局部投影估计

嘉树

2026 年 3 月

说明：本文是对局部投影估计的简要介绍，着重强调了和 VAR 的联系以及稳健性。这个方法本质上就是 OLS，但正是因为简单，才有了相比 VAR 更佳的稳健性。

局部投影法 (local projection, LP) 是一种估计脉冲响应的方法，由 Jordà (2005) 提出，在近十年受到越来越多的关注和应用。在介绍这个方法之前，我们先回顾一下什么是脉冲响应。

1 脉冲响应

所谓脉冲响应 (impulse response, IR)，就是指一个变量 y 受到另一个变量 x 的变化而产生的变化，x 的变化是未预期到的，因此是一个冲击或脉冲[1]，而 y 的变化是动态的，即自冲击发生开始，y 会持续地发生变化，因而这种响应量是时间位移的函数，也就是*脉冲响应函数* (impulse response function, IRF)。从因果效应的角度看，脉冲响应函数描述了 x 的冲击对 y 的*动态*因果效应。在应用宏观经济学中，脉冲响应无疑是最核心的研究对象之一，被用来研究各种经济冲击对经济变量的影响，如货币政策冲击对产出、通货膨胀、利率等的影响。

[1] 「脉冲」一词原本来自信号处理领域，表示一个瞬间的、突然的信号变化。因宏观经济学有研究各种经济冲击效应的需要，「脉冲响应」被顺势引入，其含义用来对应「冲击」也恰如其分。

考虑一个简单模型，假设冲击序列 $\{x_t\}$ 独立同分布，响应变量 y_t 是由冲击序列合成的 MA(∞) 过程：

$$y_t = \sum_{j=0}^{\infty} \beta_j x_{t-j}$$

那么 y_{t+h} 对 x_t 的脉冲响应正是 β_h。

更一般地，当不知道 y_t 的具体生成过程时，定义其对 δ 单位冲击的脉冲响应为

$$\mathrm{IR}(h,\delta) := \mathbb{E}(y_{t+h}|x_t=\delta) - \mathbb{E}(y_{t+h}|x_t=0) \qquad h=0,1,2,\dots$$

通常将冲击强度标准化为 $\delta=1$，即 $\mathrm{IR}(h,1)=\mathrm{IR}(h)$。大多数情况下，还有其他变量 $\boldsymbol{w}_t$ 影响 y_t，此时，需要控制这些变量以隔离出冲击的偏效应，因而脉冲响应就是

$$\mathrm{IR}(h) = \mathbb{E}(y_{t+h}|x_t=1,\boldsymbol{w}_t) - \mathbb{E}(y_{t+h}|x_t=0,\boldsymbol{w}_t) \qquad h=0,1,2,\dots$$

这个表达式很自然地使我们用线性回归的方式来刻画变量关系：

$$y_{t+h} = \alpha + \beta_h x_t + \boldsymbol{\gamma}'\boldsymbol{w}_t + v_{t+h}$$

这便是*局部投影方程*。当 x_t 满足外生条件 $\mathbb{E}(x_t v_{t+h})=0$ 时，就可以使用 OLS 得到 β_h 的一致估计量，即 LP 的 h 期脉冲响应估计量。

2 LP 和 VAR 的联系

自 Sims (1980) 以来，向量自回归 (vector autoregression, VAR) 模型成为估计脉冲响应的标准范式，它将宏观经济变量归纳为一个动力系统，所有的变量都是内生的。这个范式如今依旧是宏观经济研究的主流方法，但就估计脉冲响应而言，它并非唯一选择，也不是在任何情况下都最好的选择，LP 就有一些独特的优势。我们先介绍 LP 和 VAR 的联系。

在结构上，VAR 和 LP 的关系，好比一般均衡和局部均衡的关系，前者描述了所有变量在所有 horizon 上的动态演进，而后者只是归纳了某一个变量在每个特定 horizon 上同其他变量（和其历史）的动态关系❷。所谓「局部」就是特定的意思，局部投影即将特定 horizon 的结果变量投影到冲击变量上。

❷ 在这个意义上，局部投影又相当于 reduced form，而 VAR 是 structural form。

2.1 由 VAR 导出 LP

当数据生成过程确为 VAR 模型时，我们事实上可以从 VAR 模型推导出 LP 方程。假设有 n 个内生变量，写成一个向量 $\boldsymbol{y}_t = (y_{1t}, \dots, y_{nt})'$，它们服从一个 VAR($p$) 过程：

$$\boldsymbol{y}_t = \sum_{\ell=1}^{p} \boldsymbol{A}_\ell \boldsymbol{y}_{t-\ell} + \boldsymbol{u}_t \quad t = 1, \dots, T$$

为简化讨论，我们暂不考虑结构型 VAR (SVAR)，假设 $\boldsymbol{u}_t$ 就是结构冲击，并设 $\boldsymbol{\Sigma}_u := \mathbb{E}[\boldsymbol{u}_t \boldsymbol{u}_t'] = \boldsymbol{I}_n$。

事实上，任意 VAR(p) 模型都重写为 VAR(1) 的形式。定义

$$\boldsymbol{Y}_t = \begin{bmatrix} \boldsymbol{y}_t \\ \vdots \\ \boldsymbol{y}_{t-p+1} \end{bmatrix} \quad \underset{(np \times np)}{\boldsymbol{\Phi}} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_{p-1} & \boldsymbol{A}_p \\ \boldsymbol{I}_n & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_n & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I}_n & \boldsymbol{0} \end{bmatrix} \quad \boldsymbol{U}_t = \begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix}$$

不难验证

$$\boldsymbol{Y}_t = \boldsymbol{\Phi} \boldsymbol{Y}_{t-1} + \boldsymbol{U}_t$$

这被称为 VAR(p) 模型的伴随形式 (companion form)。因此，通常我们只需关注 VAR(1)。迭代此式可得

$$\boldsymbol{Y}_{t+h} = \boldsymbol{\Phi}^h \boldsymbol{Y}_t + \sum_{\ell=0}^{h-1} \boldsymbol{\Phi}^\ell \boldsymbol{U}_{t+h-\ell}$$

定义 $\boldsymbol{J}_k := \boldsymbol{e}_k' \otimes \boldsymbol{I}_n$❸，其中 $\boldsymbol{e}_k$ 是第 k 个 p 维单位向量，$\otimes$ 表示 Kronecker 积，$\boldsymbol{I}_n$ 是 n 维单位矩阵。对这个式子左乘 $\boldsymbol{J}_1$，得到

❸ 例如，$\boldsymbol{J}_1 := [\boldsymbol{I}_n, \boldsymbol{0}_{n \times n(p-1)}]$，表示选择 $\boldsymbol{Y}_{t+h}$ 的前 n 个分量。

$$\begin{aligned} \boldsymbol{y}_{t+h} &= \boldsymbol{J}_1 \boldsymbol{\Phi}^h \boldsymbol{Y}_t + \sum_{\ell=0}^{h-1} \boldsymbol{J}_1 \boldsymbol{\Phi}^\ell \boldsymbol{U}_{t+h-\ell} \\ &= \boldsymbol{J}_1 \boldsymbol{\Phi}^h (\boldsymbol{J}_1' \boldsymbol{y}_t + \boldsymbol{J}_2' \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{J}_p' \boldsymbol{y}_{t-p+1}) + \sum_{\ell=0}^{h-1} \boldsymbol{J}_1 \boldsymbol{\Phi}^\ell \boldsymbol{J}_1' \boldsymbol{u}_{t+h-\ell} \\ &= \boldsymbol{J}_1 \boldsymbol{\Phi}^h \boldsymbol{J}_1' \boldsymbol{y}_t + \sum_{k=1}^{p-1} \boldsymbol{J}_1 \boldsymbol{\Phi}^h \boldsymbol{J}_{k+1}' \boldsymbol{y}_{t-k} + \sum_{\ell=0}^{h-1} \boldsymbol{J}_1 \boldsymbol{\Phi}^\ell \boldsymbol{J}_1' \boldsymbol{u}_{t+h-\ell} \end{aligned}$$

从这个式子可以看出，对于来自 u_{jt} 的一单位冲击，在 ℓ 期后造成 $y_{i,t+\ell}$ 的响应，等于 $\boldsymbol{J}_1\boldsymbol{\Phi}^\ell\boldsymbol{J}_1'$ 的第 i 行第 j 列的元素。注意到 $\boldsymbol{y}_t$ 的系数矩阵恰为 $\boldsymbol{J}_1\boldsymbol{\Phi}\boldsymbol{J}_1'$，因此，欲得到 $y_{i,t+h}$ 对 $\boldsymbol{u}_t$ 的脉冲响应，只需以 $y_{i,t+h}$ 对 $\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \dots, \boldsymbol{y}_{t-p+1}$ 做线性回归，所得 $\boldsymbol{y}_t$ 的回归系数即 h 期脉冲响应。这便是局部投影。

2.2 LP 和 VAR 在总体上的等价性

事实上，LP 和 VAR 具有更广泛意义上的等价性，这种等价并不依赖于数据的生成过程必须为 VAR 模型。这个性质由 Plagborg-Møller and Wolf (2021) 做了正式讨论，这里给出其主要结论。

假设数据 $\boldsymbol{w}_t = (\boldsymbol{r}_t', x_t, y_t, \boldsymbol{q}_t')'$，其中 $\boldsymbol{r}_t$ 是 n_r 维向量，$\boldsymbol{q}_t$ 是 n_q 维向量，x_t 和 y_t 都是标量。我们想要估计 y_t 对 x_t 的脉冲响应，$\boldsymbol{r}_t$ 和 $\boldsymbol{q}_t$ 将作为控制变量（可以为空）❹。

假设 1. *数据 $\{\boldsymbol{w}_t\}$ 协方差平稳且不能被其历史完全预测（存在一个创新成分使之具有不可预测的随机性），此外有一个处处非奇异的谱密度矩阵和绝对可加的 Wold 分解。*

假设 2. *$\{\boldsymbol{w}_t\}$ 是一个联合高斯过程*❺。

LP 的脉冲响应来自如下投影方程：

$$y_{t+h} = \mu_h + \beta_h x_t + \boldsymbol{\gamma}_h'\boldsymbol{r}_t + \sum_{\ell=1}^{\infty}\boldsymbol{\delta}_{h,\ell}'\boldsymbol{w}_{t-\ell} + \xi_{h,t} \tag{1}$$

脉冲响应就是 $\{\beta_h\}_{h\geq 0}$❻。换言之

$$\beta_h = \mathbb{E}(y_{t+h}|x_t = 1, \boldsymbol{r}_t, \{w_\tau\}_{\tau<t}) - \mathbb{E}(y_{t+h}|x_t = 0, \boldsymbol{r}_t, \{w_\tau\}_{\tau<t})$$

另一方面，我们有（简约型）VAR(∞) 投影方程：

$$\boldsymbol{w}_t = \boldsymbol{c} + \sum_{\ell=1}^{\infty}\boldsymbol{A}_\ell\boldsymbol{w}_{t-\ell} + \boldsymbol{u}_t \tag{2}$$

其中 $\boldsymbol{u}_t := \boldsymbol{w}_t - \mathbb{E}[\boldsymbol{w}_t|\{w_\tau\}_{\tau<t}]$ 是投影残差。我们可以考虑 SVAR 的递归识别，令 $\boldsymbol{\Sigma}_u := \mathbb{E}[\boldsymbol{u}_t\boldsymbol{u}_t']$ 并对其进行 Cholesky 分解 $\boldsymbol{\Sigma}_u = \boldsymbol{B}\boldsymbol{B}'$，于是 SVAR 可以写成：

$$\boldsymbol{A}(L)\boldsymbol{w}_t = \boldsymbol{c} + \boldsymbol{B}\boldsymbol{\eta}_t$$

其中 $\boldsymbol{A}(L) := \boldsymbol{I} - \sum_{\ell=1}^{\infty}\boldsymbol{A}_\ell L^\ell$，$\boldsymbol{\eta}_t := \boldsymbol{B}^{-1}\boldsymbol{u}_t$。定义滞后多项式 $\sum_{\ell=0}^{\infty}\boldsymbol{C}_\ell L^\ell = \boldsymbol{C}(L) := \boldsymbol{A}(L)^{-1}$。于是，$y_t$ 对 x_t 的脉冲响应为

$$\theta_h := \boldsymbol{C}_{n_r+2,\bullet,h}\boldsymbol{B}_{\bullet,n_r+1}$$

这里 $\boldsymbol{C}_{i,\bullet,h}$ 是 $\boldsymbol{C}_h$ 的第 i 行，$\boldsymbol{B}_{\bullet,j}$ 是 $\boldsymbol{B}$ 的第 j 列。

命题 1. *若假设 1 和 2 成立，则 LP 和 VAR 导致的脉冲响应是等价的：对每个 $h = 0, 1, 2, \dots$，有 $\theta_h = \sqrt{\mathbb{E}(\tilde{x}_t^2)}\beta_h$，其中 $\tilde{x}_t := x_t - \mathbb{E}(x_t|\boldsymbol{r}_t, \{w_\tau\}_{\tau<t})$。*

需要注意，假设 1 和 2 并未对实际的数据生成过程做出任何参数化或线性的要求，LP 和 VAR 在这里只是两种线性投影方式，并不必然是数据的真实生成过程。此外，LP 方程 (1) 和 VAR 方程 (2) 都包含了数据的所有历史

❹ 这种排序是规定了 VAR 的递归识别，x_t 排在 y_t 之前，这在 Cholesky 分解时能意味着 x_t 的冲击可以当期影响 y_t，反之则不能。如果 x_t 排在 y_t 之后，则意味着 x_t 的冲击不能当期影响 y_t，那么在 (1) 中就还必须控制当期 y_t。r_t 是那些可以当期影响 x_t 的变量，因此它当期值必须被纳入控制变量。q_t 和 y_t 的排序并不重要。

❺ 这个假设只为了简化记号，将线性投影表示为条件期望，因为高斯变量之间的线性投影就是条件期望。没有这个假设也能证明等价性。

❻ 根据 FWL 定理，β_h 可由 y_{t+h} 对残差 $\tilde{x}_t := x_t - \text{proj}(x_t|\boldsymbol{r}_t, \{w_\tau\}_{\tau<t})$ 做线性回归得到，而残差 $\tilde{x}_t$ 是 x_t 的新息成分，代表了冲击。

信息，因而有别于实际应用中通常使用的有限滞后的估计方程。命题 1 是在说，只要控制了全部的历史信息，LP 投影和 VAR 投影得到的脉冲响应是等价的，只存在一个常数倍数的差别[7]。

还需要注意，这个结论只是表明了两种投影所得脉冲响应在总体上的等价性，并不意味着样本估计结果会相同。关于样本估计结果的差异，我们将在下一节讨论。

Plagborg-Møller and Wolf (2021, Proposition 2) 还论证了有限维滞后 LP 和 VAR 的等价性：对于 $h \leq p$，LP 和 VAR 投影所得脉冲响应是等价的，但对于 $h > p$，两者可能相异。

一般认为，除了直截了当的估计形式，局部投影的优势在于对模型误设的稳健性。换言之，其得到的脉冲响应估计量相比 VAR 模型有更小的偏误，但代价是更大的方差。当 VAR 模型是真实的数据生成过程时，LP 估计量的方差会显著高于 VAR 模型（尽管仍然是一致的），在推断中，这表现为更宽的置信区间。这是一种 bias-variance trade-off，LP 通过牺牲效率换取稳健性，后者的重要性常常高于效率，所以 LP 越来越受到实证研究者的青睐。

[7] 这个常数倍数存在的原因是，VAR 中 x_t 的新息成分 $\eta_{x,t}$（一单位冲击）被标准化为单位方差，而 LP 中的一单位冲击（即 $\tilde{x}_t$）并不一定有单位方差。

3 LP 的稳健性

在数值模拟中，可以更清楚地观察到 LP 和 VAR 之间的 bias-variance trade-off。我们用一个简单的实验来展示。设数据服从一个 ARMA(1,1)：

$$y_t = \rho y_{t-1} + \varepsilon_t + \alpha \varepsilon_{t-1}$$

其中 $\varepsilon_t \sim$ i.i.d. $\mathcal{N}(0,1)$，$\rho = 0.85$，$\alpha = 0.1$。这里 α 表示对 AR(1) 模型的轻度偏离。换言之，我们想考察当真实模型并非 AR(1) 时，LP 和 AR(1) 模型估计的脉冲响应的差异。

取数据长度 $T = 250$，分别用 LP 和 AR(1) 模型估计 $h = 2$ 的脉冲响应[8]。图 1 展示了两种估计量的分布。可以看到，LP 估计量的分布更宽，但更加围绕真实值对称，而 AR(1) 估计量的分布更窄，但更偏离真实值。这表明了，在模型误设下，LP 比 AR 的偏误更小，但方差更大。

[8] 真实值为 $\theta_h = \rho^h + \alpha\rho^{h-1}$。

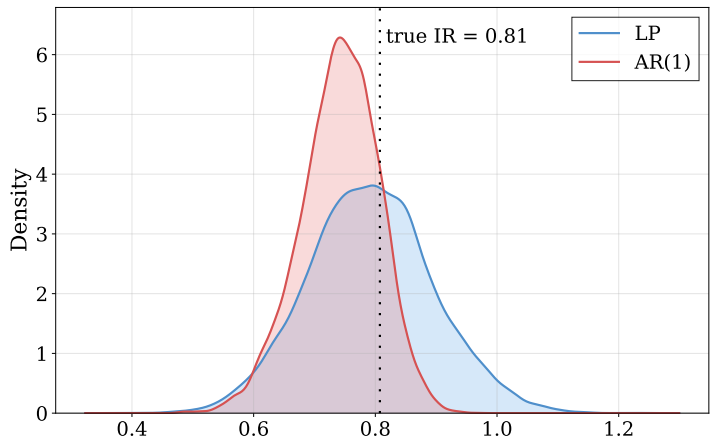

图 1: LP 和 AR(1) 模型估计的脉冲响应估计量的分布

Li et al. (2024) 做了大量更细致的数值模拟，结果表明（LP 和 VAR 控制相同的滞后项数），对于短期脉冲响应，两者差异较小，但对中远期脉冲响应，

LP 有更小的偏误和更大的方差，而 VAR 则相反。Montiel Olea et al. (2024) 则从理论上建立了 LP 对局部模型误设的稳健性：当真实模型是一个带有 MA 项的 VARMA 模型时，LP 估计量的置信区间仍具有良好的覆盖概率，而使用误设的 VAR 所得置信区间则存在显著的覆盖概率不足，换言之，LP 仍然可以被用于对脉冲响应的有效推断，而 VAR 则不能。

在实际应用的有限样本中，两者的估计量更是可能出现显著差异。利用 Ramey (2016) 中的多个实证例子，Montiel Olea et al. (2025, Section 3.1) 计算了 LP 和 VAR 的脉冲响应估计量及其标准误，发现两者的估计量差异很大，特别是在远期脉冲响应上，而 VAR 脉冲响应的标准误常常比 LP 更小。这意味着两种方法可能导致截然不同的实证结论。一般而言，不推荐使用短滞后的 VAR 模型，如果模型必须精简，通常推荐 LP，如果不限制滞后期数，则 VAR 往往可以给出更精确的估计。

4 LP 的推断

在实证研究中，除了获得脉冲响应的点估计，更重要的还是对估计量进行推断，即构造置信区间。这里的推断指的是逐点推断 (pointwise inference)，即对每个 horizon 的脉冲响应估计量构造置信区间[9]。

[9] 相对应的是联合推断 (joint inference)，即对整个脉冲响应函数构造置信区间。

LP 本质上就是 OLS，因而其推断也（理应）基于 OLS 的推断理论。只需要计算 LP 估计量的标准误，然后给估计量加减 1.96 倍的标准误，就可以得到 95% 的置信区间。这再简单不过了。但问题是，我们应该采用什么标准误？Jordà (2005) 的推荐是使用 HAC (heteroskedasticity and autocorrelation consistent) 稳健标准误，如 Newey-West 标准误，原因是 LP 的误差项一般是自相关的。

多数情况下这已经足够了，但在某些情况下，事情并没有这么简单。首先是数据中可能存在的非平稳性，此时寻常的推断方法会失效。对此，Montiel Olea and Plagborg-Møller (2021) 建议使用「滞后扩充」(lag-augmented) 回归，然后使用 White 异方差稳健标准误就够了。另外，在有限样本中，Herbst and Johannsen (2024) 发现，LP 的估计量在小样本（如 $T < 200$ 或更小）中存在偏误，足以使采用 HAC 稳健标准误的推断失效。

5 小结

本文简要介绍了局部投影估计的原理，和 VAR 的联系，以及推断的一些问题。自然，这只是关于 LP 的一小部分，其他主题如和工具变量的结合、正则化估计、面板数据、与因果推断的联系等都未有涉及。欲更深入了解 LP 的方方面面，Jordà and Taylor (2025) 和 Montiel Olea et al. (2025) 是不得不读的两篇综述。

最后，我们梳理一下实证研究中关于如何选择 LP 和 VAR 的建议[10]。如果样本量较小，推荐使用 VAR 和 Herbst and Johannsen (2024) 的误差纠正 LP；如果样本量不是担心的问题，推荐使用 LP 或者滞后项足够多的 VAR；如果存在非平稳性，则需要使用滞后扩充的 LP。

[10] 不过通常的做法是两者均报告在研究中。

参考文献

Herbst, E. P. and Johannsen, B. K. (2024): "Bias in Local Projections," *Journal of Econometrics*, 240(1), 105655.

Jordà, Ò. (2005): "Estimation and Inference of Impulse Responses by Local Projections," *American Economic Review*, 95(1), 161–182.

Jordà, Ò. and Taylor, A. M. (2025): "Local Projections," *Journal of Economic Literature*, 63(1), 59–110.

Li, D., Plagborg-Møller, M., and Wolf, C. K. (2024): "Local Projections vs. VARs: Lessons from Thousands of DGPs," *Journal of Econometrics*, 244(2), 105722.

Montiel Olea, J. L. and Plagborg-Møller, M. (2021): "Local Projection Inference is Simpler and More Robust Than You Think," *Econometrica*, 89(4), 1789–1823.

Montiel Olea, J. L., Plagborg-Møller, M., Qian, E., and Wolf, C. K. (2024): "Double Robustness of Local Projections and Some Unpleasant VARithmetic," *arXiv preprint*, arXiv: 2405.09509.

Montiel Olea, J. L., Plagborg-Møller, M., Qian, E., and Wolf, C. K. (2025): "Local Projections or VARs? A Primer for Macroeconomists," *arXiv preprint*, arXiv: 2503.17144.

Plagborg-Møller, M. and Wolf, C. K. (2021): "Local Projections and VARs Estimate the Same Impulse Responses," *Econometrica*, 89(2), 955–980.

Ramey, V. A. (2016): "Macroeconomic Shocks and Their Propagation," in *Handbook of Macroeconomics*, Vol. 2, ed. by J. B. Taylor and H. Uhlig, Chapter 2, 71–162.

Sims, C. A. (1980): "Macroeconomics and Reality," *Econometrica*, 48(1), 1–48.