半参数模型的影响函数

嘉树

2026 年 3 月

说明: 本文是对「影响函数」(influence function) 的介绍，讲述了它在半参数模型中扮演的角色，并以构造正交矩条件为例展示了其威力。影响函数本质上是对「梯度」概念在泛函空间中的推广，刻画了分布扰动对估计量的一阶影响。由此，它是分析估计量效率、对模型误设的稳健性等性质的利器。

1 影响函数的定义

我们考虑半参数意义上的估计量，即那些依赖于对非参数部分估计的对参数的估计量。记 $\hat{\theta}$ 就是一个半参数意义的估计量，它是数据 $W_1, \ldots, W_n$ 的函数。我们始终假设数据是独立同分布的，且有一个累积分布函数 F_0。当 F_0 是数据 W_i 的累积分布函数时，记 θ_0 是 $\hat{\theta}$ 的概率极限。

我们关注的是这样一类线性估计量，满足如下的渐进展开[1]

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i) + o_p(1)$$

其中 $\psi(\cdot)$ 满足

$$\mathbb{E}[\psi(W)] = 0 \quad 和 \quad \mathbb{E}[\psi(W)'\psi(W)] < \infty$$

于是 $\hat{\theta}$ 的渐进方差就是 $\mathbb{E}[\psi(W)\psi(W)']$。函数 $\psi(\cdot)$ 就被称为「影响函数」(influence function)[2]，它描述了每个数据点对估计量的影响，同时我们还将看到，它也刻画了数据分布的一个微小扰动对 $\hat{\theta}$ 的概率极限的影响。有许多 $\sqrt{n}$ 一致的估计量都有这样的渐进线性展开，如 M 估计量、Z 估计量等。

其实影响函数不一定需要通过渐进展开来构造。设 F 是任意分布（只需满足一些正则条件），令 $\theta(F)$ 表示估计量 $\hat{\theta}$ 在数据的累积分布函数为 F 时的概率极限。由此 $\theta(F)$ 就刻画了当模型误设时，估计量会收敛到什么伪真值。为了研究分布的微小扰动，我们考虑一个参数化的分布族 $\{F_\beta\}$，当 $\beta = 0$ 时，$F_\beta = F_0$。得分函数（对数似然函数关于参数 β 在 $\beta = 0$ 处的导数）表示为 $S_\beta(w)$。可以证明，影响函数 $\psi(W)$ 满足如下微分方程[3]

$$\left. \frac{\partial \theta(F_\beta)}{\partial \beta} \right|_{\beta = 0^+} = \mathbb{E}_{F_0}[\psi(W) S_\beta(W)]$$

其中期望是关于分布 F_0 的。可以通过求解这个微分方程得到影响函数。

Ichimura and Newey (2022) 展示了一种更方便的求解影响函数的方法，他们使用 Gâteaux 导数来定义泛函微分。设 H 是一个累积分布函数，考虑分布扰动 $F_\tau = (1-\tau)F_0 + \tau H$，其中 $\tau \in [0, C)$，$C \in (0,1)$ 是某个常数。F_τ

[1] 严格来说，除了渐进线性展开，我们还要求估计量 $\hat{\theta}$ 是正则的，粗略的含义是，让真实分布 F_0 产生微小扰动后，估计量的极限分布保持不变，见 (1) 的定义。满足正则性和渐进线性展开的估计量被称为正则渐进线性估计量 (regular asymptotically linear estimator, RAL)。

[2] 这其实是一个稳健统计学中的经典概念，最早来自 Hampel (1974)。核心思想就是看数据分布的一阶扰动对估计量的影响，和半参数效率理论有紧密而深刻的联系。

[3] $\theta(F_\beta)$ 是关于分布 F_β 的泛函，其导数定义为 $\lim_{\beta \to 0^+} \beta^{-1}[\theta(F_\beta) - \theta(F_0)]$，它的存在性依赖于前面所说的正则渐进线性条件（还潜在地依赖于二次均值可微条件，见技术附录），当这个条件满足时，导数就存在，且一定可以表示成影响函数和得分函数的内积，见 van der Vaart (1991) 的引理 A.1 或 Newey (1994) 的定理 2.1。我们还可以更进一步，我们可以将这个泛函的导数看作是切空间（因为得分函数属于切空间）上的泛函，那么根据 Riesz 表示定理，存在该切空间上的一个唯一的元素 ψ^*，使得此泛函可以表示成 ψ^* 和得分函数的内积。这个 ψ^* 就是「有效率的影响函数」(efficient influence function, EIF)。其他的影响函数在切空间上的投影就是 ψ^*。

的意思是说，数据大部分来自于 F_0，但有很小的概率 τ 来自于 H[4]。设对选定的 H，$\theta(F_\tau)$ 存在且关于 τ 可导（导数就是 Gâteaux 导数），则影响函数 $\psi(W)$ 满足

$$\left.\frac{d\theta(F_\tau)}{d\tau}\right|_{\tau=0^+} = \int \psi(w)H(dw) = \mathbb{E}_H[\psi(W)]$$

其中 $\mathbb{E}[\psi(W)] = 0$ 且 $\mathbb{E}[\psi(W)^2] < \infty$。上式左边就衡量了一个轻微污染的分布对于估计量的极限值的影响，等式右边定义了一个线性泛函。

[4] 这种形式的分布扰动来自 Hampel (1974)。可以证明这种偏离路径是二次均值可微的。

2 影响函数的威力：以构造正交矩条件为例

泛函导数 $d\theta(F_\tau)/d\tau$ 及影响函数是一个大有可为的概念级工具。首先其基本含义是，它衡量了分布 F_τ 的局部扰动对 $\theta(F_\tau)$ 的影响。我们可以赋予它经济学上的意义，譬如，它可以表示数据分布的局部变化带来的政策效应；又如，当 $\theta(F_\tau)$ 是一个估计量的概率极限时，这个导数就可以看作模型误设对估计量的影响，从而可以应用在估计量的局部敏感性和稳健性分析中；再如，当 $\theta(F_\tau)$ 是矩条件的期望时（在代入非参数估计量的概率极限后），影响函数可以被用来构造正交矩条件，正交的含义是矩条件的成立不受对非参数部分的估计好坏影响[5]。我们以构造正交矩条件为例来展示其威力，前两个用途的示例见 Ichimura and Newey (2022)。构造正交矩条件是最近几年大热的双重机器学习的两个基石之一（另一个是样本拆分），而这一切都离不开影响函数。关于影响函数在双重机器学习中所扮演角色的精彩介绍，见 Kennedy (2023)。

[5] 这就是 Neyman 正交化。

假设我们有总体矩条件

$$\mathbb{E}_{F_0}[g(W, \gamma_0, \theta_0)] = 0$$

其中 $g(\cdot,\cdot,\cdot)$ 是形式未知的矩函数，γ_0 是未知的非参数部分，θ_0 是未知的参数部分。我们感兴趣的是对 θ_0 的估计。

如果我们用机器学习的算法获得了对 γ_0 的第一步估计 $\hat{\gamma}$，那么获得样本矩函数的一个最直接的方式就是直接代入 $\hat{\gamma}$，得到 $\sum_{i=1}^n g(W_i, \hat{\gamma}, \theta)/n$，然后对其使用 GMM 得到 θ_0 的估计量。然而，这种 plug-in 的方法会给 θ_0 的估计带来严重的偏误，因为我们的第一步估计 $\hat{\gamma}$ 本就不够精确——毕竟机器学习算法总是会用一些正则化手段来避免过拟合，从而带来较大的偏误。我们是否能重新构造一个矩条件，使得它不受第一步估计 $\hat{\gamma}$ 好坏的影响呢？答案是我们需要利用矩条件的影响函数。

记 $\hat{\gamma}$ 在数据分布为 F 时的概率极限为泛函 $\gamma(F)$，从而 $\gamma_0 = \gamma(F_0)$。使用前文所介绍的 F_τ 形式的分布扰动，假设存在一个函数 $\phi(w, \gamma, \alpha, \theta)$，使得对任意的 θ 和 H 都有

$$\left.\frac{\partial}{\partial\tau}\mathbb{E}_{F_0}[g(W, \gamma(F_\tau), \theta)]\right|_{\tau=0^+} = \int \phi(w, \gamma_0, \alpha_0, \theta)H(dw) = \mathbb{E}_H[\phi(W, \gamma_0, \alpha_0, \theta)]$$

其中 $\mathbb{E}_{F_0}[\phi(W, \gamma_0, \alpha_0, \theta)] = 0$ 且 $\mathbb{E}_{F_0}[\phi(W, \gamma_0, \alpha_0, \theta)^2] < \infty$。这里的 ϕ 就是泛函 $\mu(F) := \mathbb{E}_F[g(W, \gamma(F), \theta)]$ 的影响函数，或者按照文献中的说法，称为「第一步影响函数」(first step influence function, FSIF)。

正交矩函数就可以构造为

$$\psi(W, \gamma, \alpha, \theta) = g(W, \gamma, \theta) + \phi(W, \gamma, \alpha, \theta)$$

即原矩条件加上其影响函数。这个做法事实上是利用了影响函数的梯度性质，将分布扰动的一阶影响加回去。显然有 $\mathbb{E}_{F_0}[\psi(W,\gamma_0,\alpha_0,\theta)]=0$。

通常我们要求影响函数满足规范化条件，即任意扰动分布下都有零均值，$\mathbb{E}_{F_\tau}[\phi(W,\gamma(F_\tau),\alpha(F_\tau),\theta)]\equiv 0$，两边关于 τ 在 $\tau=0^+$ 处求导：

$$
\begin{aligned}
0 &\equiv \frac{\partial}{\partial\tau}\int \phi(W,\gamma(F_\tau),\alpha(F_\tau),\theta)[F_0+\tau(H-F_0)](dw)\bigg|_{\tau=0^+} \\
&= \frac{\partial}{\partial\tau}\int \phi(w,\gamma(F_\tau),\alpha(F_\tau),\theta)F_0(dw)\bigg|_{\tau=0^+} + \int \phi(w,\gamma_0,\alpha_0,\theta)(H-F_0)(dw) \\
&= \frac{\partial}{\partial\tau}\mathbb{E}_{F_0}[\phi(W,\gamma(F_\tau),\alpha(F_\tau),\theta)]\bigg|_{\tau=0^+} + \mathbb{E}_H[\phi(W,\gamma_0,\alpha_0,\theta)] \\
&= \frac{\partial}{\partial\tau}\mathbb{E}_{F_0}[\phi(W,\gamma(F_\tau),\alpha(F_\tau),\theta)]\bigg|_{\tau=0^+} + \frac{\partial}{\partial\tau}\mathbb{E}_{F_0}[g(W,\gamma(F_\tau),\theta)]\bigg|_{\tau=0^+} \\
&= \frac{\partial}{\partial\tau}\mathbb{E}_{F_0}[\psi(W,\gamma(F_\tau),\alpha(F_\tau),\theta)]\bigg|_{\tau=0^+}
\end{aligned}
$$

第二步利用了链式法则，第三步利用了 $\mathbb{E}_{F_0}[\phi(W,\gamma_0,\alpha_0,\theta)]=0$。这个结论意味着 γ 的局部扰动不会影响矩条件的成立，也就达到了正交化的目的。

技术附录：泛函的可微性

本附录来自 van der Vaart (1991)，为泛函的可微性做出定义。泛函的导数是方向导数的一般化。

设 P 是一个概率测度，它对应真实的模型，$\mathscr{P}$ 是一个概率测度族，它包含了对 P 的某些偏离的概率测度，一个偏离以 P_t 表示，$t \in (0, \varepsilon) \subseteq \mathbb{R}$。令 $\mathscr{P}(P)$ 表示这种偏离路径的集合，即映射 $t \mapsto P_t$ 的集合，每个映射从一个区间 $(0, \varepsilon) \subseteq \mathbb{R}$ 映射到 $\mathscr{P}$，因此每个映射都代表了 P 的一个偏离路径，而且每一个偏离路径都需要满足：存在某个 $g \in L^2(P)$ 使得

$$\lim_{t \to 0^+} \int \left[\frac{1}{t}(dP_t^{1/2} - dP^{1/2}) - \frac{1}{2} g dP^{1/2} \right]^2 = 0$$

其中 dP_t 和 dP 分别是 P_t 和 P 的（关于某个支配测度的）密度函数。这个条件事实上对偏离路径 $t \mapsto P_t$ 做出了一些限制，它要求偏离路径具有某种意义上的光滑性。满足这个条件的路径也被称为「二次均值可微」(quadratic mean differentiability)。函数 g 相当于我们熟悉的得分函数❻。

从几何意义上说，得分函数 g 可以理解为在 $t = 0$ 处切于偏离路径 $t \mapsto P_t$ 的一个切向量。这使得 $\mathscr{P}(P)$ 可以引致一个所谓的「切空间」(tangent space) $T(P)$，它是由所有满足上述条件的函数 g 张成的线性空间。

设 $(\boldsymbol{B}, \|\cdot\|)$ 是一个巴拿赫空间❼。称泛函 $\kappa \colon \mathscr{P} \to (\boldsymbol{B}, \|\cdot\|)$ 在 P 处相对于 $\mathscr{P}(P)$ 可微，如果它满足：存在一个连续的线性映射 $\kappa'_P \colon (T(P), \|\cdot\|_P) \to (\boldsymbol{B}, \|\cdot\|)$ 使得对 $\mathscr{P}(P)$ 中的每一条路径 $t \mapsto P_t$ 及其相应的 $g \in T(P)$ 使得

$$\lim_{t \to 0^+} \frac{1}{t}[\kappa(P_t) - \kappa(P)] = \kappa'_P(g)$$

κ'_P 被称为 κ 在 P 处相对于 $\mathscr{P}(P)$ 的导数❽。这个导数是一个泛函，而且是连续的线性泛函，它将切空间 $T(P)$ 中的元素 g 映射为 $(\boldsymbol{B}, \|\cdot\|)$ 中的元素 $\kappa'_P(g)$。连续线性是一个很关键的要求，只有这样才会有后面的 Riesz 表示。需要强调，这个定义不仅仅是定义左边极限的存在性，还说明了它能被表示成切空间上的泛函。

对于巴拿赫空间 $\boldsymbol{B}$，有其对偶空间 $\boldsymbol{B}^*$，它是由所有连续线性实值泛函 $b^* \colon \boldsymbol{B} \to \mathbb{R}$ 构成的空间。对每个 $b^* \in \boldsymbol{B}^*$，映射 $b^* \circ \kappa'_P$ 是从 $(T(P), \|\cdot\|_P)$ 到 $\mathbb{R}$ 的连续线性泛函。根据 Riesz 表示定理，存在一个唯一的 $\tilde{\kappa}_{P,b^*} \in \bar{T}(P)$ 使得，对每个 $g \in T(P)$ 都有❾

$$b^* \circ \kappa'_P(g) = \langle g, \tilde{\kappa}_{P,b^*} \rangle_P$$

特别地，对于正文中的泛函 $\theta(F_\beta)$，当 θ 是标量时，$\boldsymbol{B}$ 就是实数集 $\mathbb{R}$，因此我们就可以取 b^* 为恒等映射，于是存在一个唯一的 $\psi \in \bar{T}(F_0)$ 使得，对每个 $S_\beta \in T(F_0)$ 都有

$$\theta'(F_\beta) = \langle \psi, S_\beta \rangle_{F_0} = \mathbb{E}_{F_0}[\psi(W) S_\beta(W)]$$

这里的 Riesz 表示元 ψ 正是 EIF。

以下定理是对 Hájek-Le Cam 卷积定理的推广，见 van der Vaart (1991) 的定理 2.1。

定理 1. *设对任意路径 $\{P_t\} \in \mathscr{P}(P)$ 和实数列 $h_n \to h \in \mathbb{R}$，有*

$$\sqrt{n}\big[T_n - \kappa(P_{h_n/\sqrt{n}})\big] \rightsquigarrow_{P_{h_n/\sqrt{n}}} L \tag{1}$$

❻ 怎么理解？它可以看成 $dP_t^{1/2} - dP^{1/2} \approx \frac{t}{2} g dP^{1/2}$。为什么要用平方根密度呢？而且还带有一个 1/2？我们考虑普通的得分函数 $s(w) = \partial \log p_t(w)|_{t=0}$，对 $p_t(w)$ 在 $p_0(w)$ 处做一阶展开：

$$p_t(w) = p_0(w)[1 + t \cdot s(w) + o(t)]$$

两边开根号

$$\sqrt{p_t(w)} \approx \sqrt{p_0(w)}\left[1 + \frac{t}{2} s(w)\right]$$

所以（当似然函数可导时）g 就是得分函数。不过这里其实我们并不要求似然函数可导，因此 g 是更一般的得分函数。

❼ 即完备的线性赋范空间。就我们考察的半参数问题而言，当参数 θ 是标量时，$\boldsymbol{B}$ 就是实数集 $\mathbb{R}$；当参数 θ 是向量时，$\boldsymbol{B}$ 就是欧几里得空间 $\mathbb{R}^k$。

❽ 如果 $\mathscr{P}(P)$ 是所有满足前面光滑性条件的路径的集合，那么这个导数就是 Hadamard 导数。但一般情况下，比如在半参数模型中，由于模型限制，我们并不会考虑这么多的路径。

❾ 通常我们还要加上规范化条件 $\mathbb{E}_P[\tilde{\kappa}_{P,b^*}] = 0$ 来保证唯一性，因为 $\mathbb{E}_P[g] = 0$。

其中 L 是 $\boldsymbol{B}$ 上的一个固定的紧密分布。此外，设对任意 $g \in T(P)$ 都有

$$\left(\sqrt{n}(T_n - \kappa(P)), n^{-1/2}\sum_{j=1}^{n} g(X_j)\right)$$

依 P 在 $\boldsymbol{B} \times \mathbb{R}$ 上联合弱收敛。那么 *(i)* $\kappa\colon \mathscr{P} \to (\boldsymbol{B}, \|\cdot\|)$ 在 P 处相对于 $\mathscr{P}(P)$ 可微，且 *(ii)* 存在一个 $\boldsymbol{B}$ 上的紧密的概率测度 N 满足：对任意 $b^* \in \boldsymbol{B}^*$

$$N \circ b^{*-1} = \mathcal{N}(0, \|\tilde{\kappa}_{P,b^*}\|_P^2)$$

即 N 是巴拿赫空间上的一个高斯分布，其由 b^* 投影到实数轴上的一维边缘分布都是均值为 0，方差为 $\|\tilde{\kappa}_{P,b^*}\|_P^2$ 的高斯分布。此外，极限分布 L 可以表示为 N 和另一个 $\boldsymbol{B}$ 上的一个概率测度的卷积。

这里，满足条件 (1) 的估计量 T_n 称为「正则」(regular) 估计量。这个条件和联合弱收敛条件一起保证了 κ 的可微性，且导数能表述成影响函数和得分函数的内积，见 van der Vaart (1991) 的引理 A.1。结论 (ii) 即卷积定理，描述了估计量的渐进效率下界，N 就代表了最优极限分布。

参考文献

Hampel, F. R. (1974): “The Influence Curve and its Role in Robust Estimation,” *Journal of the American Statistical Association*, 69(346), 383–393.

Ichimura, H. and Newey, W. K. (2022): “The Influence Function of Semiparametric Estimators,” *Quantitative Economics*, 13(1), 29–61.

Kennedy, E. H. (2023): “Semiparametric Doubly Robust Targeted Double Machine Learning: A Review,” *arXiv preprint*, arXiv: 2203.06469.

Newey, W. K. (1994): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62(6), 1349–1382.

van der Vaart, A. (1991): “On Differentiable Functionals,” *Annals of Statistics*, 19(1), 178–204.