doi: 10.12011/SETP2020-0037 中图法分类号: F224; O212 文献标志码: A

带有变量选择的协变量平衡倾向得分的估计: 基于 GMM-LASSO 方法

范菊逸[1], 詹铭峰[2], 蔡宗武[3], 方 颖[2,4], 林 明[2,4]

(1. 厦门大学 经济学院, 厦门 361005; 2. 厦门大学 王亚南经济研究院, 厦门 361005;
3. 堪萨斯大学 经济系, 堪萨斯 66045; 4. 福建省统计科学重点实验室 (厦门大学), 厦门 361005)

摘 要 本文基于 Imai 和 Ratkovic (2014) 协变量平衡倾向得分的估计方法, 提出了带协变量平衡的 GMM-LASSO 估计方法. 该方法既利用了协变量平衡的性质, 同时又解决了如何基于数据来选取协变量的问题. 理论上, 文章证明了该估计方法是相合的. 模拟显示, 在满足一定的稀疏性的条件下, 该方法可以显著地降低平均处理效应估计的绝对误差的中位数. 最后, 该方法被应用于研究 2000 年代初期意大利托斯卡纳地区的劳务派遣机制是否有助于工人寻找一份稳定的工作.

关键词 协变量平衡; GMM-LASSO; Logistic 回归; 倾向得分; 处理效应

Covariate balancing in propensity score estimation with variable selection: Based on GMM-LASSO approach

FAN Juyi[1], ZHAN Mingfeng[2], CAI Zongwu[3], FANG Ying[2,4], LIN Ming[2,4]

(1. School of Economics, Xiamen University, Xiamen 361005, China; 2. The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen 361005, China; 3. Department of Economics, University of Kansas, KS 66045, USA; 4. Fujian Key Laboratory of Statistical Sciences, Xiamen University, Xiamen 361005, China)

Abstract Based on the covariate balancing propensity score method introduced by Imai and Ratkovic (2014), this paper proposes a new method to combine the GMM-LASSO type estimation method with covariate balancing approach. The proposed method not only utilizes the property of covariate balancing, but also solves the problem of how to select covariates based on data. Also, it is shown that the proposed estimator is consistent and simulations show that under the condition of sparsity, the proposed method indeed can significantly reduce the median of the absolute errors of the average treatment effect. Finally, the proposed method is applied to the data from the Tuscany region of Italy in the early 2000s to study whether temporary work agency mechanism helps workers to find a stable job in the future.

Keywords covariate balancing; GMM-LASSO; logistic regression; propensity score; treatment effect

1 引言

政策评估是经济学家和政策制定者迫切关心的问题, 例如: 评估培训项目对收入的影响, 评估最低工资对就业的影响等, 准确估计这些处理效应为合理实施某项政策或项目提供了重要的科学依据. 在处理效应的

收稿日期: 2020-01-10

作者简介: 范菊逸 (1979–), 女, 汉, 湖北浠水人, 博士研究生, 研究方向: 因果推断, 机器学习, E-mail: fanjuyi@stu.xmu.edu.cn; 通信作者: 詹铭峰 (1988–), 男, 汉, 福建闽清人, 博士研究生, 研究方向: 计量经济, 数理统计, E-mail: zhanmf@stu.xmu.edu.cn; 蔡宗武 (1960–), 男, 汉, 福建莆田人, 博士, 教授, 研究方向: 计量经济, 金融计量, E-mail: caiz@ku.edu; 方颖 (1973–), 男, 汉, 上海人, 博士, 教授, 研究方向: 计量经济, E-mail: yifst1@xmu.edu.cn; 林明 (1978–), 男, 汉, 福建连江人, 博士, 教授, 研究方向: 贝叶斯统计, 蒙特卡罗方法, E-mail: linming50@xmu.edu.cn.

基金项目: 国家自然科学基金 (71631004, 72033008); 国家杰出青年科学基金 (71625001)
Foundation item: National Natural Science Foundation of China (71631004, 72033008); National Natural Science Foundation for Distinguished Young Scholars (71625001)
中文引用格式: 范菊逸, 詹铭峰, 蔡宗武, 等. 带有变量选择的协变量平衡倾向得分的估计: 基于 GMM-LASSO 方法 [J]. 系统工程理论与实践, 2021, 41(10): 2631–2639.
英文引用格式: Fan J Y, Zhan M F, Cai Z W, et al. Covariate balancing in propensity score estimation with variable selection: Based on GMM-LASSO approach[J]. Systems Engineering — Theory & Practice, 2021, 41(10): 2631–2639.

估计中, 尽管随机对照试验方法是确定因果效应的黄金标准, 但是在实际应用中进行随机试验很多时候是困难的, 甚至是不现实的. 在这种情形下, 研究者不得不依赖于观测数据 (observational data), 即接受处理的个体只能被观测而不是由研究者预先指定. 从 20 世纪 70 年代初期开始, Rubin[1−4] 在一系列论文中提出了现在占有主导地位的方法来分析处理效应, Rubin 对处理效应问题的描述被 Holland[5] 标记为 Rubin 因果模型 (RCM).

在观测性因果推断中一个典型的问题是混淆 (confounding), 即个体特征可能与处理分配 (treatment assignment) 和结果变量相关. 此时简单地比较处理组和对照组的样本均值可能会导致严重的估计偏差. 为了解决这一问题, Rubin[6] 引入了无混淆性 (unconfoundedness) 假设, 该假设意味着在给定混淆因素的条件下, 个体进入处理组和对照组的过程可以看成是随机的. 在无混淆假设下, 主要有两大类方法来估计平均处理效应: 一种是基于结果变量和协变量之间的关系而建立的结果回归 (outcome regression) 方法; 另一种是基于处理变量和协变量之间的关系, 即倾向得分 (propensity score), 而建立的逆概率加权 (inverse probability weighting) 方法. 如果采取参数模型对上述两种方法进行建模, 则模型存在误设的可能, 这种误设将会导致估计上的偏差. Robins 等 [7] 提出的增广逆概率加权方法在一定意义下可以减轻模型误设导致的偏差, 基于这种方法的估计量也被称为双重稳健估计量 [8]. 双重稳健估计量同时结合了结果回归模型和倾向得分模型, 使得只要在这两个模型之中有一个模型是正确设定的, 则该估计量仍是平均处理效应的渐近无偏估计量. 这种估计量给了研究者一次犯错的机会, 在实际应用中备受青睐.

除了上述提到的逆概率加权估计量和双重稳健估计量, 基于倾向得分的方法还有匹配 (matching) 和子分类 (subclassification) 等. 可见, 倾向得分在处理效应的估计中扮演着十分重要的角色. 倾向得分的概念首先由 Rosenbaum 和 Rubin[9] 提出, 它被定义为在给定协变量的条件下个体进入处理组的概率. 倾向得分的一大优点是可以将多维的协变量简化为一个标量. 尽管上述估计平均处理效应的方法很受研究者欢迎, 但是在实际应用中倾向得分往往是未知的, 研究者需要根据观测的数据对其进行估计. Kang 和 Schafer[10] 发现倾向得分模型的轻微误设可能导致处理效应估计的显著偏差. 这一挑战突出了倾向得分的矛盾性, 即倾向得分旨在减少协变量的维度, 但它的估计又需要对高维协变量进行建模. 倾向得分的估计可以采用参数模型或非参数模型. 常用的参数模型是线性 Logistic 回归, 但此时的倾向得分模型存在误设的可能. 非参数模型虽然可以避免模型误设, 但在高维的时候会面临 “维度灾难” 的问题. 因此在协变量维度较高的情况下, 研究者通常被迫采用参数模型来估计倾向得分. 一种流行的方法是先假设一个线性 Logistic 模型, 并通过极大似然法估计未知参数, 然后检查所得到的倾向得分估计值是否平衡协变量的特定矩. 如果不平衡, 则重新设定倾向得分模型 (包含更高阶和交互项), 并重复该过程, 直到实现协变量平衡 [11]. 尽管这种方法能够改善倾向得分模型误设的问题, 但仍可能会导致参数的有偏估计 [12].

鉴于这些实际问题, 近年来研究者提出了一类新的方法 —— 协变量平衡 (covariate balancing) 法. 该方法旨在直接平衡处理组和对照组的协变量, 本质上是一种类随机化的方法. 特别地, Hainmueller[13] 提出了熵平衡 (entropy balancing) 法, 该方法是一种预处理程序, 在估计处理效应之前, 为每个个体分配标量权重, 使得加权后的组满足平衡约束, 再把构造的平衡样本用于处理效应的估计. 此外, Imai 和 Ratkovic[14] 引入了协变量平衡倾向得分 (covariate balancing propensity score, CBPS) 的方法, 该方法利用了倾向得分的协变量平衡性质, 并在熟悉的广义矩估计 (GMM) 的框架内估计倾向得分. 该方法在估计倾向得分的同时实现了协变量的平衡, 这有别于以往先估计倾向得分, 再检查协变量是否平衡的作法. 倾向得分具有协变量平衡的性质, 对这一性质的充分利用使得这些方法往往能够改善估计量的有限样本性质. 需要注意的是, Imai 和 Ratkovic[14] 的协变量平衡倾向得分方法只平衡了协变量的有限矩. 而倾向得分协变量平衡的性质可以平衡协变量的任意可测函数. 为了充分利用倾向得分协变量平衡的性质, Sant'Anna 等 [15] 在协变量平衡倾向得分的基础上, 提出了综合倾向得分 (integrated propensity score) 方法, 该方法使得倾向得分可以平衡协变量的分布函数, 即可以平衡协变量的任意阶矩. 蒙特卡罗模拟的结果显示, Sant'Anna 等 [15] 提出的综合倾向得分方法的表现, 在倾向得分模型轻微误设的情况下, 要优于 Imai 和 Ratkovic[14] 的协变量平衡倾向得分方法.

尽管 Imai 和 Ratkovic[14] 提供了一些经验数据表明协变量平衡倾向得分方法可以显著改善逆概率加权

和匹配估计量的结果. 但是协变量平衡倾向得分方法的成功应用, 需要研究者指定一系列完整的混淆因素. 在实际应用中, 协变量的选取并不是显而易见的. 特别是在大数据环境下, 开发一种数据驱动的方法恰当的选取混淆因素, 显得尤为重要. 为了解决这一问题, Ning 等 [16] 提出了高维协变量平衡倾向得分估计的方法, 他们分别对倾向得分模型和结果回归模型建立带惩罚的广义线性模型进行双重选择, 然后嵌入协变量平衡条件得到了倾向得分的估计量.

本文中, 我们提出了一种有别于 Ning 等 [16] 的方法来考虑带变量选择的协变量平衡倾向得分的估计问题. 基于 Imai 和 Ratkovic[14] 的协变量平衡倾向得分方法, 本文提出了 GMM-LASSO 的方法来估计倾向得分. GMM-LASSO 的方法是在广义矩估计的框架内考虑变量选择的问题, 这方面的文献可以参见 Caner[17]. 我们的方法在估计倾向得分的时候, 既利用了倾向得分的协变量平衡性质, 同时又解决了如何基于数据来选取混淆因素的问题, 这是对 Imai 和 Ratkovic[14] 的协变量平衡倾向得分方法的推广. 我们提出的带协变量平衡的 GMM-LASSO 估计方法将变量选择引入协变量平衡问题, 这一方法有助于提高协变量平衡法在研究中的实用性. 模拟部分显示, 在满足一定的稀疏性的条件下, 本文的方法可以显著地降低平均处理效应估计的绝对误差的中位数.

2 方法

2.1 背景

设 $\boldsymbol{X}$ 是 p 维协变量, T 是取值为 0 和 1 的二元处理变量, $T=1$ 表示个体进入处理组, $T=0$ 表示个体进入对照组. 令 $Y(1)$ 和 $Y(0)$ 分别是处理组和对照组的潜在结果变量, 则观测到的结果变量为 $Y=TY(1)+(1-T)Y(0)$. 定义倾向得分 $\pi(\boldsymbol{x})=P(T=1|\boldsymbol{X}=\boldsymbol{x})$. 假设 $\{(Y_i,T_i,\boldsymbol{X}_i)\}_{i=1}^n$ 是来自总体 $(Y,T,\boldsymbol{X})$ 的一个独立同分布的样本. 处理效应研究的主要问题是评估处理变量 T 对感兴趣的结果变量 Y 的效应, 实际中常用的处理效应有平均处理效应、分布处理效应、分位数处理效应等, 这里我们关注平均处理效应. 定义总体的平均处理效应 $\Delta=E[Y(1)-Y(0)]$. 注意到, 这里参加项目的同一个体不能同时进入处理组和对照组, 即 $Y(1)$ 和 $Y(0)$ 不能被同时观测到. 这样处理效应实际上是一个带有缺失数据的问题, 所以平均处理效应的估计并不是显而易见的.

为了获得平均处理效应的相合估计量, 研究者通常假设进入处理组和对照组的选择机制完全基于可观测的特征, 并且所有个体都有进入处理组和对照组的可能, 这就是所谓的强可忽略假设 [9]. 具体地, 强可忽略假设指的是: (a) 给定 $\boldsymbol{X}$ 的条件下, $(Y(1),Y(0))$ 和处理变量 T 独立; (b) 存在 $\varepsilon>0$, 使得 $\varepsilon<\pi(\boldsymbol{x})<1-\varepsilon$. Rosenbaum[18] 证明在该假定下, 平均处理效应可以被表示为

$$\Delta=E\left[\frac{TY}{\pi(\boldsymbol{X})}-\frac{(1-T)Y}{1-\pi(\boldsymbol{X})}\right].$$

如果倾向得分 $\pi(\boldsymbol{X})$ 是已知的, 则我们可以通过样本矩估计总体矩的方法获得平均处理效应的相合估计量:

$$\tilde{\Delta}(\pi)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_iY_i}{\pi(\boldsymbol{X}_i)}-\frac{(1-T_i)Y_i}{1-\pi(\boldsymbol{X}_i)}\right).$$

然而, 在观测性研究中, 倾向得分 $\pi(\boldsymbol{X})$ 通常是未知的. 研究者往往采取参数或非参数的方法对倾向得分进行估计. 当协变量 $\boldsymbol{X}$ 的维度较高时, 为了避免 "维数灾难", 研究者一般采用参数化模型进行拟合, 常用的参数模型有线性 Logistic 回归模型:

$$\pi_{\boldsymbol{\beta}}(\boldsymbol{X})=\frac{\exp(\boldsymbol{X}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}'\boldsymbol{\beta})},\tag{1}$$

其中, $\boldsymbol{\beta}\in\Theta\subseteq\mathbb{R}^p$ 是 p 维未知参数. 通常, 研究者采用最大似然法估计未知参数

$$\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}=\arg\max_{\boldsymbol{\beta}\in\Theta}\sum_{i=1}^{n}T_i\log\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)+(1-T_i)\log\{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)\}.$$

在观测性研究中, 上述的 Logistic 回归模型存在误设的可能性, 而处理效应的估计对倾向得分模型的误设十分敏感. 为了提高模型估计的精度, 近年来有研究者提出通过优化协变量平衡的方法来估计倾向得分. Rosenbaum 和 Rubin[9] 证明了倾向得分具有平衡协变量的性质:

$$T\perp\boldsymbol{X}|\pi(\boldsymbol{X}),\tag{2}$$

即在给定倾向得分的条件下, 处理变量 T 和观测的协变量 $\boldsymbol{X}$ 是独立的. 一个好的倾向得分的估计值理论上要满足 (2) 式. 也就是说估计的倾向得分能够使得处理组和对照组在观测的协变量上具有相同的分布. 实际应用中, 研究者一般通过平衡处理组和对照组关于协变量 $\boldsymbol{X}$ 的任意可测函数的矩来实现这一性质:

$$E\left[\frac{Tf(\boldsymbol{X})}{\pi(\boldsymbol{X})}\right]=E\left[\frac{(1-T)f(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right]=E\left[f(\boldsymbol{X})\right], \tag{3}$$

其中, $f(\cdot):\mathbb{R}^p\to\mathbb{R}^m$. (3) 式可以由倾向得分的定义和条件期望的重期望公式得到. 它的直观含义是, 对协变量的任意函数, 经过倾向得分调整的处理组的一阶矩, 和经过倾向得分调整的对照组的一阶矩, 以及总体的一阶矩, 三者之间是相等的. 换言之, 一个真实的倾向得分应该具备平衡处理组和对照组协变量任意阶矩的能力.

一个理想的 Logistic 回归模型应当最小化总体样本的预测误差, 使得整体样本在观测值和真实值之间的差异达到最小. 但在估计倾向得分的时候, 一个最佳的 Logistic 回归模型应该同时将协变量平衡纳入考虑. 基于这样的观察, Imai 和 Ratkovic[14] 提出了协变量平衡倾向得分法, 该方法在估计倾向得分模型的同时考虑了协变量平衡的条件, 从而提高了参数估计的精度. 由 (3) 式, Imai 和 Ratkovic[14] 在估计倾向得分的时候利用了处理组和对照组之间的协变量平衡关系:

$$E\left[\frac{Tf(\boldsymbol{X})}{\pi(\boldsymbol{X})}-\frac{(1-T)f(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right]=0. \tag{4}$$

(4) 式对应的样本矩形式为:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}-\frac{1-T_i}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\right)f(\boldsymbol{X}_i)=0. \tag{5}$$

显而易见, (5) 式保证了处理组和对照组关于函数 $f(\boldsymbol{X}_i)$ 有相同的样本均值. 当 (5) 式中方程的个数等于未知参数的个数时, 即恰好识别的情况, 例如取 $f(\boldsymbol{X}_i)=\boldsymbol{X}_i$, 则可以直接通过求解下列方程组来估计倾向得分模型中的未知参数 $\boldsymbol{\beta}$:

$$\frac{1}{n}\sum_{i=1}^{n}g_{\boldsymbol{\beta}}(T_i,\boldsymbol{X}_i)=0,$$

其中,

$$g_{\boldsymbol{\beta}}(T_i,\boldsymbol{X}_i)=\left(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}-\frac{1-T_i}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\right)f(\boldsymbol{X}_i).$$

当 (5) 式中方程的个数大于未知参数的个数时, 即过度识别的情况, 例如取 $f(\boldsymbol{X}_i)=(\boldsymbol{X}_i',(\boldsymbol{X}_i^2)')'$, 此时 (5) 式中未知参数的解并不唯一. 为了解决这个问题, 我们可以应用广义矩估计的方法 [19] 来求解未知参数 $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}=\arg\min_{\boldsymbol{\beta}\in\Theta}\left(n^{-1/2}\sum_{i=1}^{n}g_{\boldsymbol{\beta}}(T_i,\boldsymbol{X}_i)\right)'\boldsymbol{W}_n(\boldsymbol{\beta})\left(n^{-1/2}\sum_{i=1}^{n}g_{\boldsymbol{\beta}}(T_i,\boldsymbol{X}_i)\right), \tag{6}$$

为了得到更好的有限样本性质, Imai 和 Ratkovic[14] 在估计时采用连续更新的广义矩估计量 [20], 即在 (6) 式中选取

$$\boldsymbol{W}_n^{-1}(\boldsymbol{\beta})=\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(T,\boldsymbol{X})=\frac{1}{n}\sum_{i=1}^{n}E\{g_{\boldsymbol{\beta}}(T_i,\boldsymbol{X}_i)g_{\boldsymbol{\beta}}(T_i,\boldsymbol{X}_i)'|\boldsymbol{X}_i\}.$$

Imai 和 Ratkovic[14] 的协变量平衡倾向得分方法主要是通过广义矩估计求解使得协变量在处理组和对照组之间尽可能地保持平衡, 这样即使在倾向得分误设的情况下, 研究者也能得到一个使得处理组和对照组之间的协变量保持相对平衡的倾向得分的估计量.

2.2 可选择变量的协变量平衡方法

与 Logistic 回归模型相比, 尽管 Imai 和 Ratkovic[14] 的协变量平衡倾向得分方法减小了平均处理效应估计量的均方误差和绝对误差的中位数, 但遗憾的是该方法不能进行变量选择. 也就是说该方法的实施需要研究者预先指定进入模型的协变量, 然而基于经验的协变量的选择要求研究者具有很强的专业背景. 特别是在大数据蓬勃发展的背景下, 高维协变量的问题是研究者经常要面对的情况. 于是, 开发一种数据驱动的可选择变量的协变量平衡方法显得十分必要.

本节, 我们考虑带变量选择的协变量平衡倾向得分的问题. 我们在 Imai 和 Ratkovic[14] 的协变量平衡倾向得分方法的基础上, 结合了 Caner[17] 的 GMM-LASSO 的思想, 提出了带协变量平衡的 GMM-LASSO 估计方法. 我们的方法在估计倾向得分的时候, 既利用了倾向得分的协变量平衡性质, 同时又考虑了变量选择的问题. 这一方法是对 Imai 和 Ratkovic[14] 的协变量平衡倾向得分方法的推广, 有助于提高协变量平衡法在实际中的应用性. 这里, 我们仍然考虑用 (1) 式的线性 Logistic 回归模型对倾向得分进行建模. 利用 (4) 式的矩条件, 我们可以构造 (6) 式的广义矩估计量. 在 (6) 式的基础上, 对未知参数 $\boldsymbol{\beta}$ 加上 L_1 范数惩罚, 于是我们得到了带协变量平衡的 GMM-LASSO 估计量

$$\hat{\boldsymbol{\beta}}_{\text{GMM-LASSO}} = \underset{\boldsymbol{\beta}\in\Theta}{\arg\min} \left(n^{-1/2}\sum_{i=1}^{n} g_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)\right)' \boldsymbol{W}_n(\boldsymbol{\beta}) \left(n^{-1/2}\sum_{i=1}^{n} g_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)\right) + \lambda\sum_{i=1}^{p}|\boldsymbol{\beta}_i|, \tag{7}$$

其中, λ 是惩罚系数, 权重矩阵

$$\boldsymbol{W}_n^{-1}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} E\{g_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) g_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)'|\boldsymbol{X}_i\} = \frac{1}{n}\sum_{i=1}^{n}\left([\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)\{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)\}]^{-1} f(\boldsymbol{X}_i) f(\boldsymbol{X}_i)'\right).$$

为了求解 (7) 式中 GMM-LASSO 估计量, 这里我们采用了近端梯度算法 [21], 详细的描述见附录 B. 值得一提的是, (7) 式中的惩罚函数可以换成其他形式, 比如 Fan 和 Li[22] 的 SCAD (smoothly clipped absolute deviation) 惩罚函数. 另外, 惩罚系数 λ 的选择, 我们采用了 Fan 和 Li[22] 的思想. 当然也可以采用 Caner[17] 的方法.

2.3 估计量的渐近性质

本节, 我们讨论基于 GMM-LASSO 方法的平均处理效应估计量的相合性结果. 首先假设真实的倾向得分模型为

$$\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}) = \frac{\exp(\boldsymbol{X}'\boldsymbol{\beta}_0)}{1+\exp(\boldsymbol{X}'\boldsymbol{\beta}_0)},$$

和真实的平均处理效应为 $\Delta = E[Y(1)-Y(0)]$, 其中, $\boldsymbol{\beta}_0$ 代表 $\boldsymbol{\beta}$ 的真实值. 为简化符号, 记 (7) 式中的 GMM-LASSO 估计量为 $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\text{GMM-LASSO}}$. 将 $\hat{\boldsymbol{\beta}}$ 代入 $\pi_{\boldsymbol{\beta}}$, 则得到 $\hat{\pi} = \pi_{\hat{\boldsymbol{\beta}}}$, 于是, 基于 GMM-LASSO 方法的平均处理效应的估计量为

$$\hat{\Delta} = \tilde{\Delta}(\hat{\pi}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} - \frac{(1-T_i)Y_i}{1-\pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)}\right). \tag{8}$$

在 (7) 式中, 若记

$$Z_n(\boldsymbol{\beta}) = \left(n^{-1}\sum_{i=1}^{n} g_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)\right)' \boldsymbol{W}_n(\boldsymbol{\beta}) \left(n^{-1}\sum_{i=1}^{n} g_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)\right) + \frac{\lambda}{n}\sum_{i=1}^{p}|\boldsymbol{\beta}_i|,$$

则有 $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Z_n(\boldsymbol{\beta})$. 为了得到相合性结论, 我们需要如下假设:

1) 对任意的 $1 \le i \le n$, $E|f(\boldsymbol{X}_i)| < \infty$.

2) 设 $\boldsymbol{W}_n(\boldsymbol{\beta})$ 是关于 $\boldsymbol{\beta}$ 连续的正定矩阵, $\boldsymbol{W}_n(\boldsymbol{\beta})$ 依概率收敛于 $\boldsymbol{W}(\boldsymbol{\beta})$ 关于 $\boldsymbol{\beta}$ 一致成立, 其中 $\boldsymbol{W}(\boldsymbol{\beta})$ 是对称非随机正定矩阵且关于 $\boldsymbol{\beta}$ 连续.

3) 惩罚系数满足 $\lambda = o(n)$.

4) 倾向得分模型是正确设定的, 即真实的倾向得分模型由 Logistic 回归模型给出.

5) 强可忽略条件成立, 即 (a) 给定 $\boldsymbol{X}$ 的条件下, $(Y(1), Y(0))$ 和处理变量 T 独立; (b) 存在 $\varepsilon > 0$, 使得 $\varepsilon < \pi_{\boldsymbol{\beta}}(\boldsymbol{x}) < 1-\varepsilon$.

6) $E|Y(1)| < \infty$ 且 $E|Y(0)| < \infty$.

下面我们陈述本文的主要定理, 证明过程详见附录 A.

定理 1 若假设 1) ∼ 6) 成立, 则 (i) $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ 且 (ii) $\hat{\Delta} \xrightarrow{p} \Delta$.

3 蒙特卡罗模拟

本节, 我们通过蒙特卡罗试验来研究本文提出的 GMM-LASSO 估计方法. 我们比较了最大似然法 (MLE), Imai 和 Ratkovic[14] 的协变量平衡倾向得分 (CBPS) 方法和本文的 GMM-LASSO 方法在基于逆概率加权

的平均处理效应估计量中的表现. 假设我们有 p 维协变量 $(\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \cdots, \boldsymbol{X}_{ip}) \sim N(\boldsymbol{0}, \boldsymbol{I_p}), i = 1, \cdots, n$, 其中 $\boldsymbol{I_p}$ 为 $p \times p$ 单位矩阵. 真实的倾向得分为:

$$\pi(\boldsymbol{X}_i) = \frac{\exp(0.3 - \boldsymbol{X}_{i1} + 0.5\boldsymbol{X}_{i2})}{1 + \exp(0.3 - \boldsymbol{X}_{i1} + 0.5\boldsymbol{X}_{i2})}.$$

处理变量 $T_i \sim B(1, \pi(\boldsymbol{X}_i))$, 其中 $B(1, \pi(\boldsymbol{X}_i))$ 是以概率为 $\pi(\boldsymbol{X}_i)$ 的 0-1 分布. 假设潜在的结果变量 $Y_i(0)$ 和 $Y_i(1)$ 为

$$Y_i(1) = 210 + 27.4\boldsymbol{X}_{i1} + 13.7\boldsymbol{X}_{i2} + 13.7\boldsymbol{X}_{i3} + 13.7\boldsymbol{X}_{i4} + \varepsilon_i(1),$$

$$Y_i(0) = 27.4\boldsymbol{X}_{i1} + 13.7\boldsymbol{X}_{i2} + 13.7\boldsymbol{X}_{i3} + 13.7\boldsymbol{X}_{i4} + \varepsilon_i(0),$$

其中, $\varepsilon_i(1)$ 和 $\varepsilon_i(0)$ 独立同分布于 $N(0, 1)$. 则观测到的结果变量 $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$. 显然, 真实的平均处理效应 $\Delta = 210$. 这里我们考虑 (8) 式的逆概率加权的平均处理效应估计量 $\hat{\Delta}$. 下面我们将用三种不同的方法估计倾向得分, 并把估计的倾向得分代入 (8) 式计算平均处理效应. 三种方法分别为: 1) 对倾向得分建立 Logistic 回归模型并用最大似然法估计未知参数 (MLE); 2) Imai 和 Ratkovic[14] 的协变量平衡倾向得分法 (CBPS); 3) 本文提出的带变量选择的协变量平衡倾向得分法: GMM-LASSO 方法 (gmmlso).

对上述的三种方法, 我们在样本量 n 等于 200, 500 和 1000 的设定下, 分别对协变量维度 p 等于 10, 30 和 50 的情况进行了 1000 次蒙特卡罗模拟, 并计算了每种方法所对应的平均处理效应估计量 $\hat{\Delta}$ 的绝对误差的中位数 (median of absolute error) 和绝对误差的标准差 (standard deviation of absolute error). 由于两步法 (two-step) 的 GMM 估计量比连续更新 GMM 估计量的程序运行速度要快得多, 模拟部分和后续的实证部分的 CBPS 和 GMM-LASSO 方法均采用两步法的 GMM 进行计算. 模拟的结果如表 1 所示.

从表 1 的结果可以看到: 首先, 随着样本量的增大, 三种方法的估计结果都得到了改善; 其次, 在小样本高维度下, 本文的 GMM-LASSO 方法具有明显的优势; 最后, 因为基于 LASSO 的变量选择方法的估计量是有偏估计, 所以, 在样本量较大且协变量维度较低的情况下, 我们的结果并不优于 CBPS 估计量. 但是在一定的样本量下, 随着协变量维度的增高, 本文的 GMM-LASSO 方法的优势会逐渐显现. 总之, 在满足一定的稀疏性的前提下, 基于 GMM-LASSO 方法的逆概率加权估计量可以有效改善平均处理效应的估计结果.

本文, 我们使用 CR (coverage rate) 来反映 GMM-LASSO 方法变量选择的精确度. CR 称为覆盖率, 在本文中表示系数为零的协变量被正确识别出来的比例. CR 的最终取值为 1000 次模拟的中位数, 模拟结果如表 2 所示. 模拟结果表明, 模型越稀疏, 本文的 GMM-LASSO 方法对于系数为零的协变量的正确识别率就越高.

表 1 基于不同倾向得分估计方法的 ATE 估计量的表现

		p=10			p=30			p=50		
		MLE	CBPS	gmmlso	MLE	CBPS	gmmlso	MLE	CBPS	gmmlso
n=200	MAE	5.223	3.521	2.962	9.508	21.169	4.679	15.554	49.300	4.635
	sd	9.382	3.804	2.406	43.229	13.689	3.827	95.028	13.921	4.784
n=500	MAE	2.992	1.383	2.003	3.799	4.113	3.099	4.350	9.335	3.776
	sd	3.865	1.292	1.544	4.677	2.450	1.599	6.009	4.219	2.120
n=1000	MAE	2.075	0.777	1.551	2.312	1.794	2.517	2.778	3.224	3.105
	sd	2.224	0.743	1.221	2.526	1.149	1.261	3.835	1.547	1.174

表 2 不同设定下 CR 的取值

	p=10	p=30	p=50
$n = 200$	0.500	0.750	0.771
$n = 500$	0.625	0.786	0.854
$n = 1000$	0.625	0.857	0.917

4 实证研究

本节, 我们将本文提出的方法应用到实际案例中, 这里考虑了 Ichino 等 [23] 文中用到的数据集. 我们使用 2000 年代初期意大利托斯卡纳地区的数据来研究劳务派遣 (temporary work agency, TWA) 机制是否有

利于工人在今后找到一份稳定的工作.

在劳务派遣机制中, 劳务派遣机构负责雇佣工人并把这些工人派遣到需要员工的公司. 与传统的工作不同, 招募员工的公司直接与劳务派遣机构签订合同, 并负责培训和指导工人, 而工人的工资和福利则由劳务派遣机构负责. 引入劳务派遣机制的主要目的之一是为了帮助面临就业困难的工人在今后找到一份稳定的工作. 从理论上说, 有两个原因可以支持劳务派遣的方式是一种有利的机制: 一是, 工人可以通过劳务派遣的方式来表明自己的工作能力; 二是, 劳务派遣可以为工人提供获取额外人力资本, 社会交往和职位空缺信息的机会.

为了评估劳务派遣机制是否对就业有积极影响, Ichino 等 [23] 在 2000 年代初期收集了意大利托斯卡纳和西西里这两大地区的数据. 该数据集包含 2030 个样本, 其中处理组有 511 个, 而对照组有 1519 个. 其中, 处理组包括在 2001 年的前 6 个月内参与劳务派遣分配的个人; 而对照组则包括年龄在 18 至 40 岁之间的人群, 这部分群体属于劳动力, 但 2001 年 1 月没有稳定的工作, 并且在 2001 年前半年没有参与劳务派遣分配. 这样, 两个组的群体均来自同一劳动力市场. 同时, 该数据集具有一个丰富的变量集合, 包括人口特征、家庭背景、学历和工作经验等.

本文, 我们重点分析托斯卡纳地区男性工人 (包含 339 个样本) 的情况. 我们感兴趣的结果变量是工人在 2002 年底是否有一份稳定的工作 (以是否签订一份长期合同来衡量). 处理变量为工人是否参加劳务派遣. 协变量与 Ichino 等 [23] 文中用来识别倾向得分的协变量一致, 包含 Ichino 等 [23] 文中表 1 的全部, 距离的平方和一个交叉项 (共 29 个维度). 我们使用基于倾向得分加权的估计量来评估平均处理效应. 这里考虑了三种不同的倾向得分的估计方法: 最大似然估计法 (MLE), 协变量平衡倾向得分法 (CBPS) 以及本文提出的 GMM-LASSO 方法 (gmmlso). 表 3 给出了基于不同倾向得分估计方法的 ATE 的点估计和标准差. 这里, 我们采用 Bootstrap 方法计算标准差: 首先, 从原始样本中有放回重复抽取 339 个样本, 根据抽出的样本计算 ATE 估计量; 重复上述步骤 500 次, 得到 500 个 ATE 的估计量, 最后计算上述估计量的样本标准差.

本例中, GMM-LASSO 方法共选出了 24 个相关变量. 此外, 从表 3 可以得出, 当置信水平 $\alpha = 0.05$ 时, 基于 MLE, CBPS 和 GMM-LASSO 方法的平均处理效应的置信区间分别为 $(-0.0932, 0.381)$, $(0.0812, 0.375)$ 和 $(0.114, 0.480)$. 基于 MLE 方法的置信区间包括 0, 基于 CBPS 和 GMM-LASSO 方法的置信区间不包含 0, 且都在 0 的右侧. 这说明在 0.05 的置信水平下, 基于 MLE 方法的平均处理效应不显著; 但基于 CBPS 和 GMM-LASSO 方法的平均处理效应显著为正.

表 3　基于不同倾向得分估计方法的 ATE 估计

MLE	CBPS	gmmlso
0.144	0.228	0.297
(0.121)	(0.0749)	(0.0932)

5　结论

一个好的倾向得分的估计, 需要考虑两个方面的问题: 一方面, 倾向得分应使得处理组和对照组在每个协变量上达到相同的分布, 即协变量平衡; 另一方面, 正确选择建立倾向得分模型的协变量. 现有的估计倾向得分的方法更多地专注于协变量平衡方面, 并没有很好地解决协变量的选择问题. 本文提出了基于 GMM-LASSO 方法的协变量平衡倾向得分估计量. 该估计量在估计倾向得分时, 除了进行协变量平衡, 还兼顾了变量选择. 模拟结果表明, 在样本量小且协变量具有高维稀疏的特征时, 该估计量对平均处理效应的估计具有更好的结果. 理论上, 文章证明了基于该估计量的平均处理效应估计量具有相合性. 因此, 本文的方法有助于提高协变量平衡方法在研究中的实用性. 最后, 值得一提的是, 我们提出的估计量的其他大样本性质 (比如渐进正态性) 以及如何选择 (5) 中最佳的函数 $f(\cdot)$ 等, 这些问题需要我们进一步去研究.

参考文献

[1] Rubin D B. Matching to remove bias in observational studies[J]. Biometrics, 1973, 29(1): 159–183.
[2] Rubin D B. Estimating causal effects of treatments in randomized and nonrandomized studies[J]. Journal of Educational Psychology, 1974, 66(5): 688–701.
[3] Rubin D B. Assignment to treatment group on the basis of a covariate[J]. Journal of Educational Statistics, 1977, 2(1): 1–26.
[4] Rubin D B. Bayesian inference for causal effects: The role of randomization[J]. The Annals of Statistics, 1978, 6(1): 34–58.
[5] Holland P W. Statistics and causal inference[J]. Journal of the American Statistical Association, 1986, 81(396): 945–960.
[6] Rubin D B. Inference and missing data[J]. Biometrika, 1976, 63(3): 581–592.
[7] Robins J M, Rotnitzky A, Zhao L P. Estimation of regression coefficients when some regressors are not always observed[J]. Journal of the American Statistical Association, 1994, 89(427): 846–866.
[8] Scharfstein D O, Rotnitzky A, Robins J M. Adjusting for nonignorable drop-out using semiparametric nonresponse models[J]. Journal of the American Statistical Association, 1999, 94(448): 1096–1120.
[9] 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.
[10] Kang J D, Schafer J L. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data[J]. Statistical Science, 2007, 22(4): 523–539.
[11] Dehejia R H, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs[J]. Journal of the American Statistical Association, 1999, 94(448): 1053–1062.
[12] Leeb H, Pötscher B M. Model selection and inference: Facts and fiction[J]. Econometric Theory, 2005, 21(1): 21–59.
[13] 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.
[14] Imai K, Ratkovic M. Covariate balancing propensity score[J]. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 2014, 76(1): 243–263.
[15] Sant'Anna P H C, Song X, Xu Q. Covariate distribution balance via propensity scores[J]. Available at SSRN 3258551, 2019.
[16] Ning Y, Peng S, Imai K. Robust estimation of causal effects via high-dimensional covariate balancing propensity score[J]. Biometrika, 2020, 107(3): 533–554.
[17] Caner M. Lasso-type GMM estimator[J]. Econometric Theory, 2009, 25(1): 270–290.
[18] Rosenbaum P R. Model-based direct adjustment[J]. Journal of the American Statistical Association, 1987, 82(398): 387–394.
[19] Hansen L P. Large sample properties of generalized method of moments estimators[J]. Econometrica, 1982, 50(4): 1029–1054.
[20] Hansen L P, Heaton J, Yaron A. Finite-sample properties of some alternative GMM estimators[J]. Journal of Business & Economic Statistics, 1996, 14(3): 262–280.
[21] Boyd S, Parikh N. Proximal algorithms[J]. Foundations and Trends in Optimization, 2013, 1(3): 123–231.
[22] Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348–1360.
[23] Ichino A, Mealli F, Nannicini T. From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?[J]. Journal of Applied Econometrics, 2008, 23(3): 305–327.
[24] Van Der Vaart A W, Wellner J A. Weak convergence and empirical processes[M]. New York: Springer, 1996.

附录

A. 数学证明

定理 1 的证明: (i) 因为 $(T_i, \boldsymbol{X}_i)$, $1 \le i \le n$ 是独立同分布的, 所以由大数定律, 关于 $\boldsymbol{\beta}$ 一致地有

$$\frac{1}{n}\sum_{i=1}^{n} g_{\boldsymbol{\beta}}\left(T_i, \boldsymbol{X}_i\right) \xrightarrow{p} m(\boldsymbol{\beta}),$$

其中,

$$m(\boldsymbol{\beta}) = E\left[\frac{Tf(\boldsymbol{X})}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X})} - \frac{(1-T)f(\boldsymbol{X})}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X})}\right].$$

由假设 2 和 3, 我们得到

$$Z_n(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\Theta} \left(n^{-1}\sum_{i=1}^{n} g_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)\right)' \boldsymbol{W}_n(\boldsymbol{\beta}) \left(n^{-1}\sum_{i=1}^{n} g_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)\right) + \frac{\lambda}{n}\sum_{i=1}^{p}|\boldsymbol{\beta}_i|$$
$$\xrightarrow{p} m(\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{\beta})m(\boldsymbol{\beta}) =: Z(\boldsymbol{\beta}).$$

最后, 由 Van Der Vaart 和 Wellner[24] 书中的推论 3.2.3, 我们得到

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Z_n(\boldsymbol{\beta}) \xrightarrow{p} \arg\min_{\boldsymbol{\beta}} Z(\boldsymbol{\beta}) = \boldsymbol{\beta}_0.$$

(ii) 利用 (i) 的结论, 我们有 $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$. 由假设 5)(b), 可以得到

$$\left|\frac{\mathrm{TY}}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X})} - \frac{(1-T)Y}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X})}\right| \le \frac{2|Y|}{\varepsilon}.$$

由假设 6, 我们有 $\mathbb{E}|Y| < \infty$. 于是, 根据控制收敛定理可得

$$\widehat{\Delta} = E\left(\frac{\mathrm{TY}}{\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X})} - \frac{(1-T)Y}{1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X})}\right) + o_p(1).$$

注意到, $Y = \mathrm{TY}(1) + (1-T)Y(0)$. 在假设 5)(a) 下, 由条件期望的重期望公式, 上式可简化为,

$$\widehat{\Delta} = E\left(\frac{\mathrm{TY}}{\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X})} - \frac{(1-T)Y}{1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X})}\right) + o_p(1) = E\left(\frac{\mathrm{TY}(1)}{\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X})} - \frac{(1-T)Y(0)}{1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X})}\right) + o_p(1)$$
$$= E\left(\frac{E(T|\boldsymbol{X})E(Y(1)|\boldsymbol{X})}{\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X})} - \frac{(1-E(T|\boldsymbol{X}))E(Y(1)|\boldsymbol{X})}{1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X})}\right) + o_p(1).$$

在倾向得分模型正确设定的条件下, 我们有

$$\widehat{\Delta} = E\left[E\left(Y(1)|\boldsymbol{X}\right) - E\left(Y(0)|\boldsymbol{X}\right)\right] + o_p(1) = E\left(Y(1) - Y(0)\right) + o_p(1) = \Delta + o_p(1).$$

这就完成了定理的证明.

B. 算法

近端梯度法是一种特殊的梯度下降方法, 主要用于求解目标函数不可微的最优化问题. 如果目标函数在某些点是不可微的, 那么该点的梯度无法求解. 此时, 传统的梯度下降法就无法使用. 近端梯度算法的基本思想是, 使用临近算子作为近似梯度进行梯度下降. 此算法解决凸优化问题模型如下:

$$\min F(\boldsymbol{x}) = G(\boldsymbol{x}) + h(\boldsymbol{x}), \tag{9}$$

其中, $G(\boldsymbol{x})$ 是凸函数且可微, $h(\boldsymbol{x})$ 是凸函数但不可微. 定义凸函数 $h(\boldsymbol{x})$ 的近端投影如下:

$$\mathrm{prox}_h(\boldsymbol{z}) := \arg\min_{\boldsymbol{\theta}\in R^p}\left\{\frac{1}{2}\|\boldsymbol{z}-\boldsymbol{\theta}\|_2^2 + h(\boldsymbol{\theta})\right\}. \tag{10}$$

为了求解 (9) 式所示问题, 近端梯度法中 $\boldsymbol{x}$ 的迭代递推公式为:

$$\boldsymbol{x}^{t+1} = \mathrm{prox}_{s^t h}(\boldsymbol{x}^t - s^t \bigtriangledown G(\boldsymbol{x}^t)).$$

当不可微函数 $h(\boldsymbol{x})$ 是 l_1 范数, 即 $h(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|_1$, 近端投影 (10) 式为:

$$\mathrm{prox}_{\mathrm{sh}}(\boldsymbol{z}) = \arg\min_{\boldsymbol{\theta}\in R^p}\left\{\frac{1}{2s}\|\boldsymbol{z}-\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1\right\} = \arg\min_{\boldsymbol{\theta}\in R^p}\left\{\frac{1}{2}\|\boldsymbol{z}-\boldsymbol{\theta}\|_2^2 + \lambda s\|\boldsymbol{\theta}\|_1\right\}$$
$$= \arg\min_{\boldsymbol{\theta}\in R^p}\left\{\sum_{j=1}^{p}\left\{\frac{1}{2}(\boldsymbol{z}_j-\boldsymbol{\theta}_j)^2 + \lambda s\,|\boldsymbol{\theta}_j|\right\}\right\}.$$

用软阈值算子表示上式的解为

$$[\mathcal{S}_\tau(\boldsymbol{z})]_j = \mathrm{sign}(\boldsymbol{z}_j)(|\boldsymbol{z}_j| - \tau)_+, j = 1, 2, \cdots, p,$$

其中, 阈值 $\tau = s\lambda$, $(x)_+$ 表示 $\max\{x, 0\}$.

因此, 当 $h(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|_1$ 时, 近端梯度法可等价通过以下两个步骤来实现: 1) 在第 t 步迭代时, 通过梯度下降法得到一个点 $\boldsymbol{z} = \boldsymbol{x}^t - s^t \bigtriangledown G(\boldsymbol{x}^t)$, s^t 表示第 t 步迭代的步长; 2) 对该点 $\boldsymbol{z}$ 的元素应用软阈值算子得到第 $t+1$ 步的解 $\boldsymbol{x}^{t+1} = \mathcal{S}_{s^t\lambda}(\boldsymbol{z})$. 用线性搜索法确定步长 s^t, 算法如下: 给定 $\boldsymbol{x}^k$, s^{k-1} 和参数 $\beta \in (0, 1)$. 令 $s = s^{k-1}$, 重复

1. $\boldsymbol{z} = \mathrm{prox}_{\mathrm{sh}}(\boldsymbol{x}^t - s \bigtriangledown G(\boldsymbol{x}^t))$
2. break if　$G(\boldsymbol{z}) \le \hat{G}_s(\boldsymbol{z}, \boldsymbol{x}_k)$
3. update　　$s = \beta s$

返回 $s^k = s$, $\boldsymbol{x}^{k+1} = \boldsymbol{z}$, 其中 $\hat{G}_s(\boldsymbol{x}, \boldsymbol{y}) = G(\boldsymbol{y}) + \bigtriangledown G(\boldsymbol{y})^\top(\boldsymbol{x}-\boldsymbol{y}) + \frac{1}{2s}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$.