Rejoinder

Michal Kolesár
Princeton University

Mikkel Plagborg-Møller
Princeton University

July 15, 2025

We thank the discussants for their thoughtful and stimulating comments. The goal of our paper was to highlight a robustness wedge: vector autoregression (VARs) and linear local projections onto observed shocks or proxies (LLPs) are robust to nonlinearities in the data generating process (DGP) in that they deliver a meaningful causal summary—a convex average of marginal effects—regardless of the extent of nonlinearities. By contrast, identification approaches that exploit heteroskedasticity or non-Gaussianity of latent shocks are fragile: they do not generally deliver a meaningful causal summary once the DGP is nonlinear. We were gratified to see that virtually all discussion focused on interpreting and extending our positive results pertaining to LLPs (and VARs, though we will henceforth refer only to LLPs for brevity). While absence of evidence is not evidence of absence, we hope this is a signal that we were at least partially successful in our goal.

This rejoinder addresses three issues raised in the discussions. First, is the LLP estimand indeed a sufficiently interesting causal summary? Second, are negative weights always a problem? Third, can our results be extended to cover non-stationarity DGPs?

1 Is "good" good enough?

All four discussions raised the point that we were perhaps not sufficiently ambitious in our goal: while it is good that LLPs identify an average marginal effect, this fact alone may not be a strong enough argument to use them, for two reasons. First, all discussants questioned the empirical usefulness of average marginal effects. Second, Herbst and Johannsen (HJ) and Gonçalves, Herrera, and Pesavento (GHP) both raised the point that LLPs use a weighting function that differs from the historic shock density, which they argued may provide a more natural weighting.

To evaluate these critiques, one first has to define the goal of the exercise. Most commonly, as Jordà's comment discusses, impulse responses are used to validate or (quantitatively or qualitatively) calibrate structural macro models. For this goal, it is crucial to know the weighting function, so that we can weight marginal effects in the model using the same weights. As we show in the paper, LLPs have the advantage that the weighting function can easily be estimated using regression methods; its estimation is much easier than that of the historical shock density f_X, for instance. The convex weighting ensures we get the sign right if the marginal effect is uniformly positive or negative, which is useful when the calibration is qualitative. But robustness is not limited to the sign: convexity also ensures that if the marginal effects $\Psi_h'(x)$ all lie in some bounded interval $[\underline{\theta}, \overline{\theta}]$, then the LLP estimand will also lie in that interval. This implies that if the DGP is only moderately non-linear, so that the interval $[\underline{\theta}, \overline{\theta}]$ is not too wide, we automatically get the *magnitude* of the causal effects broadly right. The other important property of this estimand is that it can be estimated relatively precisely even in small samples, and its estimation doesn't involve tuning parameters; noisy or sensitive estimates make for poor calibration targets.[1] In our view, these properties lend theoretical support to the common practice of using LLPs for validation of structural models.

Should LLPs be the only tool? Clearly, exploring nonlinearities in the data can be informative and provide further insights. The results table of any careful empirical analysis typically features multiple columns; researchers don't just report a single specification. In our view, the properties of LLPs are compelling enough that they make a good candidate for the first column; the other columns can explore how the results change with alternative weighting schemes or nonlinear specifications. Our results in Section 6 discuss how to build weighting-based and doubly robust estimators targeting any alternative weighting scheme; one should just be careful to examine their sensitivity to functional form and tuning parameters. Jordà and HJ suggest reporting state-dependent impulse responses: our framework easily accommodates this by conditioning on states. As we discuss at the end of Section 3.1, this can be implemented either via full interactions with state dummies or by estimation on subsamples.[2]

Another way of exploring nonlinearities is to estimate the effects of large shocks. GHP

[1] Goldsmith-Pinkham, Hull, and Kolesár (2024) show, building on the work of Crump, Hotz, Imbens, and Mitnik (2006), that for a binary treatment, linear regression targets the easiest-to-estimate treatment effect. It would be interesting to know whether the average marginal effect identified by LLPs shares this property.

[2] Jordà questions whether state dependence may break down the equivalence between VARs and local projections. Provided the state variable is discrete, the equivalence still holds since we can simply estimate the VAR or local projection on a subsample; Section 3.1 discusses this point when the state is binary.

motivate this by empirical applications in which the researcher is interested in estimating the response of GDP growth to oil price shocks that exceed one standard deviation, and suggest estimating average causal effects. In their conclusion, GHP posit that to examine the impact of large shocks, it may be necessary to resort to nonparametric estimation or a flexible parametric model. We agree that such empirical applications are interesting. However, average causal effects identify the effect of shifting the shock distribution by a given amount, such as *exactly* one standard deviation, not the effect of shocks *exceeding* one standard deviation. The latter is simply estimated using our framework by running a regression on a binary indicator that the shock exceeds one standard deviation. To obtain finer comparisons within our framework, we can discretize the shock into more than two groups (e.g., by rounding a monetary shock to the nearest 25 basis points) and run a regression on these dummies: it is not necessary to resort to full nonparametric estimation.

In addition to validation of structural models, a second possible use of impulse responses is to identify particular causal effects, or, more ambitiously, directly inform policy. Kitagawa, Wang, and Xu (KWX) suggest directly formulating the policy choice problem, and using it to form the estimation target. In the current context, this would be an ambitious exercise since most policies will have general equilibrium effects that the average structural function does not capture (the so-called Lucas critique). Barnichon and Mesters (2023) and McKay and Wolf (2023) show that impulse response functions can be used as inputs into evaluating counterfactuals and optimal policy rules in certain classes of linear structural macroeconomic models, but their results heavily leverage the linearity assumption and the resulting certainty equivalence. To our knowledge, it is an open question how to tackle the Lucas critique in nonlinear settings without imposing a complete structural model.

Let us therefore consider the less ambitious goal of identifying causal effects. Such causal effects (impulse responses) are policy relevant in the specific context where policy-makers are considering whether to surprise the public with a one-off policy intervention, and by how much. For concreteness, consider the example from GHP's discussion: what is the magnitude of the government spending multiplier and does it exceed unity? As alternatives to the LLP weighting, GHP consider targeting the average marginal effect weighted by the historical shock density f_X and estimating the average causal effect when we shift the shock distribution by a given fixed amount.

As we discuss in the paper, average causal effects can be represented as average marginal effects, but using a different weighting function. For example, reporting the average causal effect of a one standard deviation government spending shock, relative to a baseline of no shock, is tantamount to choosing a particular weighting of marginal effects. Thus, for estimating a particular causal effect, such as the government spending multiplier, one needs to choose how to weight the marginal effects: there is no single "government spending multiplier" when the average structural function is nonlinear, so one cannot ask what is the magnitude of the marginal effects without specifying the weighting function. In the cross-sectional binary treatment effect literature, the average treatment effect is a natural target, comparing the world in which everyone receives treatment with one in which nobody does. Since in the presence of covariates, regression does not generally yield the average treatment effect, it is commonly argued that regression is "biased" for the average treatment effect (e.g., Aronow and Samii, 2016). HJ and GHP both argue that LLPs are similarly biased relative to the average marginal effect that uses the historical shock density f_X as weights.[3]

But with a continuously distributed treatment or a shock, it is less clear to us why the density f_X should be the default weighting. If the impulse responses are to be used to inform policymaking going forward, there is no reason to privilege the historical shock distribution. For this reason, it seems reasonable to report results for alternative weighting schemes in different columns of the results table to see how they impact the magnitude of the effect—say, whether the government spending multiplier exceeds unity. Given the properties of LLPs, they make a good candidate for the first column. Our proposal of reporting the estimated weight function allows the reader to gauge whether the first column is relevant for their specific decision problem.

One factor that should influence the choice of estimand is the variance of the estimator. Linear local projections can be viewed as a local linear kernel regression, as used by GHP to estimate the density-weighted average marginal effect, but with a bandwidth set to infinity. Since the first-order effect of increasing the bandwidth is to decrease the variance of the estimator, we expect linear local projections to typically be more precise than nonparametric local linear alternatives.

One limitation of our analysis is that we abstract from finite-sample issues in order to establish clean identification results. HJ show, using simulations, that when the data is very persistent, LLPs can be biased and consequently subject to coverage distortions. We agree that this concern merits serious attention in empirical work. In the context of linear DGPs, Montiel Olea, Plagborg-Møller, Qian, and Wolf (2025) find that the coverage distortions

[3] GHP make this point using simulations with highly heavy-tailed or skewed shocks. We remark that in the specification in panel (a) of Figure 4, the shock x_t is so heavy-tailed that the expectation $E[y_t x_t]$ does not exist, so it is unclear whether the bias of the LLP estimator is well-defined in that DGP.

of LLPs can be greatly ameliorated by employing a residual block bootstrap based on an auxiliary VAR (Brüggemann, Jentsch, and Trenkler, 2016), combined with analytical bias correction (Herbst and Johannsen, 2024). Alternatively, the persistence of the data could be reduced by differencing, as suggested by Piger and Stockwell (2025). It would be worthwhile investigating whether these fixes are equally successful in nonlinear DGPs.

2 Should we worry about negative weights?

KWX point out that the presence of negative weights $\omega(x)$ in an estimand is not necessarily a cause for alarm. Formally, their comment shows that if each point x where ω is negative can be matched with a unique point $Q(x)$ such that (i) the weight $\omega(Q(x))$ is positive and larger in magnitude than $\omega(x)$ and (ii) the marginal effect g' takes on the same value at x and $Q(x)$, then we can simply zero out the negative weight at x by subtracting it off the positive weight at $Q(x)$, obtaining a new weighting function that places zero weight over the region where ω is negative and is non-negative otherwise.

This is a very nice result! Thinking of points x where ω is negative as "defier" points and points x where ω is positive as "complier" points, the conditions (i) and (ii) generalize the "more compliers than defiers" condition of de Chaisemartin (2017), made in the context of an instrumental variables regression with a binary treatment and a binary instrument when the Imbens and Angrist (1994) monotonicity condition is violated. We expect this generalization to be useful not just for salvaging estimands with negative weights in the settings we focused on, but also well beyond it, such as alleviating the negative weighting issue in two-way fixed effects regressions (e.g., de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021).

Does this result diminish the value of knowing that LLPs deliver positive weights? In our view, no: it is clearly useful to know that LLPs deliver a meaningful causal summary regardless of the shape of the marginal effect function. We agree with the broader point of KWX that positive weights are sufficient, but not always necessary for an estimand to provide a meaningful causal summary once one is willing to place restrictions on the marginal effect function. But for alternative estimands where the weights may be negative, reasonable restrictions are necessarily context-dependent, so that one needs to consider KWX's conditions (i) and (ii) on a case-by-case basis. Are these conditions applicable in the context of identification via heteroskedasticity and non-Gaussianity? Although KWX do not offer examples, perhaps there exist some, and we hope future work will lay them out. However, as several analytical counterexamples in our paper show, these particular identification strategies cannot generally be salvaged by placing weak restrictions on the DGP.

3 Allowing for non-stationarity

KWX raise the interesting question of whether our results on the properties of LLPs extend to models with structural equations that are unstable over time. A first remark is that our baseline results already allow for *stationary* time-varying parameters, such as in the stochastic regime switching model in Example 1 in our paper. Non-stationarity in the form of stochastic trends caused by unit roots is also easily accommodated: simply replace the trending outcome Y_{t+h} in the local projection with the stationary long difference $Y_{t+h} - Y_{t-1}$ (equivalently, we can control for the lagged outcome on the right-hand side). Our results should then go through if we use the long-differenced outcome in the definition of the average structural function.

As hinted by KWX, we can also follow Casini and McCloskey (2024) and extend our results to environments with smooth deterministic time-variation. Assume that the DGP is *locally* stationary (Dahlhaus, 2012): for sufficiently large sample size T, the data (Y_{t+h}, X_t) can be well-approximated by a triangular array $(Y_{t+h}(t/T), X_t(t/T))$, where at each fractional time point $\tau \in [0, 1]$ the approximating process $\{Y_{t+h}(\tau), X_t(\tau)\}_{t\in\mathbb{Z}}$ is stationary, but the law of motion is allowed to change smoothly as a function of τ. Abstracting from covariates and assuming $E[X_t(\tau)] = 0$ for notational simplicity, the LLP estimator will under regularity conditions satisfy

$$\hat{\beta}_h \equiv \frac{\sum_{t=1}^{T-h} Y_{t+h} X_t}{\sum_{t=1}^{T-h} X_t^2} \xrightarrow{p} \frac{\int_0^1 \operatorname{Cov}(Y_{t+h}(\tau), X_t(\tau))\, d\tau}{\int_0^1 \operatorname{Var}(X_t(\tau))\, d\tau} \equiv \beta_h \quad \text{as} \quad T \to \infty,$$

as discussed by Casini and McCloskey (2024). If we define the local conditional expectation $g_h(x, \tau) \equiv E[Y_{t+h}(\tau) \mid X_t(\tau) = x]$, it is an immediate corollary of Proposition 1 in our paper that

$$\beta_h = \int_0^1 \int \omega_X(x, \tau) g_h'(x, \tau)\, dx\, d\tau, \quad \text{where} \quad \omega_X(x, \tau) \equiv \frac{\operatorname{Cov}(\mathbb{1}\{X_t(\tau) \geq x\}, X_t(\tau))}{\int_0^1 \operatorname{Var}(X_t(\tau))\, d\tau},$$

and $g_h'(x, \tau)$ denotes the partial derivative with respect to x. Here we require that the regularity conditions of the proposition hold for $(Y_{t+h}(\tau), X_t(\tau))$ at all $\tau \in [0, 1]$ and that the above integrals exist. Thus, in the presence of smooth deterministic time-variation in the DGP, LLPs continue to estimate a convex weighted average of marginal effects, but we now additionally average over fractional time τ. We conjecture that analogous results obtain for instrumental variable estimators, including in settings with controls.

References

Aronow, P. M., and C. Samii (2016): "Does Regression Produce Representative Estimates of Causal Effects?," *American Journal of Political Science*, 60(1), 250–267.

Barnichon, R., and G. Mesters (2023): "A Sufficient Statistics Approach for Macro Policy," *American Economic Review*, 113(11), 2809–45.

Brüggemann, R., C. Jentsch, and C. Trenkler (2016): "Inference in VARs with conditional heteroskedasticity of unknown form," *Journal of Econometrics*, 191(1), 69–85.

Casini, A., and A. McCloskey (2024): "Identification and Estimation of Causal Effects in High-Frequency Event Studies," Manuscript, University of Rome Tor Vergata.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2006): "Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand," Working Paper 330, National Bureau of Economic Research, Cambridge, MA.

Dahlhaus, R. (2012): "Locally Stationary Processes," in *Time Series Analysis: Methods and Applications*, ed. by T. Subba Rao, S. Subba Rao, and C. Rao, vol. 30 of *Handbook of Statistics*, chap. 13, pp. 351–413. Elsevier.

de Chaisemartin, C. (2017): "Tolerating Defiance? Local Average Treatment Effects without Monotonicity," *Quantitative Economics*, 8(2), 367–396.

de Chaisemartin, C., and X. D'Haultfœuille (2020): "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110(9), 2964–2996.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2024): "Contamination Bias in Linear Regressions," *American Economic Review*, 114(12), 4015–51.

Goodman-Bacon, A. (2021): "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 225(2), 254–277.

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

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

McKay, A., and C. K. Wolf (2023): "What Can Time-Series Regressions Tell Us About Policy Counterfactuals?," *Econometrica*, 91(5), 1695–1725.

Montiel Olea, J. L., M. Plagborg-Møller, E. Qian, and C. K. Wolf (2025): "Local Projections or VARs? A Primer for Macroeconomists," Working Paper 33871, National Bureau of Economic Research.

Piger, J., and T. Stockwell (2025): "Differences from Differencing: Should Local Projections with Observed Shocks be Estimated in Levels or Differences?," Working Paper 4530799, SSRN.