Recent Developments in Causal Inference with Macro Data

Mikkel Plagborg-Møller

University of Chicago

Based on joint work with:

Michal Kolesár, José Luis Montiel Olea, Eric Qian & Christian K. Wolf

January 5, 2026

Slides: `mikkelpm.com` → Literature Notes

Causal framework in macroeconomics

- Impulse-propagation paradigm: macroeconomists often find it useful to think about how their models propagate "shocks". Frisch (1933); Slutsky (1937)

$$\text{shocks} \underbrace{\longrightarrow}_{\text{propagation}} \text{outcomes}$$

- Shock = surprise changes to an external factor driving the economy, e.g., fundamentals (TFP, household discount factor, ...) or deviations from policy rules.

- Conceptually distinct shocks should be statistically independent. If two kinds of "shocks" were systematically related, there must be a third shock that's causing both.

Impulse responses = dynamic causal effects

- In the impulse-propagation paradigm, central objects of interest are the impulse response functions (dynamic causal effects)

$$\frac{\partial E[y_{t+h} \mid \varepsilon_t = \varepsilon]}{\partial \varepsilon}, \quad h = 0, 1, 2, \dots,$$

  where y_t = outcome of interest, and ε_t = shock.
  - Not a forecast! ε_t may only comprise a small fraction of the overall variance of y_t.
- Focusing attention on the propagation of shocks yields conceptual clarity and ease of matching theory with data.
  - Moreover, under some assumptions, impulse responses can be used to compute policy rule counterfactuals or evaluate policy optimality. Barnichon & Mesters (2023); McKay & Wolf (2023)

Impulse response estimators

- One estimation strategy: take a fully-specified structural model to the data.
  - If we get it right, this produces all the causal effects and counterfactuals we could dream of.
  - But typically requires numerous assumptions that are jointly implausible.
- Two popular methods allow for estimation of impulse responses without imposing a full equilibrium structure *a priori*:
  1. VAR: iterate on flexible dynamic multivariate model. Sims (1980, 22k cites)
  2. LP: direct regression of future outcome y_{t+h} on current covariates. Jordà (2005, 5k cites)
- Which one of these works best? And are these linearity-based procedures ever useful if the real world is nonlinear?

Main references for this talk

- Montiel Olea, Qian, P-M & Wolf (2025a), "Local Projections or VARs? A Primer for Macroeconomists", *NBER Macro Annual*.
- Kolesár & P-M (2025), "Dynamic Causal Effects in a Nonlinear World: the Good, the Bad, and the Ugly", *JBES*.

Outline

Local projection

- LP: linear regression, separately for each horizon $h = 0, 1, 2, \ldots$:

$$y_{t+h} = \mu_h + \theta_h^{\text{LP}} x_t + \gamma_h' r_t + \delta_{h,1}' w_{t-1} + \cdots + \delta_{h,p}' w_{t-p} + \xi_{h,t+h}.$$

  - This is a projection, not a generative model.

- **Shock:** LP estimates impulse response of y_{t+h} with respect to

$$\tilde{x}_t = x_t - \text{proj}(x_t \mid r_t, w_{t-1}, \ldots, w_{t-p}).$$

  Whether this is an interesting object depends on assumptions.

  - E.g., $\tilde{x}_t =$ narrative shock (Romer x 2) or Taylor rule residual (Christiano, Eichenbaum & Evans).

- **Projection:** LP uses autocorrelations in the data out to the horizon h of interest to compute θ_h^{LP}.

Vector autoregression

- VAR estimates reduced-form multivariate dynamic model in $w_t = (r_t', x_t, y_t, \bar{w}_t')'$:

$$w_t = c + A_1 w_{t-1} + A_2 w_{t-2} + \cdots + A_p w_{t-p} + u_t.$$

- Orthogonalize $u_t = H\varepsilon_t$. For now, assume H lower triangular (recursive/Cholesky id'n).

- Structural impulse responses $\Psi_h = \partial w_{t+h} / \partial \varepsilon_t'$ from iterative propagation:

$$\Psi_0 = H, \quad \Psi_1 = A_1 \Psi_0, \quad \Psi_2 = A_1 \Psi_1 + A_2 \Psi_0, \quad \dots \quad \Psi_h = \sum_{\ell=1}^{\min\{p,h\}} A_\ell \Psi_{h-\ell},$$

$$\theta_h^{\text{VAR}} = \partial y_{t+h} / \partial \varepsilon_{x,t} = e_y' \Psi_h e_x.$$

- **Shock:** residual in projection of $u_{x,t}$ on $u_{r,t}$ (elements of u_t). Same as LP shock $\tilde{x}_t$!

- **Projection:** VAR matches first p autocovariances of the data, but extrapolates to longer horizons $h > p$.

Outline

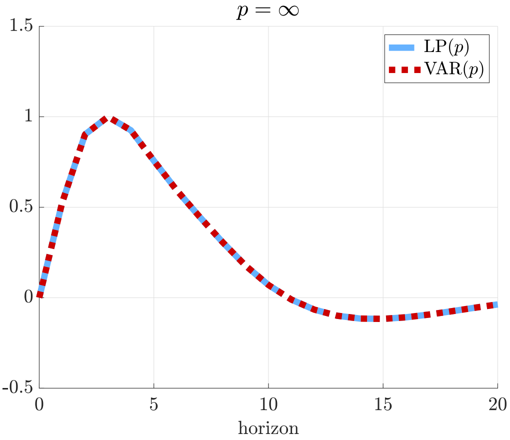
1. LP and VAR
2. Identification
3. Bias-variance trade-off
4. Uncertainty assessments
5. Taking stock
6. Nonlinear environments
7. Recommendations

Equivalence between LP and VAR

- LP & VAR project on the exact same shock.
- But the way they propagate that shock differs: LP directly uses autocorrelations in the data out to horizon h, while VAR extrapolates from the first p autocovariances.
- This suggests that if p is large, the two methods should be equivalent.
- Even if p is not large, the two methods should give similar results at horizons $h \leq p$.

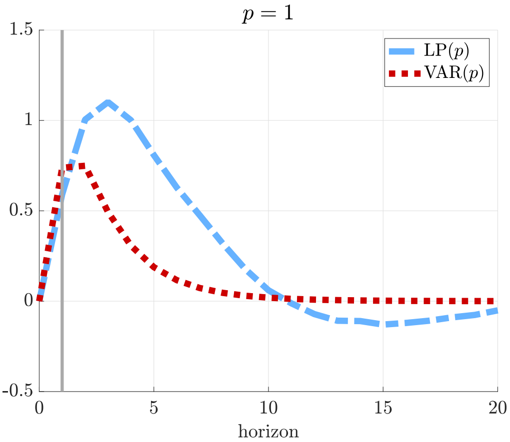
LP = VAR with very long lag length

$p = \infty$: same shock, same projection, so same impulse responses



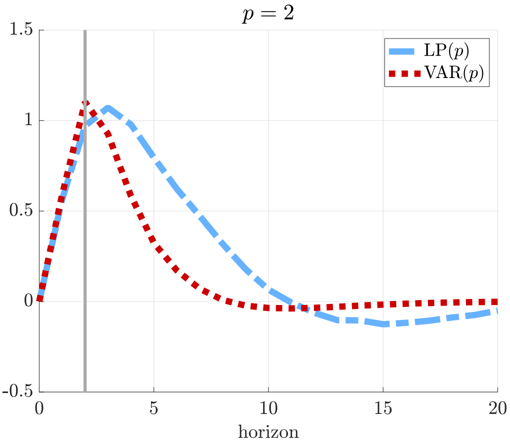
LP $\approx$ VAR up to horizon p

$p < \infty$: same shock so same responses at $h = 0$,
approx'ly same for $0 < h \leq p$, but then VAR extrapolates



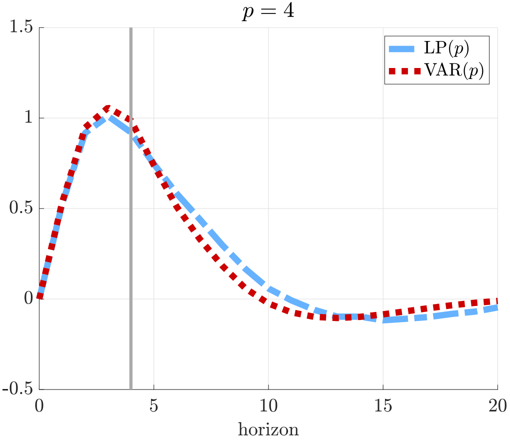
LP $\approx$ VAR up to horizon p

$p < \infty$: same shock so same responses at $h = 0$,
approx'ly same for $0 < h \leq p$, but then VAR extrapolates



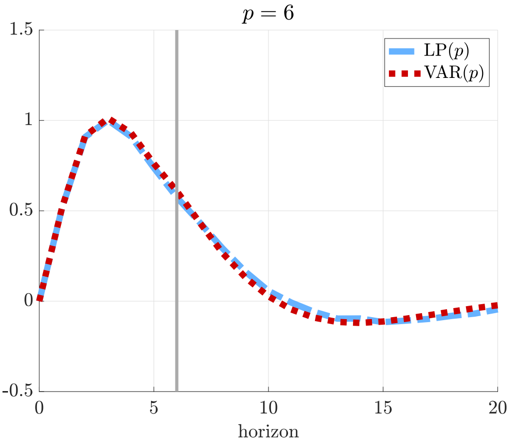
LP $\approx$ VAR up to horizon p

$p < \infty$: same shock so same responses at $h = 0$,
approx'ly same for $0 < h \leq p$, but then VAR extrapolates



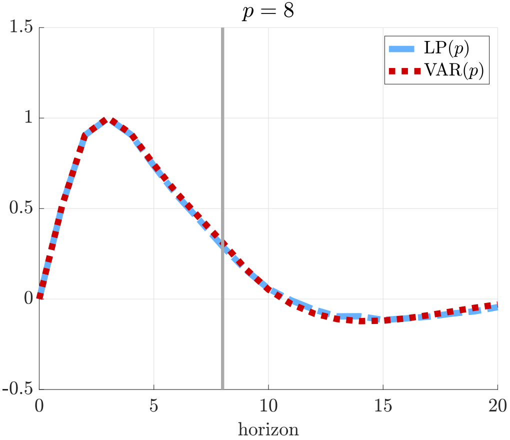
LP $\approx$ VAR up to horizon p

$p < \infty$: same shock so same responses at $h = 0$,
approx'ly same for $0 < h \leq p$, but then VAR extrapolates



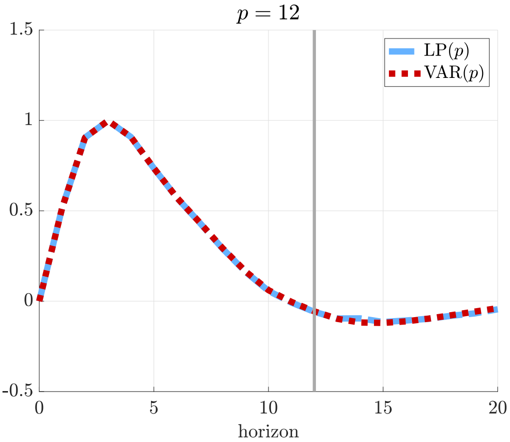
LP $\approx$ VAR up to horizon p

$p < \infty$: same shock so same responses at $h = 0$,
approx'ly same for $0 < h \leq p$, but then VAR extrapolates



LP $\approx$ VAR up to horizon p

$p < \infty$: same shock so same responses at $h = 0$,
approx'ly same for $0 < h \leq p$, but then VAR extrapolates

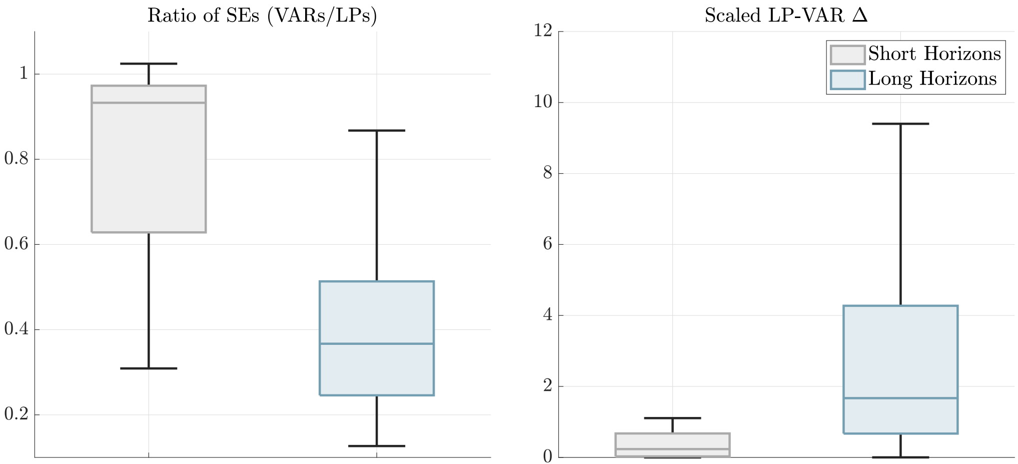


LPs and VARs share the same estimand when p is large

- The previous discussion considered recursive identification (short-run timing restrictions).
- The equivalence between LP & VAR extends to more complicated identification schemes.
  - Proxy/IV, long-run restrictions, sign restrictions, ...
  - Intuition: "shock" is still just some (potentially complicated) f'n of autocovariances of the data. With many lags, both LP and VAR approximate these well in large samples.
- Take-away: LP vs. VAR debate unrelated to questions of identification.
  - Any identification scheme that can be implemented by an SVAR can also be implemented by an LP, and *vice versa*.
  - Only difference is how a finite data set is exploited to estimate the common estimand.

Outline

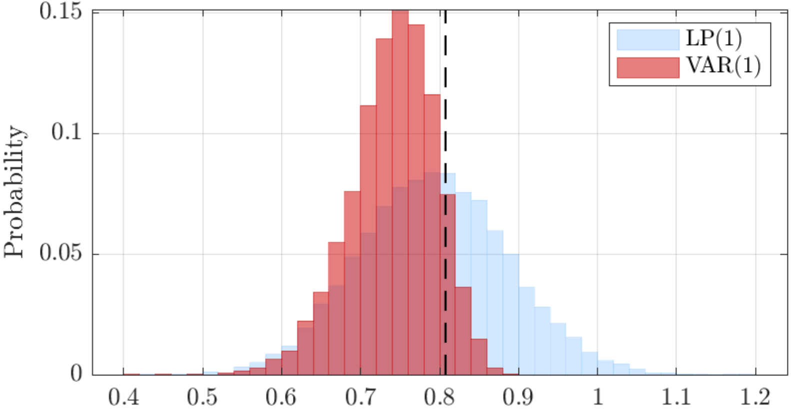
VAR vs. LP in finite samples



Replication of 4 empirical applications in Ramey (2016), total of 385 impulse responses

Illustrative simulation

$$y_t = \rho y_{t-1} + \varepsilon_t + \alpha \varepsilon_{t-1}, \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0,1)$$



$h = 2,\ \rho = 0.85,\ \alpha = 0.1,\ T = 240$

Analytics of the bias-variance trade-off

- Montiel Olea et al. (2025b) consider a structural VAR model contaminated by small MA terms:

$$w_t = A_1 w_{t-1} + \cdots + A_{p_0} w_{t-p_0} + H\left(\varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \ldots\right).$$

  - Why? Low-order VARs are known to forecast well, but not literal truth. Schorfheide (2005)
  - MA terms can arise if we slightly under-specify the lag length, forget to control for a relevant variable, aggregate the data inappropriately, or use data with measurement error.
  - Technically, assume $\alpha_\ell \propto$ std. dev. of VAR estimator.
- In this environment, estimators should control for infinitely many lags. Infeasible.

Analytics of the bias-variance trade-off

- Suppose both LP & VAR use $p \geq p_0$ estimation lags.
- Then in large samples,

$$\hat{\theta}_h^{\text{VAR}} \mathrel{\dot{\sim}} N\left(\theta_h + b_h(p), \tau_{h,\text{VAR}}^2(p)\right), \quad \hat{\theta}_h^{\text{LP}} \mathrel{\dot{\sim}} N\left(\theta_h, \tau_{h,\text{LP}}^2\right).$$

  - Benefit and cost of extrapolation: VAR more efficient ($\tau_{h,\text{VAR}}^2(p) \leq \tau_{h,\text{LP}}^2$) but biased.
  - $h \leq p - p_0$: VAR bias $b_h(p) = 0$ and variance both coincide with LP.
- Both LP & VAR require controlling for the most important predictors/lags! But the result implies that LP is robust to omitting moderately important ones, while VAR is not.
  - LP is relatively robust to model selection errors committed by BIC/AIC. Can use these to select the number of lagged controls.

How bad can the VAR bias be in theory?

- Theoretical bound on bias: letting $\mathcal{M}$ denote the fraction of the variance of the MA residual that's due to lagged shocks,

$$|b_h(p)| \;\leq\; \sqrt{T \times \mathcal{M} \times \left\{\tau^2_{h,\text{LP}} - \tau^2_{h,\text{VAR}}(p)\right\}},$$

  and there exist MA coefficients that attain the bound.

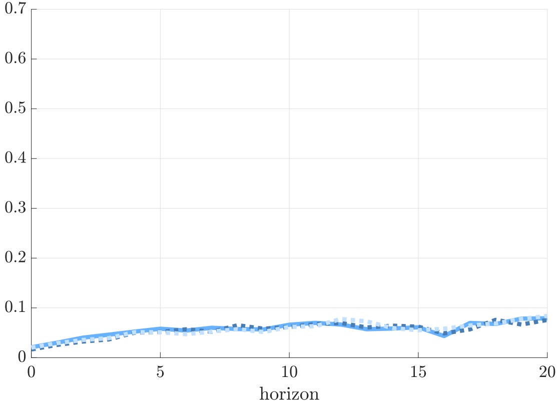
  - Example: if $T = 100$, $\mathcal{M} = 1\%$, $\tau_{h,\text{VAR}}(p)/\tau_{h,\text{LP}} = 0.5$, then bias can be $1.73 \times$ SE.

- No free lunch for VARs: if precision gain is large, then so is the potential bias.
  - VAR only robust if we use so many lags that VAR $=$ LP.

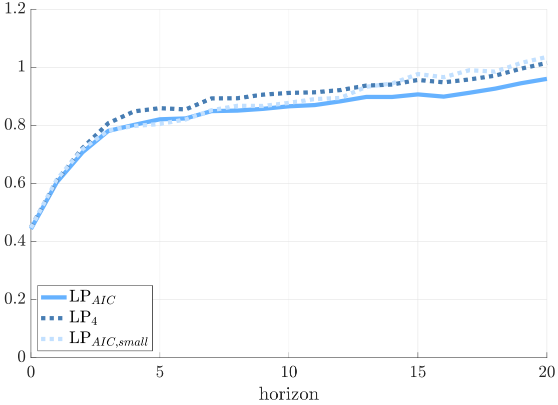
The bias-variance trade-off in practice

- Montiel Olea et al. (2025a) conduct large-scale simul'n study, extending Li et al. (2024).
- DGP: extension of Stock-Watson dynamic factor model fitted to 207 U.S. macro series.
  - Both stationary and non-stationary versions.
- Construct 100s of specifications:
  - Randomly draw subsets of 5 salient macro series from the DFM. Outcome y chosen at random from this list.
  - Additionally, econometrician observes a monetary/fiscal shock.
- Simulate data with $T = 240$, then estimate LPs, VARs, and several variants.

Simulation evidence: bias and standard deviation



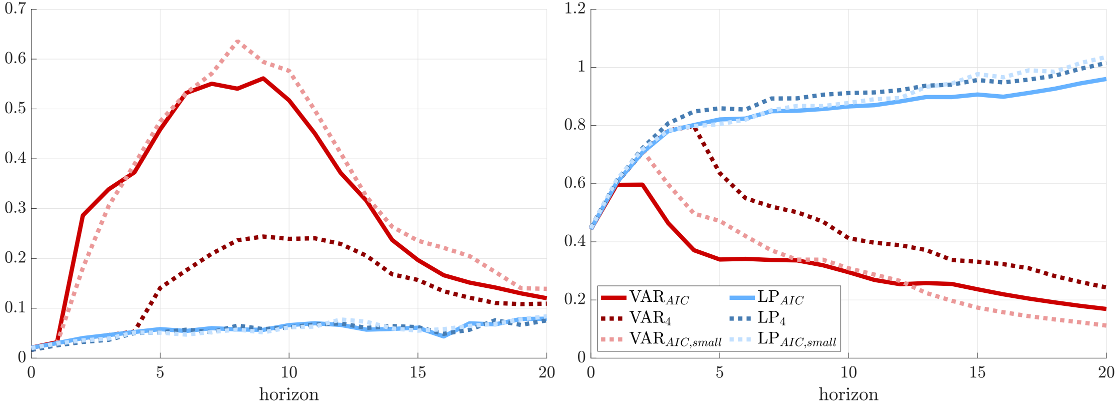
Bias



Standard deviation

average across 200 stationary and 200 non-stationary DGPs

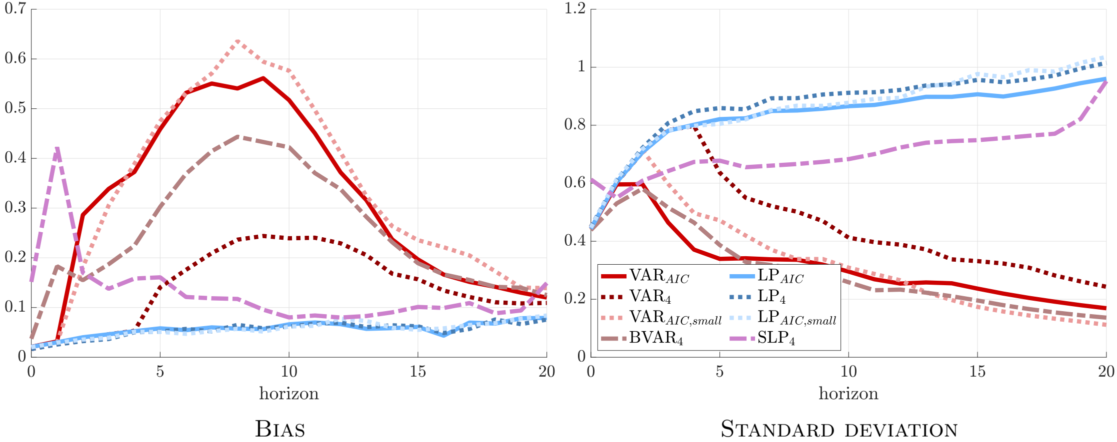
Simulation evidence: bias and standard deviation



Bias | Standard deviation

average across 200 stationary and 200 non-stationary DGPs

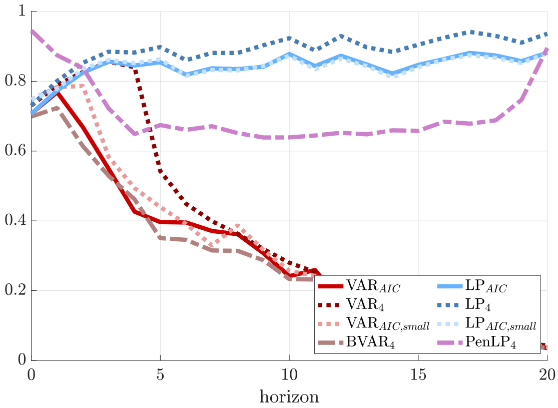
Simulation evidence: bias and standard deviation



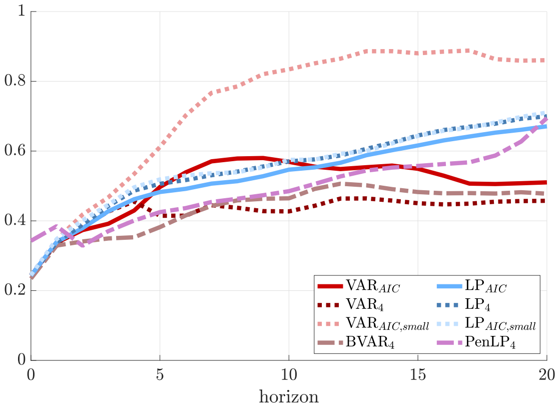
average across 200 stationary and 200 non-stationary DGPs

MSE loss: (B)VAR preferred over LP on average

Conventional way to trade off bias and variance: MSE $= \text{bias}^2 + \text{variance}$



MSE FOR STATIONARY DGPs



MSE FOR NON-STATIONARY DGPs

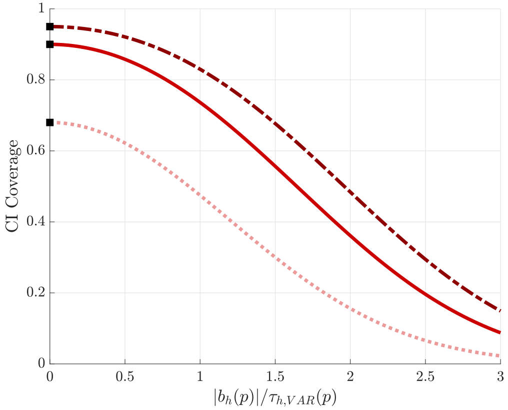
Bias-variance trade-off: recap

- Bias-variance trade-off is stark in practice.
- Robustness of LP to dynamic misspecification comes at significant variance cost.
- Under MSE loss, VAR is preferred over LP in the *average* simulation DGP.
  - Shrinkage (penalized LP or BVAR) often preferred over OLS.
- But MSE only evaluates the accuracy of the point estimate. This is not worth much without an accompanying uncertainty assessment.

Outline

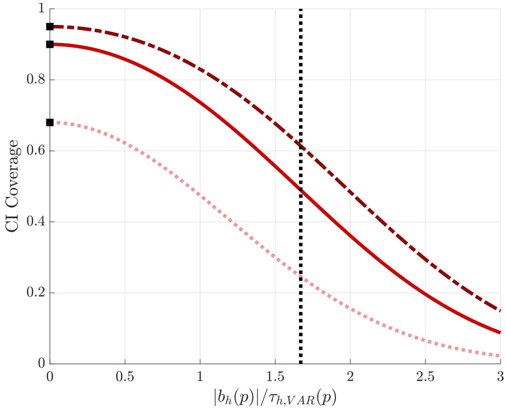
Uncertainty assessments: bias is costly

- Conventional to summarize uncertainty using confidence interval.
- Want coverage probability close to (say) 90% *regardless* of true DGP (not just for avg DGP!).
- Challenge for VARs: bias is really costly for coverage. CI has correct width, but off-center.
- Remember: easy to get worst-case bias $\approx 1.73 \times$ SE.



Uncertainty assessments: bias is costly

- Conventional to summarize uncertainty using confidence interval.
- Want coverage probability close to (say) 90% *regardless* of true DGP (not just for avg DGP!).
- Challenge for VARs: bias is really costly for coverage. CI has correct width, but off-center.
- Remember: easy to get worst-case bias $\approx 1.73 \times$ SE.



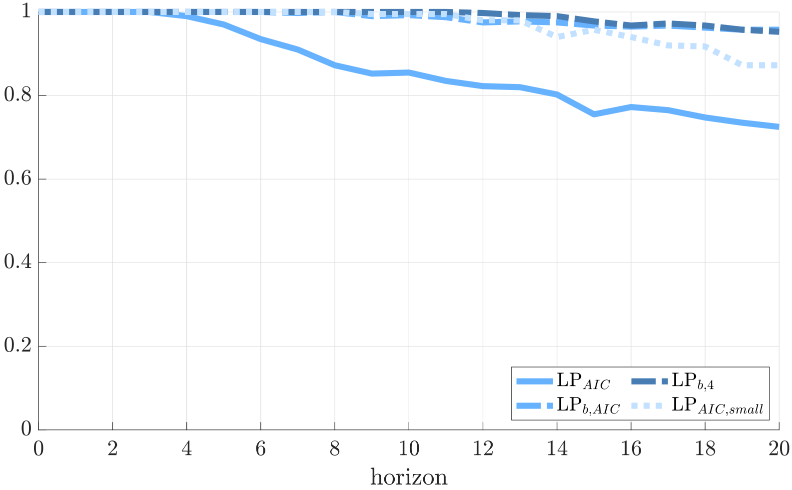
Long horizons and persistent data

- If the true DGP is a finite-order VAR, then the SVAR estimator is certainly efficient.
- But even in such settings, LP has some advantages: it's robust to persistence in the data and the length of the impulse response horizon h.
- Persistence:
  - The behavior of VAR estimators can depend sensitively on whether the data has unit roots (stochastic trends). Sometimes we get non-normally distributed estimators.
  - But the LP estimator projects on an (implicit) shock $\tilde{x}_t$, which is nicely stationary if we control for lags.
  - The LP estimator is therefore robustly (approximately) normally distributed, so the usual confidence interval works regardless of unit roots or not.

Long horizons and persistent data

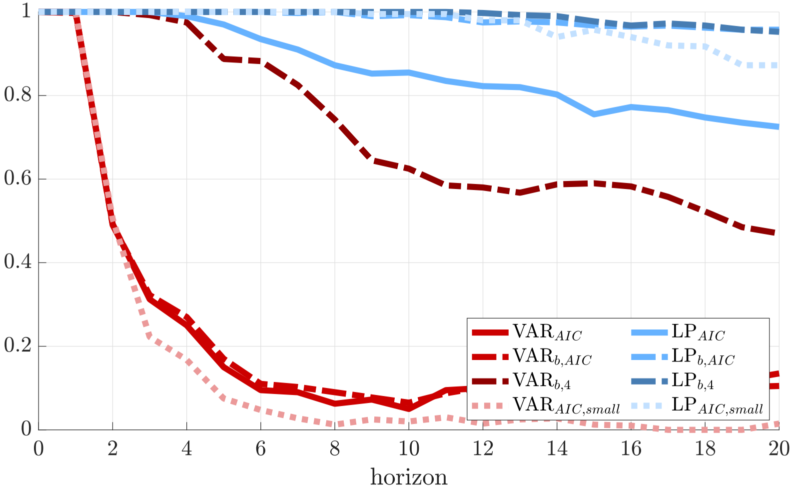
- Long horizons:
  - The VAR impulse response formula is nonlinear in the VAR coefficients.
  - E.g., AR(1): $\hat{\theta}_h^{\text{VAR}} = \hat{\rho}^h$.
  - At horizons h that are a substantial fraction of the sample size T, the nonlinearity is so severe that standard VAR confidence intervals break down.
  - But the LP estimator is just based on linear regression. No nonlinearity issues, even for large h.

Simulation evidence: confidence interval coverage



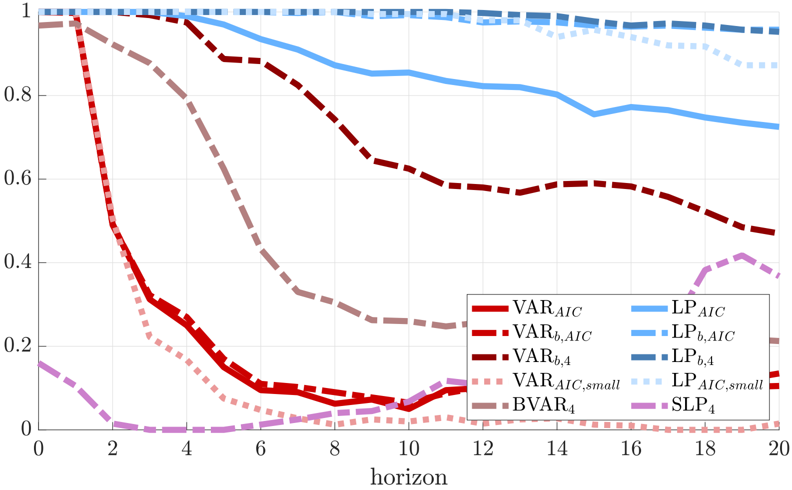
FRACTION OF DGPS WITH COVERAGE $\geq 80\%$ (TARGET COVERAGE 90%)

Simulation evidence: confidence interval coverage



Fraction of DGPs with coverage $\geq 80\%$ (target coverage 90%)

Simulation evidence: confidence interval coverage



FRACTION OF DGPS WITH COVERAGE $\geq 80\%$ (TARGET COVERAGE 90%)

Outline

VAR vs. LP: summary

- Choice of VAR vs. LP has nothing to do with identification.
  - Anything you can do with SVAR, you can do with LP, and *vice versa*.
- Bias-variance trade-off between LP (low bias) and VAR (low variance).
  - No free lunch for VARs.
  - MSE loss: VAR (or BVAR) with few lags preferred for "typical" DGPs.
- Uncertainty assessments that are reliable across a wide range of DGPs and horizons require either:
  1. LP (controlling for lags).
  2. VAR with many more lags than typically used (and minimal Bayesian shrinkage). Use LP as robustness check.

LP: importance of lagged controls

- Five good reasons to include a generous number of lagged control variables in LP:
  1. Identification of an economically meaningful shock $\tilde{x}_t$.
  2. Lowering SE (even if x_t is already unpredictable).
  3. Robustness against moderate dynamic misspecification (e.g., slightly predictable shock).
  4. Reliable CI coverage at long horizons h.
  5. Use heteroskedasticity-robust SE instead of HAC (e.g., Newey-West). The former typically works better in realistic sample sizes. Montiel Olea & P-M (2021)
- Should control for all var's and lags that are strong predictors of either outcome or impulse. OK to omit weak predictors.
- Can select lag length and/or controls var's using AIC applied to auxiliary VAR.

Outline

1. LP and VAR
2. Identification
3. Bias-variance trade-off
4. Uncertainty assessments
5. Taking stock
6. Nonlinear environments
7. Recommendations

VAR and LP in a nonlinear world

- The large-sample equivalence between VAR & LP with many lags is nonparametric: they share the same estimand regardless of how nonlinear the underlying DGP is.

- But is the common estimand of these linearity-based procedures economically meaningful if the DGP is nonlinear?

- Assume very generally that

$$y_{t+h} = \psi_h(\varepsilon_t, \nu_{h,t+h}), \quad \text{where } \varepsilon_t \text{ and } \nu_{h,t+h} \text{ are independent.}$$

  ε_t = shock of interest, $\nu_{h,t+h}$ = vector of nuisance shocks, ψ_h = structural function (arbitrarily nonlinear).

- We might be interested in the average structural function (ASF)

$$\Psi_h(\varepsilon) = E[\psi_h(\varepsilon, \nu_{h,t+h})], \quad \text{counterfactually fixing } \varepsilon_t.$$

Nonparametric identification with observed shocks

- Assume first we directly observe the shock $x_t = \varepsilon_t$.
- Then ASF can be estimated from a nonparam. regr'n of y_{t+h} on x_t. Gonçalves et al. (2024)
  - But this is challenging in typical macro data sets.
- What does a linear LP (= VAR with many lags) estimate in this case?

$$y_{t+h} = \hat{\theta}_h^{\text{LP}} x_t + \text{orthogonal controls} + \hat{\xi}_{h,t+h}.$$

Robustness of linear procedures

- In large samples, LP/VAR estimate a weighted average of nonlinear causal (marginal) effects: Yitzhaki (1996); Rambachan & Shephard (2021); Kolesár & P-M (2025)

$$\hat{\theta}_h^{\text{LP}} \overset{p}{\to} \int \omega(x)\Psi_h'(x)\,dx, \quad \text{where} \quad \omega(x) \equiv \frac{\text{Cov}(\mathbb{1}\{x_t \geq x\}, x_t)}{\text{Var}(x_t)}.$$

- The weight function is convex: $\omega(\cdot) \geq 0$, $\int \omega(x)\,dx = 1$. We get the sign right if Ψ_h is monotonic.

- Linear LP/VAR remain useful in a nonlinear world—at least as the first column of the regression table.
  - If we specifically want to characterize the nonlinearities, we should of course try to model them.
  - But if average marginal effects suffice, nonlinear modeling can be counterproductive. Even if variables have limited support, e.g., ZLB! Angrist (2001)

Estimating the weight function

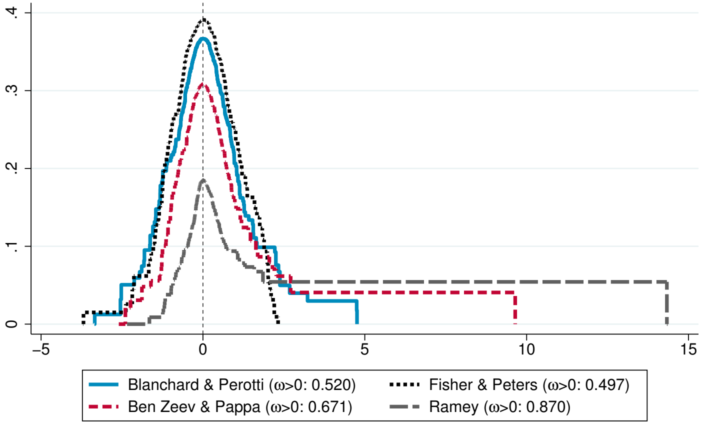
$$\hat{\theta}_h^{\text{LP}} \xrightarrow{p} \int \omega(x)\Psi_h'(x)\, dx, \quad \omega(x) \equiv \underbrace{\text{Cov}(\mathbb{1}\{x_t \geq x\}, x_t)/\text{Var}(x_t)}_{\text{regression coefficient}}$$

- $\hat{\omega}(x)$: slope in regression of $\mathbb{1}(x_t \geq x)$ on x_t.
- Easy to report, as it depends only on shock x_t, not on outcome variable or horizon.
- Weights transform empirical CDF of shocks into interpretable units.
  - Visualize asymmetry, outliers, etc.
- Special case: if $x_t \sim$ Gaussian, then $\omega(x) =$ density of x_t. Yitzhaki (1996)

Estimating the weight function

$$\hat{\theta}_h^{\mathrm{LP}} \xrightarrow{p} \int \omega(x)\Psi_h'(x)\,dx, \quad \omega(x) \equiv \underbrace{\mathrm{Cov}(\mathbb{1}\{x_t \geq x\}, x_t)/\,\mathrm{Var}(x_t)}_{\text{regression coefficient}}$$

Government spending shocks from Ramey (2016) handbook chapter:



Extensions: proxies, identification with controls

1. If we do not directly observe the shock ε_t but only a noisy proxy

$$z_t = f(\varepsilon_t, \text{independent noise}),$$

then the previous result goes through, as long as $E[z_t \mid \varepsilon_t = \varepsilon]$ is approx'ly monotonic.

   - f need not be linear. For example, z_t can be discrete (narrative proxy).

2. If we need to residualize x_t on control variables w_{t-1} to isolate a shock, then the earlier result applies only if we correctly model $E[x_t \mid w_{t-1}]$ (i.e., it's linear).

   - Might need to include nonlinear transformations of variables in w_{t-1}.
   - Verify that linearly residualized "shock" $x_t - \hat{\beta}' w_{t-1}$ cannot be predicted by *nonlinear* transformations of w_{t-1}.

Fragility of identification via heteroskedasticity

- We saw that linear LP/VAR with observed shocks/proxies estimate something meaningful in the nonparametric model
$$y_{t+h} = \psi_h(\varepsilon_t, \nu_{h,t+h}).$$
- If we don't have access to an observed shock/proxy, a popular identification scheme is identification via heteroskedasticity:
  - Assume we observe a discrete regime indicator D_t satisfying
  $$E[\varepsilon_t \mid D_t] = 0, \quad \text{but} \quad \text{Var}(\varepsilon_t \mid D_t) \neq \text{constant}.$$
  - If ψ_h is linear, impulse responses are identified. Sentana & Fiorentini (2001); Rigobon (2003)
- Unfortunately, if the true DGP is nonlinear, the linearity-based estimators in this literature do *not* estimate weighted averages of marginal effects.
  - Intuitively, marginal effects concern the effect of shifting the location of ε_t, but the observed D_t only affects its spread.

Fragility of identification via non-Gaussianity

- Another recently popular identification procedure is identification via non-Gaussianity:
  - Focusing on horizon $h = 0$, assume

$$y_t = \psi_0(\varepsilon_t, \nu_t),$$

    where y_t is now a vector of observed outcomes, and all elements of ε_t and ν_t are mutually independent and non-Gaussian.
  - If ψ_0 is linear, then impulse responses are identified. Gouriéroux, Monfort & Renne (2017); Lanne, Meitz & Saikkonen (2017)
- These procedures can break down as soon as ψ_0 is even moderately nonlinear.
  - Intuitively, independence and Gaussianity are vacuous in a nonparametric context. If we start with $\varepsilon_t \sim$ uniform, we can nonlinearly transform to any arbitrary multivariate distribution.

Outline

1. LP and VAR
2. Identification
3. Bias-variance trade-off
4. Uncertainty assessments
5. Taking stock
6. Nonlinear environments
7. Recommendations

Recommendations

- To analyze what—and how much—the data says about causal effects, use *either* (a) LPs *or* (b) VARs with very many lags and minimal shrinkage ($\approx$ LP).
  - VARs with conventional lag lengths (+ Bayesian shrinkage) remain useful for forecasting.
- Control for all var's and lags that are strong predictors of either outcome or impulse, guided by economic theory and AIC.
- Report heteroskedasticity-robust SE (no need for Newey-West).
- For highly persistent data:
  - Report bootstrap CI. Montiel Olea et al. (2025a)
  - Apply bias correction. Herbst & Johannsen (2024); Piger & Stockwell (2025)

Recommendations (continued)

- Though it's hard work constructing them, identification with observed shocks (or proxies thereof) buys robustness against nonlinearity in the DGP.
  - Report implicit weight function.
  - If controls needed for identif'n, verify that residualized shock is *nonlinearly* unpredictable.
  - Proxies should be credibly monotonic in the underlying shock, but need not be linear (e.g., discrete proxies are fine).
- Other identification schemes may not be nearly so robust.

Thank you!

Email: mikkelpm@uchicago.edu

Appendix

Proxy/IV identification

- If we observe a variable z_t that is a noisy proxy for the shock ε_t, a direct LP on z_t suffers from attenuation bias.

- Instead, it's common to use a 2SLS version of LP: Stock & Watson (2018)

$$y_{t+h} = \mu_h + \theta_h^{\text{LPIV}} m_t + \text{lags} + \xi_{h,t+h}, \quad \text{using } z_t \text{ as an IV for } m_t.$$

  This effectively scales the magnitude of the shock so that it causes a one-unit rise in the policy instrument m_t on impact.

- The above LP-IV estimand is equivalent with an "internal instrument" SVAR estimator:
  - Include the proxy z_t directly in the list of variables in the SVAR, and order it *first*.
  - Compute Cholesky-orthogonalized impulse responses with respect to the *first shock*.
  - Scale the magnitude of the shock using the same normalization as for LP-IV.

Proxy/IV identification

- The above estimation strategy differs from "external instrument" SVAR: Stock & Watson (2012); Mertens & Ravn (2013)
  - Estimate a VAR that *excludes* the proxy z_t.
  - Only use the proxy subsequently to identify the underlying structural shock.
- This latter procedure is less robust, as it requires the additional assumption that the shock is "invertible". Stock & Watson (2018); Plagborg-Møller & Wolf (2021)