并利用事后控制组的观测结果估计出如果干预组个体没有接受干预的反事实结果。

具体地，令$a = (1, -\gamma')$为B零空间中的一个向量，即a'B=0,其中$\gamma = (\gamma_2, \cdots, \gamma_{N+1})'$,模型(1-3)两边同乘以$a'$,则可消去共同因子$f_t$,从而得：

$$Y_{01t} = \gamma_1 + \gamma' Y_t + \varepsilon_1 t^* \tag{1-4}$$

其中，$\gamma_1 = \alpha' \mu$、$\varepsilon_1 t^* = \alpha' \varepsilon_t = \varepsilon_1 t - \gamma' \varepsilon_t, \varepsilon_t = (\varepsilon_{2t}, \cdots, \varepsilon_{N+1t})'$

由于ε_{1t}^*依赖于所有的误差项$\varepsilon_{jt}, j = 1, \cdots, N+1$,显然，$\varepsilon_{1t}^*$与$Y_t$相关。为此，可以将$\varepsilon_{1t}^*$分解为两部分，$\varepsilon_1 t^* = E[\varepsilon_1 t^* | Y_t] + v_t$,其中$v_{1t} = \epsilon_{1t}^* - E[\epsilon_{1t}^* | Y_t]$,因而，$E[v_{1t} | Y_t] = 0$。（1-4）可以写成：

$$Y_{01t} = \gamma_1 + \gamma' Y_t + E\left[\epsilon_{1t}^* \mid Y_t\right] + v_{1t} \tag{1-5}$$

假设6 $E(\varepsilon_1 t | Y_t) = \delta_1 + \delta' Y_t$

由假设6，（1-5）可以写成

$$\hat{Y}_{01t} = \beta_1 + \hat{\beta}' Y_t + v_{1t} \tag{1-6}$$

其中$\beta_1 = \gamma_1 + \delta_1$、$\beta = \gamma + \delta, E[v_{1t}|Y_t] = 0$ Hsiso et al.(2012)证明，如果$T_0 \to \infty$,OLS估计量将是参数的一致估计。可以利用事前数据估计模型（1-6）的参数，得到预测模型：

$$\hat{Y_{01t}} = \beta_1 + \hat{\beta}' Y_t \tag{1-7}$$

因而，相应的政策效应为：

$$\tau_{1t} = Y_{1t} - \hat{Y_{01}t}, \quad t = T_0 + 1, \cdots, T \tag{1-8}$$

反事实估计Y_{01t}，$t = T_0 + 1, \cdots, T$依赖于个体固定效应μ_1,共同因子f_t、个体对共同因子的反应b_1,以及个体特质因素ε_{1t}。然而，预测模型（1-8）并不需要这些信息，原因在于共同因子的信息已经蕴含在控制组观测结果Y_t之中，从而可以利用控制组信息Y_t代替共同因子来实现对政策效应的估计。

回归合成方法与合成控制法有很多相似之处，事实上，合成控制法也可以纳入回归合成方法的框架，在合成控制组模型中，令$\mu_i = 0$,$b_i' = (1, Z_i', \mu_i')$，$f_t' = (\delta_t, \theta_t', \lambda_t')$,则合成控制组也可以写成模型（1-2）的形式，但为了保证（1-4）的成立，需要$a'B = 0$,这一条件要求$\sum_{i=1}^{N+1} \alpha_i = 0$, $\sum_{i=1}^{N+1} \alpha_i Z_i' = 0'$，$\sum_{i=1}^{N+1} \alpha_i \mu_i' = 0$。合成控制法通过$\hat{Y}_{01t} = \sum_{i=2}^{N+1} \alpha_i Y_{it}$来构造干预组的反事实结果，不过它要求$\alpha_i \geq 0$并且$\sum_{j=2}^{N+1} \alpha_i = 1$,回归合成方法中不需要施加这一限制，从而允许外推。

回归合成控制组的选择

当事前时期远远大于控制组个体数时，即$T_0 \gg N$，$N/T_0 \to \infty$，可以利用所有控制组个体作为潜在的合成控制在应用中，T_0和N/T_0往往是有限的，进入模型的控制组个体越多，模型的自由度损失越多，估计精度会越低。因而在模型拟合时面临着拟合效果和估计精度之间的权衡。

Hsiao et al.(2012)提出了一种两步法来解决这一问题。首先，依次选择1，2，…，N个控制组个体进入模型，利用拟合优度R^2或似然值来选择模型。对于有m个控制组进入模型时，共需要估计 $C_N^m = m!/[N!(N-m)!]$个模型，利用R^2或似然值，从中选择拟合最好的一个模型，记为$M(N)^*$。依次选择下来，共得到N个模型即$M(1)^*, \cdots, M(N)^*$。然后，利用模型选择标准，选择最优的模型。Hsiao et al.(2012)建议利用AIC或AICC标准进行选择，两个模型选择标准定义如下：

$$AIC(p) = T_0 \ln\left(\frac{e_0 e_0}{T_0}\right) + 2(p+2) \tag{1-9}$$

$$\text{AICC}(p) = \text{AIC}(p) + \frac{2(p+2)(p+3)}{T_0 - (p+1) - 2} \tag{1-10}$$

其中，p为模型中包含的控制组个数，e_0为相应模型的OLS回归残差向量。

案例：对香港经济的影响

Hsiao et al.(2012)考察了1997年香港回归以及2003年在香港签署的《内地与香港关于建立更紧密经贸关系的安排》（CEPA)对香港经济的影响。作者发现，香港回归对香港经济没有明显影响，但CEPA对香港经济具有显著的正向影响。这里主要考察CEPA对香港经济的影响。主要的结果变量是经济增长率，CEPA于2004年正式实施，因而，自2004年起香港经济可能受到CEPA安排的影响。为了估计CEPA对香港经济的影响，需要估计如果没有CEPA安排，香港经济会是什么状态，即估计香港经济增长的反事实结果。

因为香港是一个城市，从经济规模上而言，相对较小，香港应该不会对其他经济体有很大的影响，为此，选择澳大利亚、奥地利、加拿大、中国大陆、丹麦、芬兰、法国、德国、印度尼西亚、意大利、日本、韩国、马来西亚、墨西哥、荷兰、新西兰、挪威、菲律宾、新加坡、瑞士、台湾地区、泰国、英国和美国等24个国家或地区作为潜在的控制组，使用1993年第一季度到2008年第1季度的季度数据进行估计。

首先，利用上文提出的两步法以及2004年之前的数据，估计模型（1-6),对于同样数目的地区进入控制组时，利用R^2来选择最优的模型$M(m)^*$，$m=1,\cdots,24$,然后再利用模型进行选择校准AICC进行比较最优的模型。通过这一方法，最终合成的控制组包括奥地利、意大利、韩国、墨西哥、挪威和新加坡等6个国家，具体权重见表1。

表1 合成控制组的权重：AICC,1993-2003:Q4

	β	std. err.	t-stat
Constant	−0.0019	0.0037	−0.5240
Austria	−1.0116	0.1682	−6.0128
Italy	−0.3177	0.1591	−1.9971
Korea	0.3447	0.0469	7.3506
Mexico	0.3129	0.0510	6.1335
Norway	0.3222	0.0358	5.9912
Singapore	0.1845	0.0546	3.3812

注：复制 Hsiao et al.(2012)的表 XX。

合成效果和政策效应见图1，图中实线为香港的实际经济增长率，虚线为合成香港的经济增长率，可以看出，2004年CEPA安排正式实施之前，两者几乎完全重合，这说明合成控制组基本上与香港的运行模型相一致。2004年起，两条曲线产生明显的差异，这说明CEPA的政策影响是显著的。

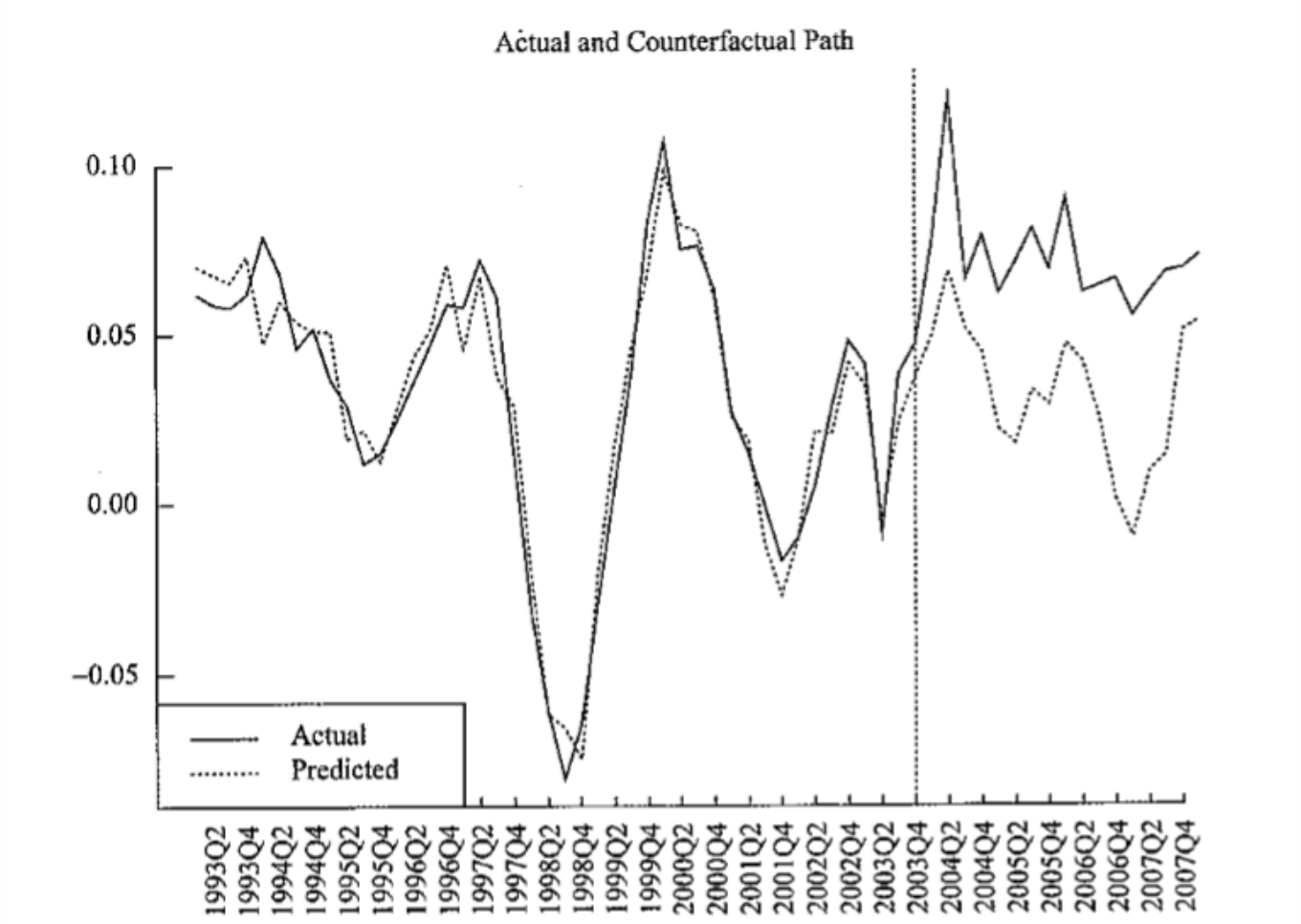

图1 香港实际的经济增长率和预测的经济增长率

基于交互固定效应的扩展合成控制

双重差分模型是最常用的因果推断方法之一。但是DID有一个关键的假设是平行趋势假设，要求处理组和控制组的潜在结果在政策实施前后遵循相同的趋势。这一假设本质上不能通过数据直接验证，因为政策实施后处理效应和潜在结果混杂在一起了。但是实证研究中通常认为，假如政策实施前处理组和控制组因变量满足平行趋势，则这一假设基本算作成立。但是，许多面板数据在政策实施前因变量都不遵循平行趋势，平行趋势无法满足的主要原因是存在了随时间变化的不可观测因素。

目前，文献中主要有两种思路来解决这一问题。第一种思路是匹配法，这主要是平衡了政策实施前可观测变量的筛选效应，这可能有助于控制处理组和控制组中未观测的时变混杂因素对因变量的影响，但是本质上并不能保证平行趋势假设成立，因为平行趋势假设是依不可观测变量的筛选效应的一种特定形式。Abadie et al.(2010)提出的合成控制法更进一步，利用处理组和控制组的截面相关性，将控制组的结果和预处理协变量取非负权重，以最小化政策实施前的均方预测误差（MSPE)。这两种方法的一个局限性是处理组只有一个，而且提供的不确定估计不容易解释。Abadie et al.(2010)通过安慰剂检验来说明处理组的平均效应（ATT)的显著性，但并未提出不确定估计的置信区间。

第二种思路是直接对未观测的时变混杂因素进行建模。一个常用的建模方法是在双向固定效应模型的基础上，加入线性时间趋势项及其平方项。这样做实质上假设了条件于双向固定效应和假定的时间趋势后，是否接受处理与潜在结果变量均值独立。但是控制了这些假定的时间趋势之后，模型通过会消耗很多自由度，并且当未观测的时变混杂因素不是模型所假定的线性或非线性趋势项形式时，这样的设定就不能解决问题。另一种建模方法是在Bai(2009)提出的交互固定效应模型（IFE)的基础上，使用半参的方法估计ATT。IFE基于因子模型，对线性模型估计后的残差进

行因子分析，提取出最具影响力的r个共同因子，r是基于数据进行交叉验证确定的。

Xu et al.(2017)提出了扩展的合成控制法（Generalized synthetic control method，以下简称GSC） 将SCM和IFE统一到一个框架下。这一模型设定允许政策处理效应随个体和时间变化，同时假定处理组和控制组被r个共同因子影响，但因子载荷（Factor loading）可随个体变化。

假设1 函数形式

$$Y_{it} = \delta_{it} D_{it} + x_{it}'\beta + \lambda' f_t + \varepsilon_{it} \tag{2-1}$$

其中， D_{it}表示个体 i 是否进入处置组，δ_{it}是在 t 期个体 i 的异质性处置效应，x_{it}是协变量（ K×1)维向量，$\beta = [\beta_1, cdots, \beta_k]'$是（ K×1)维的未知参数，$f_t = [f_{1t}, \cdots, f_{rt}]'$是（ r×1 ）维未知公共因子，$\lambda_i = [\lambda_{i1}, \cdots, \lambda_{ir}]'$是未知因子载（ r×1 ）维的向量，$\varepsilon_{it}$表示在 t 期对个体 i 未观测到的特殊冲击，并且满足均值为 0 的假设。

关于公共因子的分解如下：

$$\lambda_j' f_t = \lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} + \cdots + \lambda_{ir} f_{rt} \tag{2-2}$$

这种形式包含了广泛的未观测的异质性，传统的个体和时间的双向固定效应模型是其中的一个特例。这种形式也包含了线性和平方时间趋势的固定效应，以及可能增加的其他固定效应或趋势效应。通常，只要未被观察到的随机因素能被分解为矩阵的乘法形式，如 $U_{it} = \alpha_i \times b_t$，那么这些因素能被 $\lambda' f_t$所吸收。

假设2 严格外生的

$$\varepsilon_{it} \coprod [D_{js}, x_{js}, \lambda_j, f_s \quad \forall i, j, t, s \tag{2-3}$$

这意味着任何个体的误差项在任何时期都是独立于处置状态、可观测协变量和不可观测的横截面和时间异质性。此外，还有其他假设如：假设3 误差项的弱序列相关性；假设4 规律性条件；假设5 误差项在截面上是独立的，同方差的。

假定数据生成过程（DGP)为

$$Y_i = D_i \delta_i + X_i \beta + F \lambda_i + \varepsilon_i \tag{2-4}$$

其中，$Y_i = [Y_{i1}, Y_{i2}, \cdots, Y_{iT}]'$,$D_i = [D_{i1}, D_{i2}, \cdots, D_{iT}]$,$\delta_i = [\delta_{i1}, \delta_{i2}, \cdots, \delta_{it}]'$,$\varepsilon_i = [\varepsilon_{it}, \varepsilon_{i2}, \cdots, \varepsilon_{iT}]'$是$(K \times 1)$维矩阵，$X_i = [x_{i1}, x_{i2}, \cdots, x_{iT}]'$是$(T \times K)$维矩阵，$F = [f_1, f_2, \cdots, f_T]'$。

使用半参数方法估计ATT

GSC 基于 Bai 2009 ）提出的因子增广模型（ Factor augmented model ）来进行样本外预测 （Out of Sample Prediction），构造出每一个处理组的反事实对照。

具体分为如下三个步骤：

1.仅使用控制组进行有约束的 OLS 估计，得到外生协变量估计系数，共同因子和控制组的因子载荷 λ

$$\left(\hat{\beta}, \hat{F}, \hat{\Lambda}_{co}\right) = \text{argmin}_{\hat{\beta}, \hat{F}, \Lambda_\infty} \sum_{i \in C} \left(Y_i - X_i \tilde{\beta} - \tilde{F} \tilde{\lambda}_i\right) \left(Y_i - X_i \tilde{\beta} - \tilde{F} \tilde{\lambda}_i\right) \tag{2-5}$$

$$\frac{\tilde{F}'\tilde{F}}{T} = I_r \& \tilde{\Lambda}'_{co}\tilde{\Lambda}_{co} = \text{ diagonal} \tag{2-6}$$

2.针对每一个控制个体，通过最小化干预期前的预测结果的均方误差，以此估计每个处理个体的因子载荷 λ

$$\begin{aligned}\hat{\lambda}_i &= \underset{\tilde{\lambda}_i}{\text{argmin}} \left(Y_i^0 - X_i^0\hat{\beta} - \hat{F}^0\tilde{\lambda}_i\right)' \left(Y_i^0 - X_i^0\hat{\beta} - \hat{F}^0\tilde{\lambda}_i\right) \\ &= \left(\hat{F}^0\hat{F}^0\right)^{-1}\hat{F}^0\left(Y_i^0 - X_i^0\hat{\beta}\right), \quad i \in \mathcal{T},\end{aligned} \tag{2-7}$$

其中，$\hat{\beta}$和$\hat{F}^0$是来自于第一阶段的估计和上标 0 表明是进入处置状态之前。

3.基于以上估计系数，计算处理组接受处理后的反事实对照组

$$\hat{Y}_{it}(0) = x'_{it}\hat{\beta} + \hat{\lambda}'_i\hat{f}_t \tag{2-8}$$

最后，可以得到ATT为

$$\widehat{ATT}_t = \left(\frac{1}{N_{tr}}\right)\sum_{i\in T}\left[Y_{it}(1) - \hat{Y}_{it}(0)\right] \tag{2-9}$$

需要注意的是，只有当政策实施前时期和控制组个数足够多的时候，才能得到一致估计，否则就是有偏的。

作者的其他工作?

后续使用交叉验证的方法来确定共同因子的个数 r ，以及通过自助（ bootstrap) 的方法得到 ATT 的不确定估计。通过蒙特卡罗模拟的方法验证了 GSC 估计量良好的统计性质。而且在附录部分，作者还比较了 GSC 、 DID 、 IFE 和 SCM 的估计效果，发现 GSC 比这些方法不差，甚至更好。值得说明的是，这都是基于模型的正确设定才有的结果。

评价GSC 的优缺点

GSC的优点主要包括：1.将 SCM 推广到多个处理组和多期接受处理的一般情形。2.基于自助（ Bootstrap) 构建了 ATT 的不确定估计，如标准差和置信区间，还提高了模型正确设定下的估计效率。3.运用交叉验证的方法自动选择最优的共同因子个数，降低过度拟合的风险。

GSC的局限性主要包括：1.GSC 相较于固定效应模型，需要更多政策实施前的数据。否则可能出现偶然参数问题和有偏的ATT 估计。2.GSC 相较于 SCM, 模型假设扮演更重要的角色。GSC直接根据模型设定外推，有可能导致错误的结论，因此在使用 GSC 之前，一定要对原始数据进行画图描述、反事实拟合和诊断检验，以帮助我们认识数据的特征。

增广合成控制（ASCM)

合成控制法是对面板数据做政策评估时的一种很流行的方法，它要求适度的控制组，以及提供政策干预前有多期样本。在实际使用中，合成控制法的关键假设是干预前处置组和控制组的拟合效果很好，也就是说干预前处置组和合成控制组的观测结果近似相等。（Aadie et al.,2015），但在实际使用中很难实现。在政策干预前，处置组和合成控制组偏差很大的情况下，就不建议研究者使用合成控制法了，许多研究者很可能就会使用线性回归。但作者认为通

过对负权重和外推法的使用，就可以将干预前处置组和对照组拟合的更好，这也就是增广合成控制（ASCM）。

作者认为岭 ASCM 可以通过使用外推法允许负权重，令合成控制组和处置组的偏差减小。具体而言，当处置组位于控制组的凸包之外时，岭 ASCM可以通过允许负权重和外推法至凸包之外。

ASCM关于$Y_{1T}(0)$的估计量：

$$\begin{aligned}\hat{Y}_{1T}^{\text{aug}}(0) &= \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} Y_{iT} + \left(\hat{m}_{1T} - \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \hat{m}_{iT} \right) \\ &= \hat{m}_{1T} + \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \left(Y_{iT} - \hat{m}_{iT} \right)\end{aligned} \tag{3-1}$$

其中，权重$\hat{\gamma}_i^{scm}$是SCM的权重，标准SCM是上式的一个特例，$\hat{m}_iT$是一个截距项。

特殊形式

虽然上式的形式是通用的，但是估计量$\hat{m}$的选择对于理解这个模型的属性和实际功能非常重要。我们主要简单的介绍两种特殊情况：（1）当$\hat{m}$是线性的预处理结果时；（2）当$\hat{m}$在比较在控制组个体的干预后结果是线性的情况。

首先，考虑在预处理结果的估计量是线性的，即$\hat{m}(X) = \hat{\eta} \cdot X$.那么模型（3-1）的估计量就是

$$\hat{Y}_1 T^{awg}(0) = \sum_{W_i=0} \gamma_i^{\hat{cm}} Y_{iT} + \sum_{t=1}^{T_0} \hat{\eta}_t \left(X_{1t} - \sum_{W_i=0} \gamma_i^{\hat{scm}} X_{it} \right) \tag{3-2}$$

干预后拟合效果更好的预处理时期将有更大的（绝对的）回归系数，因此这些时期的不平衡将导致更大的调整。因此，即使我们在任何特定的预处理时间段（通过选择V_x）中没有预先确定平衡的优先级，模型（3-2）也会针对经验，将干预后拟合效果更好的时期进行调整。

其次，考虑估计量$\hat{m}$是结合了控制组干预后结果的线性组合，即$\hat{m}(X) = \sum_{W_i=0} \alpha_i \hat{(X)} Y_{iT}$,对于其中的权重$\hat{\alpha} : \mathbb{R}^{T_O} \to \mathbb{R}^{N_O}$.比如k近邻匹配和核加权以及其他“垂直”回归方法。因此增强估计量（3-1）是一个调整SCM权重的估计量。得到：

$$Y_{1T}^{\hat{aug}}(0) = \sum_{W_i=0} \left(\hat{\gamma}_i^{scm} + \hat{\gamma}_i^{adj} \right) Y_{iT} \tag{3-3}$$

其中，$\hat{\gamma}^{adj} \equiv \hat{\alpha}_i(X_i) - \sum_{W_j=0} \hat{\gamma_j}^{scm} \hat{\alpha}_i(X_j)$。在这里，个体i的调整项$\hat{\gamma_i}^{adj}$是滞后结果的个体i特殊变换的不平衡，它依赖于加权函数$\alpha(\cdot)$。当$\hat{\gamma}^{scm}$被限制为单纯形时，$\hat{\gamma}^{adj}$能表明总权重可以是负数的。

有许多可以考虑的特例，一个是等线性的线性滞后结果模型，$\hat{\eta} = 1/T_0$,它估计固定效应结果模型为$\hat{m}(X_i) = \bar{X}_i$。相应的处理结果估计对所有预处理时间段的不平衡进行相等的调整，并产生一个加权差异—差异（Difference-in-Difference)估计量。

$$\begin{aligned}\hat{\tau}^{de} &= \left(Y_{1T} - \bar{X}_1\right) - \left(\sum_{W_i=0} \hat{\gamma}_i \left(Y_{iT} - \bar{X}_i\right)\right. \\ &= \frac{1}{T_0} \sum_{t=1}^{T_0} \left[\left(Y_{1T} - X_{1T}\right) - \left(\sum_{W_i=0} \hat{\gamma}_i \left(Y_{iT} - X_{it}\right)\right.\right.\end{aligned} \tag{3-4}$$

这一模型平衡了结果$X_{it} - \bar{X}_i$而不是原始结果X_{it}。

作者其他工作

在有限样本内进行了仿真模拟。将岭回归（Ridge Regression）作为惩罚项引入ASCM，以及将ASCM与辅助的协变量结合应用而不是预处理结果。作者将辅助协变量和滞后结果平行地包含在SCM和结构模型中。作者将这一理论应用到实证中，用来研究堪萨斯州大幅减税对经济产出的影响，结果表明负向效应。

扩展应用

Matthew et al.（2020）发表在《Environmental and Resource Economics》的研究，题目是《The impact of the Wuhan Covid-19 Lockdown on Air Pollution and Health: A machine Learning and Augmented Synthetic Control Approach》

这篇论文将ASCM与机器学习相结合，研究了武汉新冠疫情导致的“封城”对空气污染和健康的影响。

参考文献

[1]Abadie, Alberto, and Javier Gardeazabal. 2003. “The Economic Costs of Conflict: A Case Study of the Basque Country.” *American Economic Review* 93 (1): 113–32.

[2]Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” *Journal* of the American Statistical Association 105 (490): 493–505.

[3]Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. “Comparative Politics and the Synthetic Control Method.” American Journal of Political Science 59 (2): 495–510.

[4]刘友金,曾小明.房产税对产业转移的影响:来自重庆和上海的经验证据[J].中国工业经济,2018,(11):98-116.

[5] Xu Y . Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models[J]. Political Analysis, 2017, 25.

[6] Hsiao, C. , Ching, H. S. , & Wan, S. K. . (2012). A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. Journal of Applied Econometrics, 27(5), 705-740.

[7] Ke X , Chen H , Hong Y , et al. Do China's high-speed-rail projects promote local economy?—New evidence from a panel data approach[J]. China Economic Review, 2017, 44:203-226.

[8] Cole M A , Elliott R , Liu B . The Impact of the Wuhan Covid-19 Lockdown on Air Pollution and Health: A Machine Learning and Augmented Synthetic Control Approach[J]. Environmental & Resource Economics, 2020, 76.

[9] Ben-Michael E , Feller A , Rothstein J . The Augmented Synthetic Control Method[J]. Papers, 2020.

[10] Bai J. Panel data models with interactive fixed effects[J]. Econometrica, 2009, 77(4): 1229-1279.