Estimator Averaging of Local Projection and VAR Impulse Responses*

Chaoyi Chen† Elena Pesavento‡ Balázs Vonnák §

This version: May 6, 2026

Abstract

Local projections (LP) and vector autoregressions (VAR) are the two standard tools for impulse response analysis, but they often display a finite-sample trade-off: LP is typically less biased but more volatile, while VAR is more precise but can be biased under misspecification. We propose an easy-to-implement estimator-averaging approach that combines LP and VAR at each horizon by minimizing the mean squared error of the impulse response itself, rather than in-sample fit. We derive closed-form oracle weights for this finite-sample risk problem, develop feasible AR-sieve-bootstrap procedures, and compare them against an R^2-based model-averaging benchmark. For a benchmark class of short-memory linear data generating processes in which LP and VAR are both consistent, we establish the consistency and limiting distribution of the feasible averaged estimator. Monte Carlo results show meaningful risk reductions relative to LP and VAR alone. In an empirical application revisiting Bauer and Swanson (2023), estimator averaging delivers stable and economically intuitive responses for yields, activity, prices, and credit spreads.

JEL Classification: C32, C52, C53, E52

Keywords: Local projections; Vector autoregressions; Impulse response functions; Estimator averaging; Model averaging; Monetary policy shocks

*We thank the participants of the AMEF 2026 and the 3rd MNB-Fudan workshop 2025 for their valuable comments. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Magyar Nemzeti Bank. Refine.ink was used to check the paper for consistency and clarity.

†Magyar Nemzeti Bank (Central Bank of Hungary), Budapest, 1054, Hungary; Email: chenc@mnb.hu.

‡Department of Economics, Emory University, Atlanta, GA, 30322–2240, USA; Email: epesave@emory.edu

§Magyar Nemzeti Bank (Central Bank of Hungary), Budapest, 1054, Hungary; Email: vonnakb@mnb.hu.

1 Introduction

Local projections (LP) and vector autoregressions (VAR) are the two workhorse approaches to estimating impulse response functions (IRF). In practice, however, the two methods often deliver noticeably different estimates, especially at intermediate and long horizons. This creates a natural problem for empirical researchers: if LP and VAR provide conflicting answers, how should one combine the information in the two estimators?

A useful starting point is that the distinction is not primarily about the population object being estimated. Plagborg-Møller and Wolf (2021) show that, with sufficiently rich lag structure, LPs and VARs estimate the same population IRFs. The key difference is therefore one of finite-sample behavior. As emphasized in recent work, LP and VAR exhibit a familiar finite-sample bias–variance trade-off (see, e.g., Li et al., 2024; Montiel Olea et al., 2025). LPs tend to have lower bias but higher variance, especially at intermediate and long horizons, because they estimate separate horizon-specific regressions. VARs, by imposing parametric dynamic structure, typically achieve lower variance by borrowing strength across horizons, but they can incur higher bias when the DGP deviates from a finite-order VAR (Li et al., 2024).

This makes combining LP and VAR naturally appealing from a mean-squared-error perspective: an averaged estimator can potentially exploit LP's low-bias properties and VAR's low-variance properties, provided the weights adapt to horizon-specific performance. Montiel Olea et al. (2026) sharpen this trade-off in a local-misspecification framework. They show that LP remains first-order robust for inference, whereas VAR can suffer first-order bias that is relevant for coverage. This highlights a practical tension: LP is attractive when robustness to misspecification is the priority, while VAR is attractive when finite-sample precision is the priority.

Existing work on combining IRF estimators has followed several related routes. One strand takes a fit-oriented model-averaging approach. For example, Hounyo and Jung (2025) propose a two-stage scheme that averages within LPs and within VARs, and then blends the two classes.

Such procedures often yield smooth, interpretable IRFs and stable weights when many horizons are pooled. However, because the objective is in-sample *fit*, they do not directly target the estimation risk of the structural IRF. A related but distinct approach is developed by Nemtyrev and Boldea (2026), who propose Targeted Local Projections (TLP), a shrinkage estimator that pulls LP impulse responses toward their SVAR counterparts in order to reduce variance at the cost of some bias. Their framework is developed explicitly under a local-misspecification asymptotic setup and is complemented by bootstrap-based inference designed to improve coverage in that setting.

Motivated by these distinctions, we take an estimator-averaging approach that selects weights to minimize the expected error of the IRF itself. Specifically, for each horizon h, we choose w_h to minimize the population or estimated risk—variance or MSE—of the convex combination

$$\hat{\theta}_h(w_h) = w_h\,\hat{\theta}_{LP,h} + (1 - w_h)\,\hat{\theta}_{VAR,h}.$$

Our approach is symmetric between LP and VAR and directly tied to the object of interest. Rather than selecting the best-fitting model (Hounyo and Jung (2025)) or shrinking one estimator toward the other as a baseline (Nemtyrev and Boldea (2026)), we ask how to combine the two estimators so as to minimize IRF estimation risk. This prioritizes precision in the IRF itself, rather than in-sample fit, by choosing weights that directly reflect the LP–VAR bias–variance trade-off and by exploiting the covariance between LP and VAR to reduce sampling noise.

Relative to the existing literature, our contribution is threefold. First, in population, we derive closed-form finite-sample (infeasible) oracle weights that minimize the MSE of the combined estimator and make transparent how the optimal LP share varies with the horizon and with the underlying LP–VAR trade–off. Second, we develop the asymptotic theory for the AR-sieve plug-in averaged estimator under a benchmark short-memory linear DGP in which both LP and VAR are consistent for the same population impulse response. In that benchmark, the limiting

risk is variance-based, and the bootstrap bias terms in our Algorithm 1 are a finite-sample refinement that is asymptotically negligible but empirically useful. This places our framework in deliberate complement to Nemtyrev and Boldea (2026), who study a closely-related linear combination of LP and VAR estimators under a local-to-VAR drifting DGP and derive its asymptotic bias–variance trade-off in that regime. We use Monte Carlo simulations to assess its finite-sample performance, compare it with a simple R^2-based model-averaging benchmark, and show that our finite-sample MSE criterion and our AR-sieve-bootstrap implementation remain meaningful when the misspecification, if any, is fixed rather than drifting. Third, in an empirical application revisiting the high-frequency monetary policy shocks of Bauer and Swanson (2023), we show that estimator averaging systematically reconciles the often volatile IV-LP and very smooth IV-VAR responses: the estimated weights put more mass on LP at short horizons and on VAR at longer horizons, and the resulting IRFs for the two-year yield, industrial production, consumer prices, and the excess bond premium are reasonably smooth, lie between LP and VAR, and are arguably more economically intuitive than either estimator on its own.

The rest of the paper is organized as follows. Section 2 presents the estimator-averaging framework, derives the oracle weights, and outlines feasible implementations alongside the R^2-based model-averaging benchmark. Section 3 presents the required assumptions and derives the large-sample properties of the estimator. Section 4 reports Monte Carlo evidence on small-sample performance and the horizon-specific comparison between estimator and model averaging. Section 5 revisits Bauer and Swanson (2023) using our methods and documents the empirical pattern of weights and IRFs. Section 6 concludes. All mathematical proofs are collected in the appendix.

2 Estimators

In this section, we review the definitions of LP- and VAR-based IRF estimators and introduce our LP–VAR averaged estimator. We derive the optimal weights by minimizing the MSE of the combined estimator—an approach known as *estimator averaging* (see, e.g., Mittelhammer and Judge, 2005). We also compare our approach with *model averaging*, which selects weights by fitting an averaged model to maximize the R^2, following Hounyo and Jung (2025). The estimator-averaging framework presented here provides the core idea and can be extended to broader settings—for example, to averaging across multiple LP estimators or across multiple VAR estimators.

2.1 LP-Based IRF

The local projection method directly regresses the future outcome of a target variable, y_t, on a current impulse variable and a set of controls. Let Y_t denote the $n \times 1$ vector of observed macroeconomic variables, which includes y_t. The horizon-h LP regression is given by

$$y_{t+h} \;=\; c_h \;+\; \theta_h\, x_t \;+\; \gamma_h' z_t \;+\; u_{t+h}, \qquad t = 1, \ldots, T-h, \tag{2.1}$$

where c_h collects deterministic terms (e.g., a constant or trends), x_t is the impulse variable of interest, and z_t is a vector of control variables. Typically, z_t consists of lagged values of the system variables, $z_t = (Y_{t-1}', Y_{t-2}', \ldots, Y_{t-p}')'$. The LP estimator of the structural impulse response is the OLS (or 2SLS) estimate of the scalar coefficient on the impulse variable, denoted $\widehat{\theta}_{LP,h}$. A key advantage of this approach is that it requires no dynamic parametric assumptions beyond the linear projection itself.

Remark 1. *The definition of x_t depends on the identification strategy. If the structural shock is observed, x_t is the shock itself. If unobserved, x_t is typically an endogenous policy indicator. Because such an indicator may react contemporaneously to other shocks, merely including lagged*

controls in z_t is insufficient. To isolate the structural shock, researchers generally rely on external proxies (IV-LP) or, when economically credible, a recursive identification scheme that adds the contemporaneous values of slow-moving variables to z_t.

2.2 VAR-Based IRF

Alternatively, the vector autoregression (VAR) approach (see, e.g., Sims, 1980) extrapolates the impulse response by iterating on a reduced-form linear dynamic system. A standard VAR(p) model for the full vector of observables Y_t is given by

$$Y_t \;=\; c \;+\; \sum_{l=1}^{p} A_l Y_{t-l} \;+\; v_t, \tag{2.2}$$

where c contains the deterministic terms, A_l are the $n \times n$ matrices of autoregressive coefficients, and v_t is the vector of reduced-form innovations. This system is typically estimated equation-by-equation via OLS. To recover the structural impulse responses, an identification scheme (e.g., recursive ordering via Cholesky decomposition, or external instruments/proxy SVARs) is specified to map the reduced-form innovations to the underlying structural shocks, $v_t = M_0 \epsilon_t$.

From the fitted dynamics and the identified structural impact matrix M_0, the implied vector moving average (VMA) representation is computed. The VAR-based estimator, $\widehat{\theta}_{VAR,h}$, is then defined as the relevant scalar element of the h-step-ahead structural VMA matrix $\widehat{\theta}_{VAR,h} \;\equiv\; g_h\big(\widehat{A}_1, \ldots, \widehat{A}_p, \widehat{M}_0\big)$ where $g_h(\cdot)$ is the non-linear function mapping the reduced-form VAR parameters and the identification matrix to the horizon-h impulse response of y_t.

2.3 Combined Estimators

As is well documented, LPs and VARs exhibit a finite-sample bias–variance trade-off, motivating averaging to improve performance. There are two distinct routes: *estimator averaging*, which chooses weights to minimize the sampling risk (variance/MSE) of the IRF estimator, and *model*

averaging, which chooses weights to optimize either in-sample or predictive fit. To represent both, we implement (i) a risk-based estimator-averaging scheme that selects horizon-specific weights by minimizing the MSE of the combined IRF, and (ii) a fit-based model averaging benchmark (e.g., Hounyo and Jung, 2025). Our implementation of this risk-based estimator-averaging scheme encompasses two distinct methods: one that approximates the asymptotically optimal weight and a second that directly minimizes the MSE, both using a bootstrap procedure.

2.3.1 Estimator Averaging with MSE-Minimizing Weights

At each horizon h, define the averaged IRF

$$\widehat{\theta}_h(w_h) \;=\; w_h\,\widehat{\theta}_{LP,h} \;+\; (1-w_h)\,\widehat{\theta}_{VAR,h}, \qquad w_h \in [0,1], \tag{2.3}$$

where $\widehat{\theta}_{LP,h}$ is the LP-based IRF estimate, $\widehat{\theta}_{VAR,h}$ is the VAR-based IRF estimate, and w_h is the weight on LP at horizon h.

We decompose the (population) MSE of $\widehat{\theta}_h(w_h)$ in (2.3) into variance plus squared bias:

$$\begin{aligned}
\mathrm{MSE}_h(w_h) &= \mathbb{E}\Big[(\widehat{\theta}_h(w_h)-\theta_h)^2\Big] \\
&= \underbrace{w_h^2 V_{L,h} + (1-w_h)^2 V_{V,h} + 2w_h(1-w_h)C_h}_{\text{variance}} \;+\; \underbrace{\big(w_h\,b_{L,h} + (1-w_h)\,b_{V,h}\big)^2}_{\text{bias}^2} \\
&= w_h^2(V_{L,h}+b_{L,h}^2) + (1-w_h)^2(V_{V,h}+b_{V,h}^2) + 2w_h(1-w_h)\big(C_h + b_{L,h}b_{V,h}\big),
\end{aligned} \tag{2.4}$$

where

$$\begin{aligned}
&b_{L,h} = \mathbb{E}[\widehat{\theta}_{LP,h}] - \theta_h, && b_{V,h} = \mathbb{E}[\widehat{\theta}_{VAR,h}] - \theta_h, \\
&V_{L,h} = \mathrm{Var}(\widehat{\theta}_{LP,h}), && V_{V,h} = \mathrm{Var}(\widehat{\theta}_{VAR,h}), && C_h = \mathrm{Cov}(\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h}).
\end{aligned}$$

Let $a_h = V_{L,h}+b_{L,h}^2$, $d_h = V_{V,h}+b_{V,h}^2$, and $f_h = C_h + b_{L,h}b_{V,h}$. Differentiating (2.4) with respect to

w_h and solving the first-order condition, assuming an interior solution, yields the *oracle* weight

$$w_h^\star = \frac{d_h - f_h}{a_h + d_h - 2f_h}, \qquad \text{with } w_h^\star \in [0,1]. \tag{2.5}$$

Because $w_h^\star$ depends on unknown population quantities, we estimate it using a semiparametric AR-sieve-bootstrap, following the time-series bootstrap literature (see, e.g., Bühlmann, 1997; Kreiss and Paparoditis, 2003; Gonçalves and Kilian, 2004; Gonçalves, 2007). Specifically, we fit an AR(p) model with p selected by BIC, resample the estimated residuals nonparametrically, simulate pseudo time series, and re-estimate the relevant objects to obtain $(\widehat{a}_h, \widehat{d}_h, \widehat{f}_h)$. This approach avoids committing to a fixed finite-order structural DGP and instead allows the bootstrap to approximate a broader class of short-memory linear processes through a growing-order autoregressive sieve, while still delivering a feasible estimator of $w_h^\star$. We refer to the procedure as the AR-sieve-bootstrap plug-in estimator, or simply the *plug-in estimator*. Algorithm 1 summarizes the procedure.[1]

The bootstrap bias terms in Algorithm 1 are used for finite-sample implementation of the MSE criterion. They are not required for the benchmark first-order asymptotic theory in Section 3. In the benchmark setting, both LP and VAR are asymptotically centered at θ_h, so the limiting risk is variance-covariance based and the bias contributions are first-order negligible. We therefore treat the bootstrap bias terms as a *finite-sample refinement* of a variance-based limit rather than as objects that the first-order theory is required to validate: they capture finite-sample features that disappear in the first-order limit but matter empirically. A higher-order justification of these terms would require additional smoothness and moment conditions on the AR-sieve approximation, which we do not impose.[2]

The weight choice is a finite-sample risk problem. For fixed T, the local projection and

[1] For ease of exposition, Algorithm 1 is written for a univariate model. It can easily be extended to a multivariate $VAR(p_T)$ sieve.

[2] A formal asymptotic justification for the bias terms used in Algorithm 1 lies outside the variance-based limit considered here; the contemporaneous and complementary work of Nemtyrev and Boldea (2026) develops such a justification in a local-to-VAR drifting framework.

VAR estimators may exhibit nonzero bias, so the MSE depends on both variance and bias components. In Section 3, we derive our asymptotic results as $T \to \infty$. Working under the assumption of a short-memory linear process where the LP and VAR estimators are both consistent, these bias terms become asymptotically negligible, and the first-order limit distribution is centered at θ_h. Within this same section, we subsequently establish the consistency of the bootstrap estimator, $\widehat{w}$, and the limiting theory for $\widehat{\theta}(\widehat{w})$.

Algorithm 1 AR-sieve-bootstrap plug-in weight $\widehat{w}_h$ (implemented jointly over h)

1: Select $\hat{p}$ (e.g., by BIC over $0 \le p \le p_{\max,T}$, where $p_{\max,T} \to \infty$ sufficiently slowly), and choose the number of bootstrap draws B.[3]
2: Fit AR($\hat{p}$): $y_t = \sum_{j=1}^{\hat{p}} \widehat{\phi}_j y_{t-j} + \widehat{\eta}_t$; set $\tilde{\eta}_t = \widehat{\eta}_t - \bar{\eta}$.
3: Construct the sieve pseudo-truth $\tilde{\theta}_h^{sieve}$ implied by $\widehat{\phi}$ (e.g., model-implied IRF or long-run simulation).
4: **for** $b = 1, \dots, B$ **do**
5: Draw $\tilde{\eta}_t^{*(b)}$ iid from the empirical distribution of $\{\tilde{\eta}_t\}_{t=1}^T$.
6: Simulate $y_t^{*(b)} = \sum_{j=1}^{\hat{p}} \widehat{\phi}_j y_{t-j}^{*(b)} + \tilde{\eta}_t^{*(b)}$ (discard burn-in).
7: Re-estimate on $\{y_t^{*(b)}\}_{t=1}^T$ to obtain $\widehat{\theta}_{LP,h}^{*(b)}$ and $\widehat{\theta}_{VAR,h}^{*(b)}$ for all h.
8: **end for**
9: **for** each h **do**
10: $\bar{\theta}_{LP,h}^* = B^{-1} \sum_{b=1}^B \widehat{\theta}_{LP,h}^{*(b)}, \quad \bar{\theta}_{VAR,h}^* = B^{-1} \sum_{b=1}^B \widehat{\theta}_{VAR,h}^{*(b)}$.
11: $\widehat{V}_{LP,h}^{SV} = B^{-1} \sum_{b=1}^B (\widehat{\theta}_{LP,h}^{*(b)} - \bar{\theta}_{LP,h}^*)^2, \quad \widehat{V}_{VAR,h}^{SV} = B^{-1} \sum_{b=1}^B (\widehat{\theta}_{VAR,h}^{*(b)} - \bar{\theta}_{VAR,h}^*)^2$.
12: $\widehat{C}_h^{SV} = B^{-1} \sum_{b=1}^B (\widehat{\theta}_{LP,h}^{*(b)} - \bar{\theta}_{LP,h}^*)(\widehat{\theta}_{VAR,h}^{*(b)} - \bar{\theta}_{VAR,h}^*)$.
13: $\widehat{b}_{LP,h}^{SV} = \bar{\theta}_{LP,h}^* - \tilde{\theta}_h^{sieve}, \quad \widehat{b}_{VAR,h}^{SV} = \bar{\theta}_{VAR,h}^* - \tilde{\theta}_h^{sieve}$.
14: $\widehat{a}_h = \widehat{V}_{LP,h}^{SV} + (\widehat{b}_{LP,h}^{SV})^2, \quad \widehat{d}_h = \widehat{V}_{VAR,h}^{SV} + (\widehat{b}_{VAR,h}^{SV})^2, \quad \widehat{f}_h = \widehat{C}_h^{SV} + \widehat{b}_{LP,h}^{SV} \widehat{b}_{VAR,h}^{SV}$.
15: $\widehat{w}_h = (\widehat{d}_h - \widehat{f}_h)/(\widehat{a}_h + \widehat{d}_h - 2\widehat{f}_h)$; clip to $[0, 1]$ (with safeguards if the denominator is small).
16: **end for**

Our MSE decomposition in (2.4) is closely related to the risk criterion in Nemtyrev and Boldea (2026). Both papers minimize the MSE of a linear combination of LP and VAR/SVAR impulse-response estimators, and both deliver weights that are in the form of a ratio in which the numerator captures a variance/covariance gap and the denominator adds a bias-squared term. Beyond this shared algebraic structure, however, the two papers are different both theoretically and methodologically, and we view the contributions as complementary rather than competing.

[3] The formal theory treats the sieve order as satisfying the standard AR-sieve growth conditions. We do not prove that the BIC-selected lag automatically satisfies these conditions.

Nemtyrev and Boldea (2026) adopt the local-to-VAR DGP of Montiel Olea et al. (2026) in which the misspecification of the VAR shrinks at rate $T^{-\zeta}$ for some $\zeta > 1/4$. Under this drifting sequence, the VAR carries an $O(T^{-\zeta})$ asymptotic bias that is first-order relevant, and the weight is determined by a bias–variance trade-off in the asymptotic limit. We instead work under a fixed short-memory Wold DGP in which both LP and VAR are root-T consistent for the same population impulse response. The bias–variance trade-off in our framework is therefore a *finite-sample* phenomenon, not an asymptotic one, and the limiting oracle weight is a pure variance/covariance ratio. Given their set up, Nemtyrev and Boldea (2026) interpret the estimators combination as shrinkage of LP toward VAR, with the VAR playing the role of a regularization target; their estimator reduces to LP when shrinkage is off and to VAR in the opposite limit. We adopt a symmetric estimator-averaging perspective, treating LP and VAR as two estimators of the same scalar IRF and choosing weights that minimize the MSE of the combined estimator that can be easily extended to K estimators (see Remark 2).

Remark 2. *The estimator-averaging framework extends directly to more than two IRF estimators. Suppose* $\widehat{\boldsymbol{\theta}}_h = (\hat{\theta}_h^{(1)}, \ldots, \hat{\theta}_h^{(K)})^\top$ *collects* K *estimators of the same scalar IRF* θ_h*, and define*

$$\boldsymbol{e}_h = \widehat{\boldsymbol{\theta}}_h - \theta_h \mathbf{1}_K, \qquad \mathbf{b}_h = \mathbb{E}[\boldsymbol{e}_h], \qquad \boldsymbol{\Sigma}_h = \mathrm{Var}(\boldsymbol{e}_h).$$

For weights $\mathbf{w}_h \in \mathbb{R}^K$ *satisfying* $\mathbf{1}_K^\top \mathbf{w}_h = 1$*, consider the averaged estimator*

$$\tilde{\theta}_h(\mathbf{w}_h) = \mathbf{w}_h^\top \widehat{\boldsymbol{\theta}}_h.$$

Its MSE is

$$\mathrm{MSE}[\tilde{\theta}_h] = \mathbf{w}_h^\top \big(\boldsymbol{\Sigma}_h + \mathbf{b}_h \mathbf{b}_h^\top\big) \mathbf{w}_h.$$

Hence, minimizing MSE subject to $\mathbf{1}_K^\top \mathbf{w}_h = 1$ *yields the oracle weight vector*

$$\mathbf{w}_h^\star = \frac{\mathbf{G}_h^{-1} \mathbf{1}_K}{\mathbf{1}_K^\top \mathbf{G}_h^{-1} \mathbf{1}_K}, \quad \mathbf{G}_h = \boldsymbol{\Sigma}_h + \mathbf{b}_h \mathbf{b}_h^\top. \tag{2.6}$$

Thus, the same logic applies to averaging across multiple LP- and VAR-based IRF estimators. As before, $\mathbf{w}_h^\star$ *depends on unknown population quantities and must be estimated in practice, for example by bootstrap methods.*

Remark 3. *The estimator-averaging framework is not tied to OLS estimation. The objects* $\widehat{\theta}_{LP,h}$ *and* $\widehat{\theta}_{VAR,h}$ *can be any two estimators of the same structural impulse response. Thus, under the usual relevance and exogeneity conditions for a valid external instrument, the framework can also combine IV-based estimators. In particular, one may combine a horizon-specific IV-LP estimator with a proxy-SVAR or IV-VAR estimator identified using the same external instrument:* $\widehat{\theta}_h(w_h) = w_h\widehat{\theta}^{IV}_{LP,h} + (1-w_h)\widehat{\theta}^{IV}_{VAR,h}$. *The MSE formula in (2.4) then applies with the bias, variance, and covariance terms interpreted for the corresponding IV-based estimators. The feasible bootstrap implementation proceeds analogously: in each bootstrap sample, both the IV-LP and proxy-SVAR/IV-VAR estimators are re-estimated under the same identification scheme, and the resulting pair of IRF estimates is used to compute the bootstrap risk components.*[4]

Remark 4. *Following Hounyo and Jung (2025), one can adopt a two–stage estimator-averaging scheme: (i) first, average within the LP–based estimators and within the VAR–based estimators separately; (ii) second, average the two class-specific aggregates (LP–avg and VAR–avg) using the same MSE–minimizing weighting rule. The second-stage weight leverages the complementary strengths of LP (lower bias at short horizons) and VAR (lower variance at longer horizons) to reduce the combined estimator's risk.*

As a complementary implementation, we also consider empirically optimal weights that are chosen by directly minimizing the bootstrap MSE of the combined estimator, rather than by first estimating the bias, variance, and covariance components entering (2.5). Concretely, for each bootstrap resample we re-estimate the LP and VAR impulse responses and numerically

[4]The formal first-order theory in Section 3 is stated for a generic pair of root-T consistent LP and VAR impulse-response estimators. Extending the primitive conditions to weak instruments, many instruments, or invalid external instruments is beyond the scope of the paper. In the empirical application, we treat the Bauer–Swanson high-frequency surprise as the external instrument and implement the same averaging procedure using the resulting IV-LP and proxy-SVAR estimates.

search over weights in $[0, 1]$ for the value that minimizes the bootstrap MSE of the combined estimator.

We also study a more flexible specification in which the weight depends on the discrepancy between the two estimators:

$$w_h^{EO} = \frac{a}{1 + b\left(\frac{\widehat{\theta}_{LP,h} - \widehat{\theta}_{VAR,h}}{\widehat{\theta}_{LP,h} + \widehat{\theta}_{VAR,h}}\right)^2}, \tag{2.7}$$

where a and b are non-negative parameters chosen numerically. The normalization by the sum ensures scale invariance, and the case $b = 0$ nests the constant-weight specification. Intuitively, larger discrepancies between LP and VAR lead this scheme to put less weight on LP, reflecting its typically higher variance. We refer to this extension as *flexible estimator averaging*. In both cases, implementation relies on the same semiparametric AR-sieve-bootstrap used for the plug-in method. Since these procedures are primarily computational alternatives and do not form the basis of our asymptotic theory, we treat them as complementary rather than central to the paper's main contribution.

2.3.2 Model Averaging with R^2-Based Weights

In this subsection, we briefly review *model averaging*, focusing on achieving the best in–sample *fit*. In the spirit of Hounyo and Jung (2025), we choose the weight between LP- and VAR-based IRF estimates using their respective in-sample coefficients of determination, R^2.

First, we calculate the R^2 for the LP regression at each horizon h, denoted as $R^2_{LP,h}$. Because the LP method estimates a separate regression for each horizon, this measure of fit is horizon-specific. Second, we calculate the R^2 for the VAR model. In the univariate case, this is the in-sample R^2 from the fitted VAR equation. In the multivariate case, we use the in-sample R^2 from the reduced-form VAR equation corresponding to the response variable whose impulse response is being averaged. Since the VAR is estimated once on the fixed sample, this measure is horizon-invariant and is denoted by R^2_{VAR}.

Using these measures of in-sample fit, we assign the weight to the LP estimator for a given horizon h proportionally to its relative explanatory power. The model averaging weight is constructed as:

$$\widehat{w}_h^M = \frac{R^2_{LP,h}}{R^2_{LP,h} + R^2_{VAR}}, \quad \widehat{w}_h^M \in [0,1]. \tag{2.8}$$

We then obtain the model-averaged IRF at horizon h as

$$\widehat{\theta}_h^M = \widehat{w}_h^M \, \widehat{\theta}_{LP,h} + \big(1 - \widehat{w}_h^M\big)\widehat{\theta}_{VAR,h}.$$

3 Asymptotic Results

Since the averaged estimator relies on a bootstrap estimate of $w_h^\star$, this section establishes the asymptotic properties of both $\hat{w}_h$ and $\hat{\theta}_h(\hat{w}_h)$. We develop the asymptotic theory for $\widehat{w}_h$ and $\widehat{\theta}_h(\widehat{w}_h)$ by focusing on the baseline environment where both the LP and VAR estimators are consistent. By analyzing the AR-sieve-bootstrap min-MSE procedure under a short-memory linear DGP in this benchmark setting, we provide the rigorous theoretical justification that anchors our methodology. We impose standard regularity conditions below.

Although the estimators in Section 2 are written as finite-lag LP and VAR regressions, the benchmark asymptotic theory does not assume that the true DGP is a fixed finite-order VAR. Instead, the true process is modeled as a short-memory Wold process. Equivalently, under standard invertibility and summability conditions, this process admits an infinite-order autoregressive representation that can be approximated by a growing-order AR/VAR sieve. The finite-lag LP and VAR specifications used in estimation should therefore be viewed as feasible approximations to this short-memory benchmark, with the sieve order increasing sufficiently slowly relative to the sample size.

Assumption 1. *The vector process $\{Y_t\}$ is covariance-stationary and purely nondeterministic,*

and admits the Wold representation

$$Y_t = \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j}, \qquad E[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0, \qquad E[\varepsilon_t \varepsilon_t'] = \Sigma_\varepsilon \succ 0.$$

Assume short memory, $\sum_{j=0}^{\infty} j\|\Psi_j\| < \infty$*, and* $E\|\varepsilon_t\|^{4+\delta} < \infty$ *for some* $\delta > 0$*.*

Assumption 2. *Fix* h*. There exists a (possibly singular)* 2×2 *matrix* $\Omega_h^{(2)}$ *such that*

$$\sqrt{T}\Big((\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})' - (\theta_h, \theta_h)'\Big) \;\Rightarrow\; \mathcal{N}(0, \Omega_h^{(2)}). \tag{3.1}$$

In addition, for some $\delta > 0$*,* $\sup_T \mathbb{E}\|Z_{T,h}\|^{2+\delta} < \infty$*, where* $Z_{T,h}$ *is defined in Assumption 3.*

Assumption 3. *For any fixed* h*, define*

$$Z_{T,h} = \sqrt{T}\Big((\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})' - (\theta_h, \theta_h)'\Big), \tag{3.2}$$

$$Z^*_{T,h} = \sqrt{T}\Big((\widehat{\theta}^*_{LP,h}, \widehat{\theta}^*_{VAR,h})' - (\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})'\Big), \tag{3.3}$$

where $(\widehat{\theta}^*_{LP,h}, \widehat{\theta}^*_{VAR,h})'$ *is generated by the AR-sieve-bootstrap in algorithm 1. Let* $\mathbb{E}^*[\cdot]$ *denote expectation conditional on the data, and let* BL_1 *be the class of real-valued functions* φ *on* $\mathbb{R}^2$ *that satisfy* $\sup_z |\varphi(z)| \le 1$ *and have Lipschitz constant at most* 1*.*

(i) There exists a tight $\mathbb{R}^2$*-valued random vector* Z_h *such that* $Z_{T,h} \Rightarrow Z_h$ *and*

$$\sup_{\varphi \in BL_1} \Big|\mathbb{E}^*[\varphi(Z^*_{T,h})] - \mathbb{E}[\varphi(Z_{T,h})]\Big| \xrightarrow{p} 0. \tag{3.4}$$

(ii) For some $\delta > 0$*,* $\mathbb{E}^*[\|Z^*_{T,h}\|^{2+\delta}] = O_p(1)$ *and* $\sup_T \mathbb{E}[\|Z_{T,h}\|^{2+\delta}] < \infty$*.*

(iii) $B = B(T) \to \infty$ *as* $T \to \infty$*.*

To connect the finite-sample MSE weight in (2.6) to the root-T asymptotic theory, it is useful to rescale the risk components. Define $a_{T,h} = V_{L,T,h} + b^2_{L,T,h}$, $\quad d_{T,h} = V_{V,T,h} + b^2_{V,T,h}$, $\quad f_{T,h} =$

$C_{T,h} + b_{L,T,h}b_{V,T,h}$, and let $A_{T,h} = Ta_{T,h}$, $D_{T,h} = Td_{T,h}$, $F_{T,h} = Tf_{T,h}$. Since multiplying both the numerator and denominator of (2.6) by T does not change the weight, we can write

$$w^{\star}_{T,h} = \frac{d_{T,h} - f_{T,h}}{a_{T,h} + d_{T,h} - 2f_{T,h}} = \frac{D_{T,h} - F_{T,h}}{A_{T,h} + D_{T,h} - 2F_{T,h}}.$$

In the benchmark case in which both the LP and VAR estimators are root-T consistent and asymptotically centered at θ_h, these scaled risk components converge to the corresponding entries of $\Omega^{(2)}_h$.

Assumption 4. *For any fixed h, the oracle min-MSE weight $w^{\star}_h$ in (2.5) is unique after clipping to $[0,1]$, and the asymptotic variance for the averaged estimator evaluated at the oracle weight is finite and well-defined.*

Assumption 5. *For any fixed h, the scaled oracle risk components satisfy $(A_{T,h}, D_{T,h}, F_{T,h}) \to (A_h, D_h, F_h)$, where $A_h + D_h - 2F_h \geq c > 0$ for some constant $c > 0$. Moreover, the limiting oracle weight $w^{\star}_h = \frac{D_h - F_h}{A_h + D_h - 2F_h}$ lies in the interior of $[0,1]$: $w^{\star}_h \in [\underline{w}, 1-\underline{w}]$ for some $\underline{w} \in (0, 1/2)$.*

Assumption 6. *For any fixed h, in addition to Assumption 3(ii), assume that conditional on the data $\mathbb{E}^*[\|Z^*_{T,h}\|^{4+\delta}] = O_p(1)$ for the same $\delta > 0$ as in Assumption 3.*

Assumption 1 assumes a short-memory Wold DGP. Under Assumptions 1–2 and routine regularity conditions—including nonsingularity of the relevant projection matrices and, for sieve-based VAR/LP approximations, a lag order p_T that grows sufficiently slowly with T—the LP and VAR IRF estimators are consistent and jointly asymptotically normal for each fixed horizon h. Moreover, AR-sieve-bootstrap approximations are valid for broad Wold-type processes for statistics whose limits depend on second-order structure. Finally, LP and VAR target the same impulse responses asymptotically when the lag length increases (Plagborg-Møller and Wolf, 2021).

Assumption 2 postulates joint root-T asymptotic normality for the LP and VAR IRF estimators at a fixed horizon h. This condition is standard for OLS-based LP and VAR estimators under Assumption 1 and routine rank and weak-dependence assumptions; for VAR IRFs

it follows by applying the delta method to the smooth map from VAR OLS coefficients to the horizon-h IRF. Writing $\Omega_h^{(2)} = \begin{pmatrix} \Omega_{11,h} & \Omega_{12,h} \\ \Omega_{12,h} & \Omega_{22,h} \end{pmatrix}$ as in (3.1), the (infeasible) oracle weight $w_h^\star = \arg\min_w e(w)'\Omega_h^{(2)} e(w)$, with $e(w) = (w, 1-w)'$, is

$$w_h^\star = \frac{\Omega_{22,h} - \Omega_{12,h}}{\Omega_{11,h} + \Omega_{22,h} - 2\Omega_{12,h}}. \tag{3.5}$$

Assumption 3 is a high-level statement that the AR-sieve-bootstrap reproduces the *first-order* limit law of the centered/scaled LP–VAR pair at a fixed horizon h (expressed via the bounded-Lipschitz metric). The moment condition in (ii) is included to justify convergence of quadratic functionals used by the MSE weight. The AR-sieve-bootstrap is designed for short-memory linear processes and may provide a poor approximation when the DGP features strong nonlinearity, structural breaks, unit roots or near-unit roots, long memory, heavy tails, or strong conditional heteroskedasticity. In the latter case, a wild or heteroskedasticity-robust bootstrap variant may be more appropriate. Assumption 4 ensures the oracle problem is well posed (a unique clipped minimizer exists) and that the asymptotic variance of the averaged estimator is finite, so plug-in variance estimation is meaningful. When LP and VAR are both root-T consistent for the same θ_h, the joint asymptotic covariance $\Omega_h^{(2)}$ of $(\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})$ may be singular. This does not affect consistency of $\widehat{w}_h$ or of the plug-in variance for $\widehat{\theta}_h(\widehat{w}_h)$; it only implies we should not require $\Omega_h^{(2)}$ to be positive definite.

Assumption 5 is a nondegeneracy condition for the scaled asymptotic risk problem. The unscaled finite-sample denominator in (2.6) is typically of order T^{-1} under root-T asymptotics, so it is the scaled denominator $A_{T,h} + D_{T,h} - 2F_{T,h}$, rather than $a_{T,h} + d_{T,h} - 2f_{T,h}$, that must be bounded away from zero. The lower bound on $A_{T,h} + D_{T,h} - 2F_{T,h}$ makes the mapping from the scaled risk components to the oracle weight locally smooth. The additional interiority condition $w_h^\star \in [\underline{w}, 1-\underline{w}]$ ensures that clipping is asymptotically inactive. While it is possible for the denominator $A_h + D_h - 2F_h$ to approach zero, for example in the lag-augmented limit

of Plagborg-Møller and Wolf (2021) when the LP and VAR estimators become asymptotically equivalent, Algorithm 1 already includes a numerical safeguard that returns a default weight when the estimated denominator falls below a small threshold, so the procedure is operationally well defined for any sample. In addition, when LP and VAR are close to asymptotically equivalent, the weight is poorly identified but the averaged estimator itself is well-behaved, because any convex combination of two nearly-coincident estimators delivers nearly the same value. The Monte Carlo evidence in Section 4 confirms this pattern.

For the benchmark first-order theory, $\widehat{w}_h$ denotes the plug-in weight computed from the scaled variance-covariance components $(\widehat{A}_{T,h}, \widehat{D}_{T,h}, \widehat{F}_{T,h})$; the finite-sample implementation in Algorithm 1 additionally includes bootstrap bias terms.

Theorem 1 (Consistency and rate of AR-sieve-bootstrap weights)**.** *For any fixed h, under Assumptions 3 and 5, $\widehat{w}_h \xrightarrow{p} w_h^\star$. If, in addition,*

$$\left\|(\widehat{A}_{T,h}, \widehat{D}_{T,h}, \widehat{F}_{T,h}) - (A_h, D_h, F_h)\right\| = O_p(r_T) + O_p(B^{-1/2}), \tag{3.6}$$

for some deterministic sequence $r_T \to 0$, then $|\widehat{w}_h - w_h^\star| = O_p(r_T) + O_p(B^{-1/2})$.

Let Ω be a 2×2 covariance matrix. Define

$$V_h(w, \Omega) = w^2\Omega_{11} + (1-w)^2\Omega_{22} + 2w(1-w)\Omega_{12}. \tag{3.7}$$

Theorem 2 (Consistency and asymptotic normality of $\widehat{\theta}_h(\widehat{w}_h)$)**.** *For any fixed h, under Assumptions 2, 3, 4, and 5, $\widehat{\theta}_h(\widehat{w}_h) \xrightarrow{p} \theta_h$, and*

$$\sqrt{T}\big(\widehat{\theta}_h(\widehat{w}_h) - \theta_h\big) \Rightarrow \mathcal{N}\Big(0,\, V_h(w_h^\star, \Omega_h^{(2)})\Big), \tag{3.8}$$

where $V_h(w_h^\star, \Omega_h^{(2)})$ is defined by equation (3.7) with $w = w_h^\star$ and $\Omega = \Omega_h^{(2)}$.

Theorem 1 shows that the AR-sieve-bootstrap plug-in weight $\widehat{w}_h$ is consistent for the limiting

oracle weight $w_h^\star$ at each fixed horizon h, provided that the scaled bootstrap risk components consistently estimate their limiting counterparts. The scaling is important because, under the benchmark root-T asymptotics, the unscaled variance and covariance terms entering the finite-sample MSE are of order T^{-1}. Thus, the relevant nondegeneracy condition is imposed on the scaled denominator $A_h + D_h - 2F_h$, rather than on the raw denominator $a_{T,h} + d_{T,h} - 2f_{T,h}$. If a rate statement is desired, it depends on the convergence rate of the scaled risk-component estimator, denoted by r_T, together with the Monte Carlo simulation error $B^{-1/2}$. We leave r_T as a high-level rate because its exact value depends on the AR-sieve approximation, the lag-order sequence, and the statistic being bootstrapped. Importantly, Theorem 2 only requires $\widehat{w}_h \xrightarrow{p} w_h^\star$, since the plug-in weight error is then first-order negligible. Theorem 2 shows that $\widehat{\theta}_h(\widehat{w}_h)$ is consistent for θ_h and asymptotically normal at the root-T rate, with asymptotic variance $V_h(w_h^\star, \Omega_h^{(2)})$. Although Theorem 1 applies to correctly specified models, our simulations show that our estimator performs well also under misspecification.

Theorem 3 (Consistency and rate of plug-in asymptotic variance)**.** *For any fixed h, let $\widehat{V}_h = V_h(\widehat{w}_h, \widehat{\Omega}_h^{(2)})$. Suppose that $\widehat{\Omega}_h^{(2)} \xrightarrow{p} \Omega_h^{(2)}$. Under the conditions of Theorem 1, $\widehat{V}_h \xrightarrow{p} V_h(w_h^\star, \Omega_h^{(2)})$. If, in addition, $|\widehat{w}_h - w_h^\star| = O_p(r_T) + O_p(B^{-1/2})$ and $\|\widehat{\Omega}_h^{(2)} - \Omega_h^{(2)}\| = O_p(s_T) + O_p(B^{-1/2})$ for some $s_T \to 0$, then $|\widehat{V}_h - V_h(w_h^\star, \Omega_h^{(2)})| = O_p(r_T + s_T) + O_p(B^{-1/2})$.*

Theorem 3 justifies plug-in inference based on $\widehat{V}_h = V_h(\widehat{w}_h, \widehat{\Omega}_h^{(2)})$. In practice, one may estimate the 2×2 matrix $\Omega_h^{(2)}$ by forming the stacked vector of estimating equations for $(\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})$ and applying a standard HAC estimator (e.g., Newey–West) to the resulting score/influence-function series. Equivalently, one can estimate $\Omega_h^{(2)}$ via the same AR-sieve bootstrap used to construct $\widehat{w}_h$: compute $Z_{T,h}^{*(b)}$ as in (3.3) and set $\widehat{\Omega}_h^{(2)} = B^{-1} \sum_{b=1}^{B} Z_{T,h}^{*(b)} Z_{T,h}^{*(b)\prime}$ (with recentering if desired).

As an alternative to plug-in Wald inference, one can use a AR-sieve-bootstrap to obtain confidence bands that are robust to conditional heteroskedasticity. Fit a stable VAR(p) sieve to $\{y_t\}$, form bootstrap innovations $u_t^{*(b)} = \eta_t^{(b)} \widehat{u}_t$ with i.i.d. Rademacher multipliers $\eta_t^{(b)} \in \{-1, +1\}$, and generate $y_t^{*(b)} = \sum_{j=1}^{p} \widehat{A}_j y_{t-j}^{*(b)} + u_t^{*(b)}$ recursively. Re-estimate LP and VAR on

each pseudo-sample and recompute the averaged estimator to obtain $\widehat{\theta}_h^{*(b)} = \widehat{\theta}_h^{*(b)}(\widehat{w}_h^{*(b)})$. A convenient centered $1-\alpha$ interval is $\left[\widehat{\theta}_h(\widehat{w}_h) - \delta_{\alpha,h},\ \widehat{\theta}_h(\widehat{w}_h) + \delta_{\alpha,h}\right]$, where $\delta_{\alpha,h}$ is the empirical $(1-\alpha)$ quantile of $\left|\widehat{\theta}_h^{*(b)} - \widehat{\theta}_h(\widehat{w}_h)\right|$ across $b = 1, \ldots, B$.

4 Monte Carlo Evidence

To assess the finite-sample performance of our methods, we consider both a simple univariate design and a richer multivariate design. The univariate exercise is useful for transparently illustrating the LP–VAR bias–variance trade-off and for evaluating how well the feasible procedures approximate the oracle weights. The multivariate exercise then assesses the same methods in a more realistic macroeconomic environment. In both the univariate and the multivariate exercises we use a fixed lag-selection rule; thus our estimators are not asymptotically equivalent in the Plagborg-Møller–Wolf sense. Although our theoretical results assume that both methods estimate the true IRF consistently, in the simulations we also explore what happens when the model is misspecified, showing that our proposed estimators still perform relatively well.

4.1 An Univariate ARMA Design

We begin with a simple univariate DGP, also considered by Montiel Olea et al. (2025). This design allows us to (i) verify that the infeasible oracle estimator averaging behaves as predicted by the LP–VAR bias–variance trade-off, (ii) evaluate how well the sieve-bootstrap and flexible implementations approximate the oracle weights in finite samples, and (iii) compare the risk properties of estimator averaging with those of R^2-based model averaging.

4.1.1 Model and Estimators

We conduct 1,000 Monte Carlo replications of a univariate ARMA(1,1) process

$$y_t \;=\; \rho\, y_{t-1} \;+\; \varepsilon_t \;+\; \alpha\, \varepsilon_{t-1}, \qquad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0,1). \tag{4.1}$$

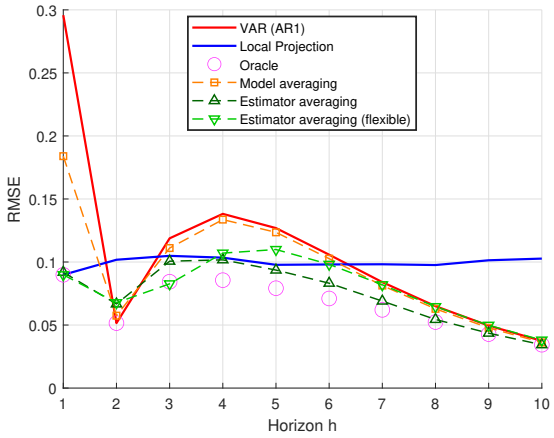
Each replication uses a burn-in period of 200 observations to reach stationarity before retaining a sample of length T. The true impulse response to a one-unit innovation at time t is $\theta_0 = 1$, $\theta_1 = \rho + \alpha$, $\theta_h = \rho\,\theta_{h-1}$ for $h \geq 2$, that is, $\theta_h = \rho^h + \alpha\rho^{h-1}$ for $h \geq 1$. We evaluate horizons $h \in \{1, \ldots, H_{\max}\}$ and set $H_{\max} = 10$. We consider $\rho \in \{0.5, 0.9\}$ and $\alpha \in \{0.5, 0.9\}$ for various sample sizes T. At each horizon, we report results for the following estimators:(1) the local projection estimator (LP), using the specification in Montiel Olea et al. (2025), (2) the VAR estimator, implemented as an AR(1), (3) the infeasible oracle averaged estimator based on MSE minimization, (4) the sieve-bootstrap plug-in estimator averaging procedure, (5) the sieve-bootstrap flexible estimator averaging procedure, and (6) the model-averaging estimator based on maximizing R^2. To approximate the oracle weight and implement the feasible averaging procedures, we use 500 bootstrap draws.

4.1.2 Horizon-Specific Risk and Weights

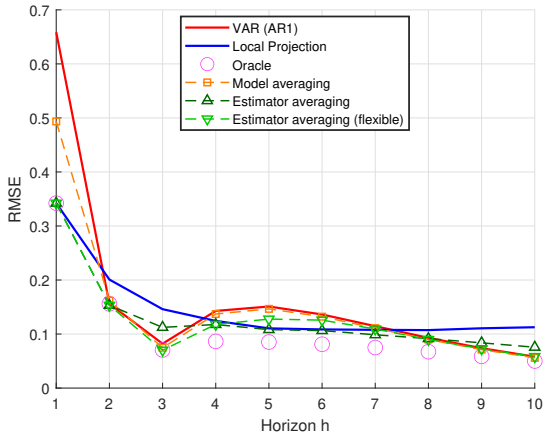
We first fix $T = 240$, an empirically relevant sample size used in Montiel Olea et al. (2025), and study horizons $h = 1, \ldots, H_{\max}$. This allows us to trace how the LP–VAR bias–variance trade-off evolves across horizons and to assess whether the oracle and feasible averaging rules shift weight from LP at short horizons toward VAR at longer horizons, as predicted by the MSE decomposition in (2.4).

Figures 1–2 report RMSE and weights across (ρ, α). The patterns line up closely with the bias–variance logic in (2.4). Under the ARMA(1,1) DGP in (4.1), $\theta_1 = \rho+\alpha$ and $\theta_h = \rho^h+\alpha\rho^{h-1}$ for $h \geq 2$, whereas the AR(1)-based VAR estimator implies $\widehat{\theta}_{VAR,h} \approx \widehat{\beta}^h$. Hence, when $\alpha > 0$,

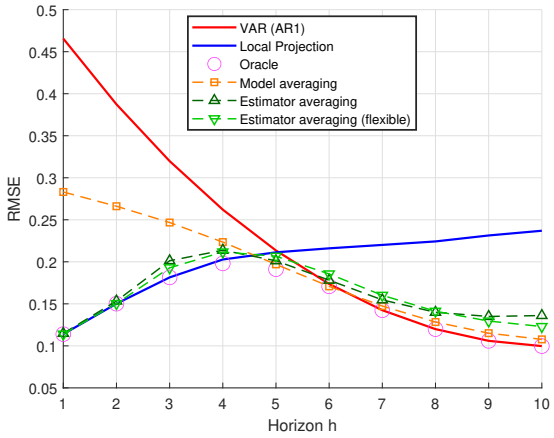
the VAR is downward biased at short horizons, with the bias decaying roughly at rate $\alpha\rho^{h-1}$, while LP is approximately unbiased at short horizons but becomes increasingly variable as h rises.



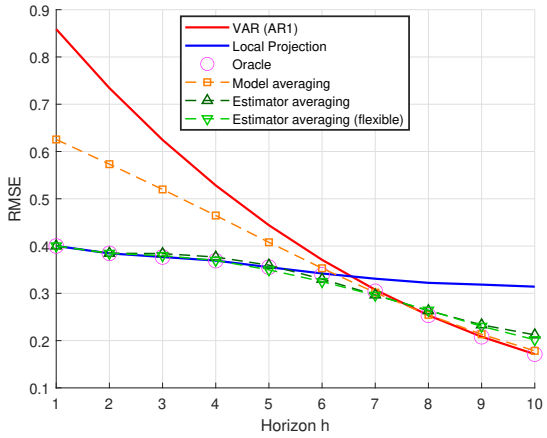
(a) $\rho = 0.5,\ \alpha = 0.5$



(b) $\rho = 0.5,\ \alpha = 0.9$



(c) $\rho = 0.9,\ \alpha = 0.5$



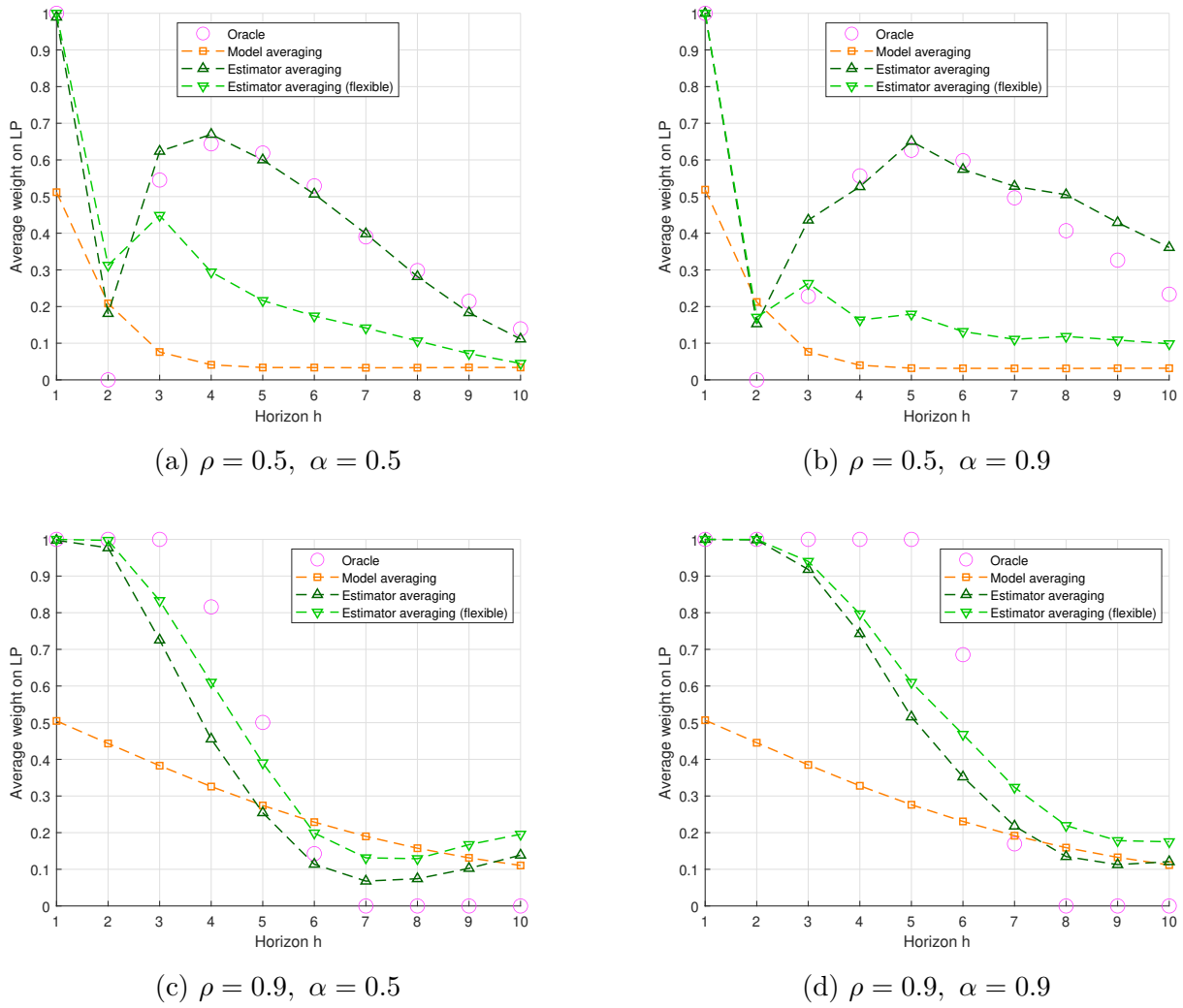
(d) $\rho = 0.9,\ \alpha = 0.9$

Figure 1: RMSE of IRF estimators across (ρ, α) designs, $T = 240$, 1,000 replications.

The figures confirm this trade-off. LP performs better at short horizons. When α is sizable, the difference in RMSE between the LP and VAR estimators is noticeably larger. Roughly after 7 periods, the VAR takes over because LP variance rises while VAR bias decays. A higher value of ρ slows the decay of the VAR bias, thus keeping its RMSE elevated longer.

The combined estimator using oracle weights yields an RMSE that is never larger—and is in fact somewhat smaller—than the minimum of the individual LP and VAR estimators at each

horizon. The feasible MSE-based estimator averaging approaches closely follow the performance of the oracle, though they naturally cannot match it perfectly. In contrast, the model averaging estimator sometimes exhibits substantially worse performance in terms of RMSE compared to the MSE-based methods, a discrepancy that is especially pronounced at short horizons.



(a) $\rho = 0.5,\ \alpha = 0.5$ (b) $\rho = 0.5,\ \alpha = 0.9$

(c) $\rho = 0.9,\ \alpha = 0.5$ (d) $\rho = 0.9,\ \alpha = 0.9$

Figure 2: Estimated weights across (ρ, α) designs, $T = 240$, 1,000 replications.

The reason for this difference in performance becomes obvious from the weights. The oracle weights in (2.5) display the benchmark pattern: they place more mass on LP at short horizons and then gradually shift toward VAR as h increases. The feasible MSE-based estimator averaging broadly reproduces this pattern, and its RMSE remains close to the oracle benchmark.

The under-performance of the model averaging estimator is due to the inflexibility of its weighting scheme based on R^2. Because this approach compares the in-sample fit of the VAR with that of the LP, it relies on an R^2_{VAR} that remains constant across all horizons and an $R^2_{LP,h}$ that exhibits very little variation as h increases. Consequently, these R^2-based weights are not flexible enough to capture the complex, shifting balance between bias and variance. While the scheme correctly recognizes that the relative advantage of the VAR estimator increases with the horizon, it adjusts too rigidly: the weight assigned to the LP estimator starts near 0.5 at $h = 1$ and declines only gradually as the horizon extends. This excessively smooth trajectory prevents the model averaging approach from adapting to the sharper, horizon-specific changes in the true risk profile.

Overall, the feasible MSE-based procedures successfully capture the bias–variance trade-off predicted by (2.4)–(2.5): they assign more weight to the LP where bias reduction matters, and shift toward the VAR where variance dominates. Across all horizons, these estimator-averaging approaches closely track the best attainable oracle benchmark, delivering modest to substantial risk reductions, depending on the horizon and the design, relative to relying on a single method. By contrast, the R^2-based model averaging is too rigid to adapt to this shifting risk profile, resulting in suboptimal performance throughout the projection horizon.

4.1.3 Finite-Sample Convergence

To complement the horizon-profile evidence above and connect the simulations to our large-sample theory, we vary the sample size $T \in \{200, 400, 800\}$ and focus on representative horizons $h \in \{1, 3, 6\}$. For each sample size, we perform 1,000 Monte Carlo replications.

Using the same ARMA(1,1) DGP in (4.1) and the same set of estimators, we report the RMSE of the estimated weights in Table 1 and the RMSE of the corresponding IRF estimators in Table 2. We again consider the four (α, ρ) designs.

Two main messages emerge. First, the estimated LP–VAR weights become more accurate

Table 1: RMSE of estimator averaging weights relative to oracle

	$h=1$		$h=3$		$h=6$	
T	$\widehat{w}_P$	$\widehat{w}_F$	$\widehat{w}_P$	$\widehat{w}_F$	$\widehat{w}_P$	$\widehat{w}_F$
			$\alpha=0.5,\ \rho=0.5$			
200	0.0656	0.0659	0.2393	0.4276	0.2907	0.3518
400	0.0079	0.0123	0.1650	0.3143	0.3330	0.3548
800	0.0017	0.0025	0.0963	0.2097	0.2782	0.2917
			$\alpha=0.5,\ \rho=0.9$			
200	0.0003	0.0000	0.3568	0.5617	0.2970	0.3367
400	0.0000	0.0000	0.2828	0.4165	0.2902	0.2885
800	0.0000	0.0000	0.1953	0.2662	0.2294	0.2096
			$\alpha=0.9,\ \rho=0.5$			
200	0.0091	0.4503	0.3643	0.0544	0.3891	0.4040
400	0.0021	0.3025	0.2637	0.0196	0.2746	0.3801
800	0.0000	0.0942	0.1692	0.0000	0.0898	0.4238
			$\alpha=0.9,\ \rho=0.9$			
200	0.0012	0.0169	0.2896	0.2417	0.4361	0.4109
400	0.0000	0.0000	0.1060	0.0970	0.6305	0.5412
800	0.0000	0.0000	0.0000	0.0000	0.4683	0.4152

Note: $\widehat{w}_P$ denotes the sieve-bootstrap plug-in estimator averaging weight from Algorithm 1. $\widehat{w}_F$ denotes the flexible sieve-bootstrap weight by choosing optimal (2.7).

Table 2: RMSE of IRF estimates relative to true IRF

	$h=1$						$h=3$						$h=6$					
T	$\widehat{\theta}_{LP}$	$\widehat{\theta}_{VAR}$	$\widehat{\theta}_{O}$	$\widehat{\theta}_{P}$	$\widehat{\theta}_{F}$	$\widehat{\theta}_{M}$	$\widehat{\theta}_{LP}$	$\widehat{\theta}_{VAR}$	$\widehat{\theta}_{O}$	$\widehat{\theta}_{P}$	$\widehat{\theta}_{F}$	$\widehat{\theta}_{M}$	$\widehat{\theta}_{LP}$	$\widehat{\theta}_{VAR}$	$\widehat{\theta}_{O}$	$\widehat{\theta}_{P}$	$\widehat{\theta}_{F}$	$\widehat{\theta}_{M}$
$\alpha=0.5,\ \rho=0.5$																		
$T=200$	0.0958	0.2972	0.0958	0.0990	0.0979	0.1149	0.1136	0.1204	0.0911	0.1084	0.1116	0.1177	0.1125	0.1070	0.0760	0.0886	0.0973	0.1044
$T=400$	0.0827	0.2929	0.0827	0.0832	0.0830	0.0924	0.0822	0.1155	0.0676	0.0785	0.0801	0.0896	0.0741	0.1029	0.0616	0.0760	0.0778	0.0808
$T=800$	0.0705	0.2881	0.0705	0.0706	0.0706	0.0754	0.0580	0.1166	0.0507	0.0554	0.0564	0.0617	0.0539	0.1035	0.0469	0.0604	0.0600	0.0620
$\alpha=0.5,\ \rho=0.9$																		
$T=200$	0.3446	0.6597	0.3446	0.3446	0.3446	0.3581	0.1525	0.0842	0.0754	0.1202	0.1377	0.1296	0.1238	0.1369	0.0874	0.1185	0.1212	0.1223
$T=400$	0.3425	0.6566	0.3425	0.3425	0.3425	0.3491	0.1266	0.0762	0.0563	0.0923	0.1025	0.1120	0.0820	0.1335	0.0686	0.0891	0.0865	0.0908
$T=800$	0.3346	0.6527	0.3346	0.3346	0.3346	0.3378	0.1034	0.0748	0.0420	0.0659	0.0703	0.0926	0.0605	0.1346	0.0512	0.0663	0.0629	0.0669
$\alpha=0.9,\ \rho=0.5$																		
$T=200$	0.1186	0.4679	0.1186	0.1197	0.1240	0.1304	0.1941	0.3261	0.1941	0.2119	0.2025	0.2367	0.2449	0.1846	0.1835	0.1909	0.1939	0.2063
$T=400$	0.1086	0.4613	0.1086	0.1087	0.1095	0.1145	0.1503	0.3080	0.1503	0.1632	0.1597	0.1882	0.1715	0.1520	0.1454	0.1505	0.1504	0.1569
$T=800$	0.0983	0.4586	0.0983	0.0983	0.0983	0.1012	0.1142	0.3002	0.1142	0.1159	0.1159	0.1365	0.1257	0.1354	0.1193	0.1239	0.1212	0.1287
$\alpha=0.9,\ \rho=0.9$																		
$T=200$	0.4025	0.8608	0.4025	0.4025	0.4029	0.4112	0.3834	0.6300	0.3834	0.3951	0.3882	0.4477	0.3676	0.3802	0.3573	0.3498	0.3424	0.3672
$T=400$	0.3995	0.8550	0.3995	0.3995	0.3995	0.4038	0.3582	0.6143	0.3582	0.3587	0.3585	0.4000	0.3006	0.3530	0.3006	0.2902	0.2832	0.3223
$T=800$	0.3922	0.8526	0.3922	0.3922	0.3922	0.3943	0.3313	0.6078	0.3313	0.3313	0.3313	0.3511	0.2677	0.3410	0.2677	0.2606	0.2582	0.2997

Note: $\widehat{\theta}_{LP}$ and $\widehat{\theta}_{VAR}$ are the LP and VAR IRF estimators. $\widehat{\theta}_{O}$ uses the infeasible oracle min-MSE weight $w_h^\star$ in (2.5). $\widehat{\theta}_{P}$ uses the sieve-bootstrap plug-in weight $\widehat{w}_P$ (Algorithm 1). $\widehat{\theta}_{F}$ uses the flexible sieve-bootstrap weight w_F in (2.7). $\widehat{\theta}_{M}$ uses the R^2-based model-averaging weight $\widehat{w}_M$ in (2.8).

as T increases. In Table 1, the RMSE of both the plug-in weight $\widehat{w}_P$ and the flexible weight $\widehat{w}_F$ (measured relative to the oracle weights) generally declines with T. It is important to note that while Theorem 1 establishes the formal consistency of these weights, and a high-level rate under an additional scaled-risk rate condition, for the baseline case where both the LP and VAR estimators are consistent, this specific simulation design features misspecified estimators. Thus, rather than merely illustrating the theorem, these numerical results demonstrate an important broader finding: the feasible weighting procedures maintain strong finite-sample performance—evidenced by a substantially shrinking RMSE—even in the presence of underlying model misspecification. This improvement is especially clear at short horizons, where the LP–VAR trade-off is most informative for risk estimation.

Second, the averaged IRF estimators based on these estimated weights inherit the same convergence and deliver strong risk performance in finite samples. In Table 2, the estimator-averaging procedures $\widehat{\theta}_P$ and $\widehat{\theta}_F$ move toward the oracle benchmark $\widehat{\theta}_O$ as T increases, in line with Theorem 2. Across designs, estimator averaging implements the intended bias–variance trade-off: it places more weight on LP where VAR misspecification bias is most relevant, especially at short horizons, and shifts toward VAR as the horizon increases and LP variance becomes more important.

Overall, these additional results reinforce the main findings from the univariate design: bootstrap-based estimator averaging converges toward oracle averaging and yields low RMSE relative to LP, VAR, and R^2-based model averaging when the objective is to minimize IRF estimation risk.

4.2 A Multivariate SVARMA Design

We next extend the analysis to a multivariate setting to evaluate the performance of the estimators in a more realistic macroeconomic environment. We consider two DGPs: an SVAR(4) model and an SVARMA(4,1) model. In the latter case, the finite-order VAR and LP models

used for estimating the impulse response functions are misspecified. In both cases, we assume that the structural shocks are observed so as to abstract from identification issues and focus directly on the estimation of impulse responses.

4.2.1 Model and Estimators

The data are generated from a three-variable system ($n = 3$) following a general SVARMA(p, q) process:

$$y_t = \sum_{j=1}^{p} A_j y_{t-j} + \sum_{k=0}^{q} M_k \varepsilon_{t-k}, \qquad \varepsilon_t \sim \mathcal{N}(0, I_3), \tag{4.2}$$

where M_0 is the structural impact matrix. We consider two specifications:

1. **SVAR(4):** $p = 4$, $q = 0$. In this case, a VAR with sufficient lag length can capture the true dynamics.

2. **SVARMA(4,1):** $p = 4$, $q = 1$. This introduces a moving-average component; consequently, finite-order VAR and LP approximations remain formally misspecified, even when lag lengths are selected via information criteria.

The specific coefficient matrices (A_j, M_k) are reported in Appendix C. We simulate 1,000 replications and, for the main finite-sample analysis, set $T = 200$.

We evaluate the following estimators:

1. **Local Projection (LP):** the lag length is set equal to the lag order chosen for the VAR.

2. **VAR:** the lag order is selected by AIC, with a maximum lag length of 8.

3. **Oracle:** the infeasible averaged estimator based on MSE minimization, computed using 500 bootstrap draws.

4. **Estimator Averaging:** the feasible MSE-based averaged estimator, where the weights are estimated via a sieve bootstrap with the DGP approximated by a VAR selected

by AIC. We use 500 bootstrap replications. To save space, we report only the plug-in estimator since the flexible estimator yields similar results.

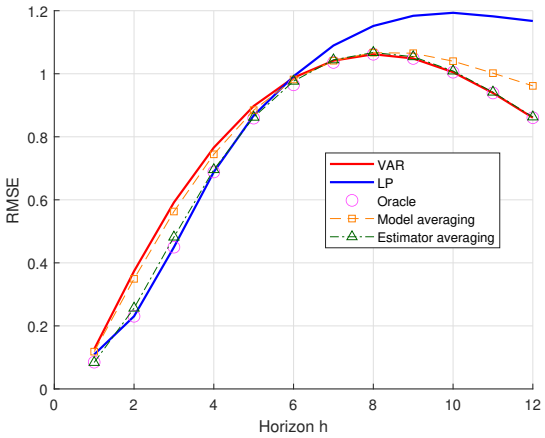
5. **Model Averaging:** the averaged estimator with weights chosen to maximize in-sample R^2 at each horizon.

4.2.2 Horizon-Specific Risk and Weights

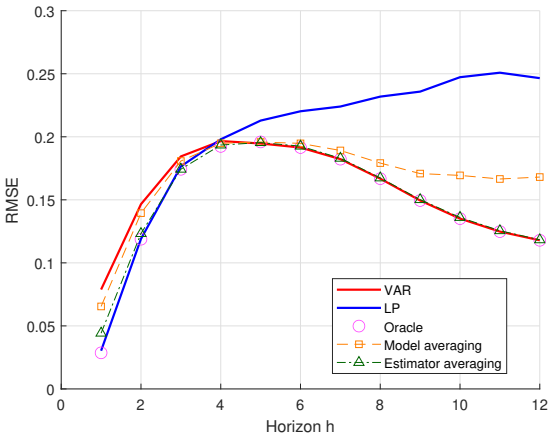
Figure 3 reports RMSE and estimated weights for the two DGPs. In the case of SVAR(4), the bias–variance trade-off is clearly visible; see Panels (a) and (c). LP delivers lower RMSE for horizons $h = 2$ to $h = 6$, after which VAR becomes superior because its structured dynamics produces greater efficiency. The oracle estimator closely tracks the lower envelope of the two individual estimators.

In terms of RMSE, the feasible estimator-averaging procedure is nearly indistinguishable from the oracle, indicating that the risk-based weighting rule is well estimated in finite samples. By contrast, model averaging performs noticeably worse, with higher RMSE at most horizons. The weight plot in Panel (c) makes the reason clear: estimator averaging tracks the shape of the oracle weights, assigning more weight to LP at short horizons and then shifting toward VAR, whereas model averaging tends to underweight LP early and overweight it later because it is driven by in-sample fit rather than IRF estimation risk.

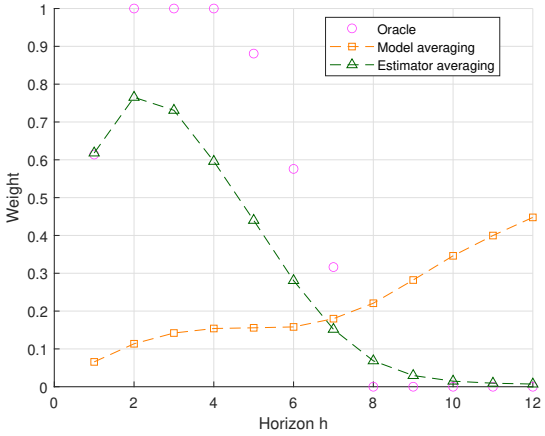
In the misspecified case (the SVARMA(4,1) DGP), LP retains its advantage at short horizons ($h < 4$), while VAR performs better thereafter; see Panels (b) and (d). The oracle again tracks the minimum RMSE, and the estimator averaging remains close to that oracle benchmark. Model averaging, however, performs less well, especially at horizons where the LP–VAR gap is large. In particular, the R^2-based rule fails to reproduce the relatively rapid decline in oracle LP weight, leading to suboptimal averaging.



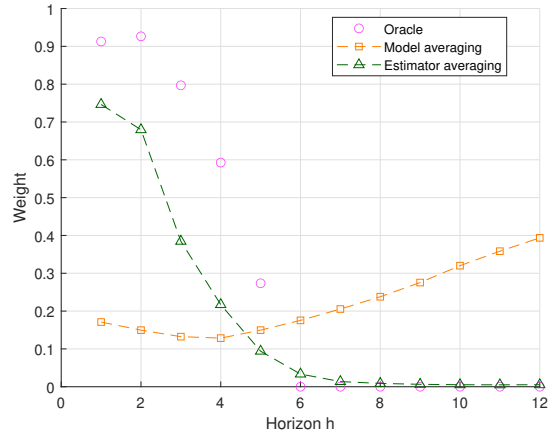
(a) RMSE: SVAR(4)



(b) RMSE: SVARMA(4,1)



(c) Weights: SVAR(4)



(d) Weights: SVARMA(4,1)

Figure 3: Root mean squared errors (RMSE) and average weights on LP for two multivariate data-generating processes. Results are based on a sample size of $T = 200$, using 1,000 Monte Carlo replications and 500 bootstrap iterations.

4.2.3 Finite-sample convergence

To examine the large-sample behavior in the multivariate setting, we vary the sample size $T \in \{200, 800, 2000\}$ and report results for horizons $h \in \{1, 6, 12\}$.

Table 3 presents the RMSE of the impulse response estimates based on 1,000 Monte Carlo replications and 500 bootstrap iterations. Consistent with the univariate evidence, the RMSE of estimator averaging declines with sample size and remains below that of model averaging across almost all specifications. Even at $T = 2000$, where LP and VAR have both become quite accurate, estimator averaging typically yields a slight improvement by optimally combining their remaining finite-sample differences.

Table 4 reports the RMSE of the estimated weights relative to the oracle weights. Unlike the IRF estimates, the weights do not converge monotonically in all cases, particularly under SVARMA(4,1). This may be related to a flat or weakly identified weighting problem, which presumably occurs when the VAR and LP estimates converge toward each other. In such cases, estimator averaging can remain close to risk-optimal even when the estimated weights display discernible finite-sample variation.

5 Empirical Application

Section V of Bauer and Swanson (2023) reassess the dynamic effects of monetary policy by addressing two key challenges in high-frequency identification: instrument relevance and exogeneity. To improve relevance, they expand the standard set of monetary policy events beyond FOMC announcements to include press conferences, speeches, and testimony by the Federal Reserve Chair, substantially increasing the variation of the surprise series. To ensure exogeneity, they argue that conventional high-frequency surprises suffer from endogeneity because they are systematically correlated with publicly available macroeconomic and financial data predating the announcements—rather than being driven by central bank "information effects." To address

Table 3: RMSE of multivariate IRF estimators

	$h=1$					$h=6$					$h=12$				
T	$\widehat{\theta}_{VAR}$	$\widehat{\theta}_{LP}$	$\widehat{\theta}_{O}$	$\widehat{\theta}_{P}$	$\widehat{\theta}_{M}$	$\widehat{\theta}_{VAR}$	$\widehat{\theta}_{LP}$	$\widehat{\theta}_{O}$	$\widehat{\theta}_{P}$	$\widehat{\theta}_{M}$	$\widehat{\theta}_{VAR}$	$\widehat{\theta}_{LP}$	$\widehat{\theta}_{O}$	$\widehat{\theta}_{P}$	$\widehat{\theta}_{M}$
SVAR(4)															
200	0.1299	0.1071	0.0847	0.0906	0.1215	1.0344	1.0445	1.0168	1.0247	1.0320	0.8433	1.1363	0.8433	0.8445	0.9390
800	0.0330	0.0562	0.0291	0.0272	0.0311	0.4253	0.4439	0.4231	0.4216	0.4240	0.4653	0.5627	0.4653	0.4662	0.4888
2000	0.0144	0.0339	0.0135	0.0126	0.0137	0.2649	0.2803	0.2647	0.2639	0.2649	0.3141	0.3428	0.3141	0.3140	0.3167
SVARMA(4,1)															
200	0.0802	0.0302	0.0289	0.0427	0.0671	0.1955	0.2205	0.1955	0.1964	0.1969	0.1182	0.2449	0.1182	0.1183	0.1656
800	0.0218	0.0148	0.0124	0.0143	0.0186	0.0985	0.1065	0.0985	0.0985	0.0985	0.0586	0.1174	0.0586	0.0587	0.0642
2000	0.0094	0.0091	0.0067	0.0071	0.0081	0.0612	0.0657	0.0612	0.0613	0.0612	0.0383	0.0699	0.0383	0.0383	0.0385

Note: $\widehat{\theta}_{VAR}$ and $\widehat{\theta}_{LP}$ are the VAR and LP IRF estimators. $\widehat{\theta}_{O}$ uses the infeasible oracle min-MSE weight $w_h^\star$ in (2.5). $\widehat{\theta}_{P}$ uses the sieve-bootstrap plug-in weight $\widehat{w}_P$ (Algorithm 1). $\widehat{\theta}_{M}$ uses the R^2-based model-averaging weight $\widehat{w}_M$ in (2.8). Results are based on 1,000 Monte Carlo replications and 500 bootstrap iterations.

Table 4: RMSE of estimator averaging weights relative to oracle

T	$h=1$	$h=6$	$h=12$
SVAR(4)			
200	0.2491	0.3027	0.0151
800	0.1404	0.1738	0.0288
2000	0.0911	0.1405	0.0477
SVARMA(4,1)			
200	0.3053	0.0894	0.0108
800	0.2882	0.1141	0.0182
2000	0.2973	0.1137	0.0184

this, they orthogonalize the high-frequency surprises with respect to these pre-announcement variables and use the resulting residual as an external instrument for the monetary policy shock. They then estimate IV-LP and IV-SVAR impulse responses of yields, activity, prices, and financial conditions, finding that this correction produces stronger and more plausible macroeconomic estimates.

We replicate their baseline setup using the same monthly data set, orthogonalized high-frequency surprise, and IV specification. We study the responses of the two-year Treasury yield (GBY), industrial production (IP), consumer prices (CPI), and the excess bond premium (EBP) to a 25-basis-point contractionary monetary policy shock. For each variable, we compute IV-LP and IV-SVAR impulse responses based on their specifications, as well as model-averaging combinations of the two. Finally, we also compute the external-instrument version of the estimator-averaging combination described in Remark 3. In the empirical application, we present only the plug-in version, as the flexible estimator averaging yields very similar estimates.

For inference, we employ a nested VAR-sieve wild bootstrap procedure. We first approximate the data-generating process by estimating a reduced-form VAR model, where the lag order is selected via the Bayesian information criterion. To handle potential heteroskedasticity and accommodate the missing observations in the instrument, we apply a wild bootstrap to the data. Specifically, to preserve the identifying contemporaneous correlation between the reduced-form VAR residuals and the external instrument, the same sequence of random wild multipliers—

drawn from a Rademacher distribution—is applied simultaneously to both the VAR residuals and the instrument during resampling. Our procedure relies on a double bootstrap architecture in which both the outer and inner loops consist of 500 iterations. The inner bootstrap loop is utilized to approximate the finite-sample variances and biases of the individual IV-SVAR and IV-LP estimators, as well as their covariance, which provides the moments necessary to calculate the optimal weights for the combined estimator. Subsequently, the outer bootstrap loop generates the empirical distribution of all the estimators. The confidence bands are computed by extracting the middle 68 percent of the bootstrap estimates centered around the point estimates from the original sample. Figure 4 displays the resulting impulse responses; Figure 5 shows the corresponding weights on LP. Appendix D also presents the impulse responses with confidence bands for all estimators.

Our baseline IV-VAR and IV-LP estimates successfully replicate the results documented in Bauer and Swanson (2023) (see their Figures 3 and A2, respectively). The IV-VAR responses (red lines) are smooth and tightly shaped: GBY and EBP jump on impact and then gradually return toward zero, IP exhibits a modest hump-shaped decline, and CPI shows a prolonged disinflation. By contrast, the IV-LP responses (blue lines) are much more volatile. For GBY and EBP, LP IRFs oscillate and change sign several times; for IP the decline is steeper and very persistent; for CPI the response eventually turns positive and drifts upward, implying an implausible long-run rise in price level after a contractionary shock. This stark contrast between low-variance but potentially biased IV-VAR estimates and low-bias but noisy IV-LP estimates mirrors the bias–variance trade-off highlighted in our simulations and motivates the use of averaging estimators.

The averaging procedures stabilize the IRFs while preserving their main qualitative features. In many cases, they tilt toward the more economically plausible trajectory. Estimator averaging (green dashed lines with triangles) pulls the paths toward the smoother IV-VAR trajectories wherever the LP estimates are extremely noisy, but still allows for some deviations where the LP and VAR disagree. For GBY, the estimator-averaged IRF shows a sharp increase in the

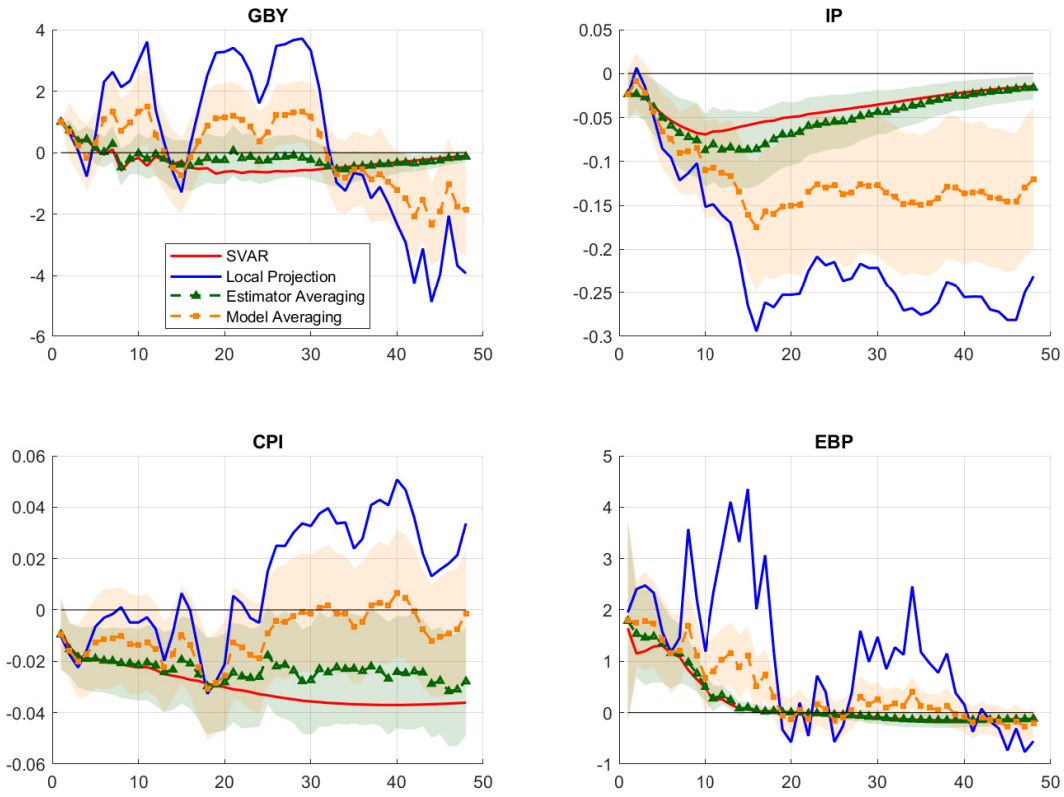


Figure 4: Point estimates of the impulse responses to a 25-basis-point contractionary monetary policy shock. Each panel compares the four estimation approaches for a given macroeconomic variable.

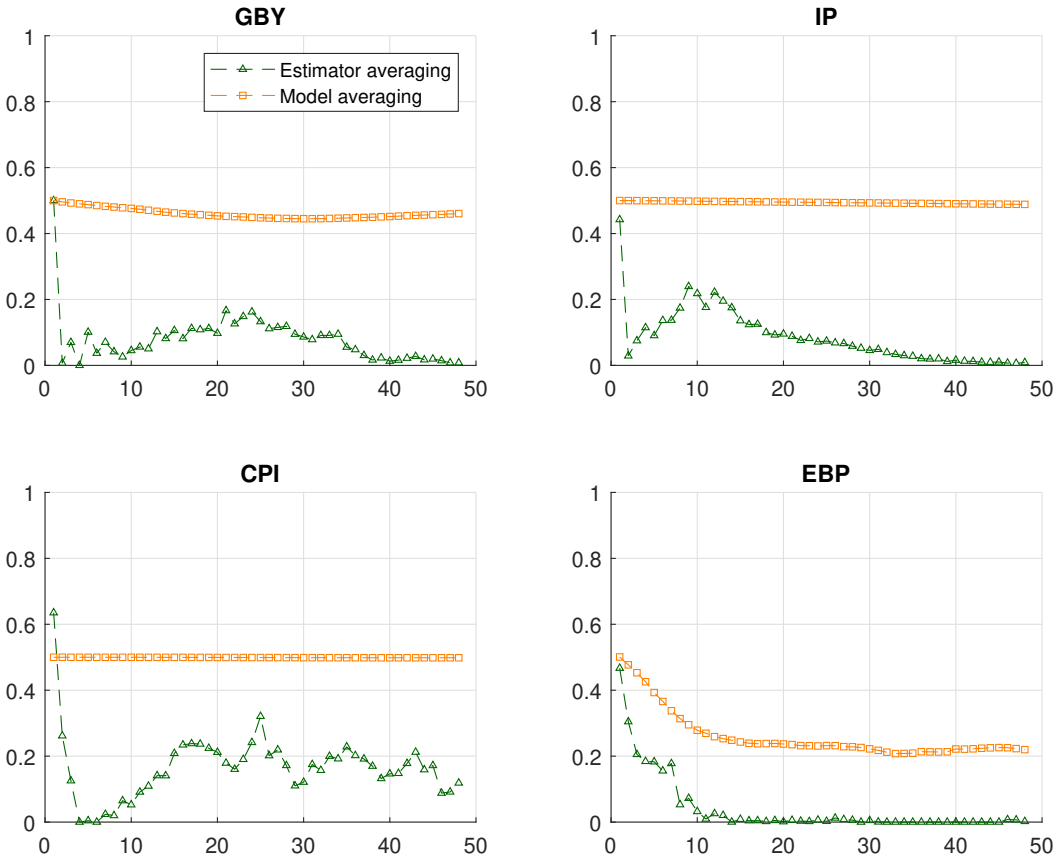


Figure 5: Estimated weights on the IV-LP estimator across the model-averaging and estimator-averaging approaches for each macroeconomic variable.

two-year yield that decays within a year, avoiding the large negative swings of the LP while not over-smoothing the near-term reaction. Furthermore, unlike the VAR response, it never sinks deeply into negative territory, a profile that is economically plausible following a tightening shock. For IP, estimator averaging yields a moderate, hump-shaped decline that lies between the highly persistent LP response and the more muted VAR response. For CPI, the combined estimator produces a smaller and less protracted disinflation than the VAR, while successfully avoiding the "price puzzle" anomaly exhibited by the LP. For EBP, estimator averaging delivers a sharp, short-run rise in credit spreads that quickly mean-reverts, eliminating the extreme volatility of the LP estimator while preserving the intuitive tightening in financial conditions after a monetary contraction.

Model averaging based on R^2 (orange lines with squares) behaves differently. As Figure 5 shows, the R^2-based weights on LP are relatively flat across horizons—around one-half for GBY, IP, and CPI, and somewhat lower for EBP. This reflects the fact that LP and VAR achieve similar in-sample fit even when the LP is estimated for longer horizons. Consequently, the model-averaged IRFs remain more heavily influenced by LP at longer horizons. For GBY and EBP, this yields more volatile responses than estimator averaging. The trajectory of the former drops substantially into negative territory after three years; for CPI, the effect disappears entirely after two years; and for IP, the decline is deeper and more persistent. Model averaging therefore improves on raw IV-LP by shrinking its most extreme movements, but it does not fully correct the long-horizon instability generated by the LP estimator and does not eliminate the puzzles.

Figure 5 also makes clear why estimator averaging tends to deliver the most intuitive IRFs. The estimator-averaging weights on LP are high only on impact and in the very first few months, then quickly decay toward zero as the horizon increases, especially for IP and EBP. This pattern reflects the empirical fact that identification is strongest and LP variance is smallest at very short horizons, whereas the VAR's parametric structure provides more reliable long-run dynamics. The resulting estimator-averaged IRFs are therefore economically appealing: a

front-loaded, temporary rise in GBY; a transitory fall in IP; a small and delayed disinflation in CPI; and a pronounced but short-lived increase in EBP. These responses sit between IV-LP and IV-VAR where the two disagree, dampen LP's long-horizon noise, and avoid the overly smooth extremes of VAR, illustrating the practical usefulness of estimator averaging in applied monetary policy analysis.

6 Conclusion

LP and VAR are the two workhorse methods for impulse response analysis, and their relative appeal is fundamentally a finite-sample question. LP is often attractive because of its robustness and comparatively low bias, whereas VAR is often attractive because of its greater precision. This paper studies how to combine these two estimators through horizon-specific estimator averaging, with weights chosen to minimize the mean squared error of the structural impulse response itself rather than the in-sample fit of the underlying regression.

We derive closed-form oracle weights that make transparent how the optimal LP share depends on the relative bias, variance, and covariance of LP and VAR, and we develop feasible AR-sieve-bootstrap implementations to estimate these weights in practice. The Monte Carlo results show that estimator averaging can deliver meaningful MSE gains relative to LP and VAR alone, especially because its inherent flexibility allows it to precisely track the horizon-specific dynamics of the bias–variance trade-off. In contrast, the fit-based model-averaging approach is less flexible and performs worse in our design.

In an empirical application revisiting the high-frequency IV monetary policy shocks of Bauer and Swanson (2023), estimator averaging yields IRFs for yields, activity, prices, and credit spreads that are stable, economically intuitive, and lie between the often volatile IV-LP estimates and the very smooth IV-VAR estimates. Overall, our results suggest that estimator averaging provides a practical and easy-to-implement complement to existing LP and VAR practice, especially for empirical researchers who want to discipline the finite-sample bias–variance

trade-off directly at the level of the impulse response of interest.

References

Bauer, M. D. and Swanson, E. T. (2023). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, 37(1):87–155.

Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, 3(2):123–148.

Gonçalves, S. (2007). Asymptotic and bootstrap inference for AR(∞) processes with conditional heteroskedasticity. *Econometric Theory*, 23(4):849–889.

Gonçalves, S. and Kilian, L. (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, 123(1):89–120.

Hounyo, U. and Jung, S. (2025). Two-stage model averaging for impulse responses: Local projections- and vars-based approaches. SSRN Working Paper No. 5309694.

Jordà, O. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, 95(1):161–182.

Kilian, L. and Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge University Press, Cambridge, UK.

Kreiss, J.-P. and Paparoditis, E. (2003). Autoregressive aided periodic bootstrap for time series. *Annals of Statistics*, 31(6):1923–1955.

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.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin.

Mittelhammer, R. C. and Judge, G. G. (2005). Combining estimators to improve structural model estimation and inference under quadratic loss. *Journal of econometrics*, 128(1):1–29.

Montiel Olea, J. L., Plagborg-Møller, M., Qian, E., and Wolf, C. K. (2025). Local projections or vars? a primer for macroeconomists. Working Paper 33871, National Bureau of Economic Research.

Montiel Olea, J. L., Plagborg-Møller, M., Qian, E., and Wolf, C. K. (2026). Double robustness of local projections and some unpleasant varithmetic. *arXiv preprint arXiv:2405.09509*.

Nemtyrev, A. and Boldea, O. (2026). Targeted local projections. *arXiv preprint arXiv:2603.00248*.

Plagborg-Møller, M. and Wolf, C. K. (2021). Local projections and vector autoregressions. *Econometrica*, 89(2):955–980.

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

Appendix A Lemmas

Lemma 1 (Consistency and joint asymptotic normality of $(\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})$)**.** *For any fixed h, under Assumption 1 and the regularity conditions summarized in Assumption 2, $(\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})' \xrightarrow{p} (\theta_h, \theta_h)'$, and there exists a possibly singular 2×2 matrix $\Omega_h^{(2)}$ such that*

$$\sqrt{T}\Big((\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})' - (\theta_h, \theta_h)'\Big) \Rightarrow \mathcal{N}(0, \Omega_h^{(2)}). \tag{A.1}$$

Moreover, if $\Omega_h^{(2)} = \begin{pmatrix} \Omega_{11,h} & \Omega_{12,h} \\ \Omega_{12,h} & \Omega_{22,h} \end{pmatrix}$, then, under the moment condition in Assumption 2,

$$T\,\mathrm{Var}(\widehat{\theta}_{LP,h}) \to \Omega_{11,h}, \qquad T\,\mathrm{Var}(\widehat{\theta}_{VAR,h}) \to \Omega_{22,h}, \qquad T\,\mathrm{Cov}(\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h}) \to \Omega_{12,h}. \tag{A.2}$$

Thus, in the benchmark centered root-T setting, $(A_{T,h}, D_{T,h}, F_{T,h}) \to (A_h, D_h, F_h) = (\Omega_{11,h}, \Omega_{22,h}, \Omega_{12,h})$.

Proof. This is standard, and we only indicate the main ingredients and references.

For fixed h, the LP(h) estimator is an OLS coefficient from a finite-dimensional regression with weakly dependent regressors and errors under Assumption 1. A law of large numbers for the sample Gram matrix and a central limit theorem for the OLS score yield root-T consistency and asymptotic normality of $\widehat{\theta}_{LP,h}$; see, e.g., Jordà (2005).

The VAR(p) IRF estimator at fixed horizon h is a smooth functional of the VAR OLS estimator, including the VAR coefficients and the innovation covariance. Under stability and standard rank conditions, the VAR OLS estimator is asymptotically normal, and the delta method yields root-T asymptotic normality of $\widehat{\theta}_{VAR,h}$; see, e.g., Lütkepohl (2005) and Kilian and Lütkepohl (2017).

Joint asymptotic normality follows by stacking the two asymptotically linear representations and applying a multivariate CLT, equivalently a joint CLT for the combined influence function. The resulting covariance matrix $\Omega_h^{(2)}$ may be singular, which is allowed.

Finally, the moment condition in Assumption 2 implies uniform integrability of the quadratic forms of $Z_{T,h} = \sqrt{T}\big((\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})' - (\theta_h, \theta_h)'\big)$. Therefore, the second moments of $Z_{T,h}$ converge to the corresponding second moments of the Gaussian limit. These second moments are precisely $T\,\mathrm{Var}(\widehat{\theta}_{LP,h})$, $T\,\mathrm{Var}(\widehat{\theta}_{VAR,h})$, and $T\,\mathrm{Cov}(\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})$. In the benchmark centered root-T setting, the first-order bias terms are negligible, so the scaled risk components converge to the corresponding entries of $\Omega_h^{(2)}$. □

Lemma 2 (Consistency and rate of scaled bootstrap risk components)**.** *For any fixed h, define the scaled bootstrap variance-covariance components by $\widehat{A}_{T,h} = T\widehat{V}^{SV}_{LP,h}$, $\widehat{D}_{T,h} = T\widehat{V}^{SV}_{VAR,h}$, and $\widehat{F}_{T,h} = T\widehat{C}^{SV}_{h}$. Under Assumption 3,*

$$(\widehat{A}_{T,h}, \widehat{D}_{T,h}, \widehat{F}_{T,h}) \xrightarrow{p} (A_h, D_h, F_h). \tag{A.3}$$

If, in addition, the scaled bootstrap risk components satisfy

$$\left\|(\widehat{A}_{T,h}, \widehat{D}_{T,h}, \widehat{F}_{T,h}) - (A_h, D_h, F_h)\right\| = O_p(r_T) + O_p(B^{-1/2}), \tag{A.4}$$

for some deterministic sequence $r_T \to 0$, then the scaled risk components converge at that rate.

Proof of Lemma 2. Before proving the lemma, we clarify the role of the AR-sieve pseudo-truth. Under the short-memory Wold assumptions and the AR-sieve lag-order condition, with $p_T \to \infty$ sufficiently slowly relative to T, the AR-sieve approximation targets the same population impulse response in the benchmark setting where LP and VAR are both consistent. This motivates the finite-sample bootstrap bias correction used in Algorithm 1. However, first-order bootstrap validity does not by itself imply consistency of the bootstrap bias estimator, which is a higher-order object. Therefore, Lemma 2 only uses AR-sieve validity to establish consistency of the scaled bootstrap variance-covariance components. A formal consistency result for the bootstrap bias terms would require additional higher-order conditions, which we do not impose.

Let $Z^*_{T,h} = \sqrt{T}\big((\widehat{\theta}^*_{LP,h}, \widehat{\theta}^*_{VAR,h})' - (\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})'\big)$, and write $Z^*_{T,h} = (Z^*_{L,T,h}, Z^*_{V,T,h})'$. The

scaled bootstrap variance and covariance estimators can be written as sample second moments of the bootstrap draws:

$$\widehat{A}_{T,h} = \frac{1}{B}\sum_{b=1}^{B}\left(Z^{*(b)}_{L,T,h} - \overline{Z}^{*}_{L,T,h}\right)^2, \tag{A.5}$$

$$\widehat{D}_{T,h} = \frac{1}{B}\sum_{b=1}^{B}\left(Z^{*(b)}_{V,T,h} - \overline{Z}^{*}_{V,T,h}\right)^2, \tag{A.6}$$

$$\widehat{F}_{T,h} = \frac{1}{B}\sum_{b=1}^{B}\left(Z^{*(b)}_{L,T,h} - \overline{Z}^{*}_{L,T,h}\right)\left(Z^{*(b)}_{V,T,h} - \overline{Z}^{*}_{V,T,h}\right). \tag{A.7}$$

Assumption 3(i) implies that the conditional distribution of $Z^{*}_{T,h}$ consistently approximates the distribution of $Z_{T,h}$ in the bounded-Lipschitz metric. Together with the conditional moment bound in Assumption 3(ii), this gives convergence of the relevant conditional quadratic moments. In particular,

$$\mathbb{E}^{*}[(Z^{*}_{L,T,h})^2] \xrightarrow{p} \Omega_{11,h}, \qquad \mathbb{E}^{*}[(Z^{*}_{V,T,h})^2] \xrightarrow{p} \Omega_{22,h}, \qquad \mathbb{E}^{*}[Z^{*}_{L,T,h}Z^{*}_{V,T,h}] \xrightarrow{p} \Omega_{12,h}. \tag{A.8}$$

Equivalently, after recentering by the bootstrap mean, the conditional covariance matrix of $Z^{*}_{T,h}$ converges in probability to $\Omega^{(2)}_{h}$.

Conditional on the data, the bootstrap draws are i.i.d. across b. Hence, as $B \to \infty$, the conditional LLN gives convergence of the bootstrap sample second moments to their conditional expectations. Combining this with (A.8) yields

$$(\widehat{A}_{T,h}, \widehat{D}_{T,h}, \widehat{F}_{T,h}) \xrightarrow{p} (\Omega_{11,h}, \Omega_{22,h}, \Omega_{12,h}) = (A_h, D_h, F_h). \tag{A.9}$$

The displayed rate is not implied by first-order bootstrap validity alone. If the high-level scaled-risk rate condition (A.4) holds, then the scaled bootstrap risk components converge at that rate. This rate is carried forward to the plug-in weight in Theorem 1. □

Appendix B Proof of Theorems

Proof of Theorem 1. Define $G(A, D, F) = (D-F)/(A+D-2F)$, and let $\bar{G}(x) = \min\{1, \max\{0, x\}\}$ denote clipping. Then $\widehat{w}_h = \bar{G}\Big(G(\widehat{A}_{T,h}, \widehat{D}_{T,h}, \widehat{F}_{T,h})\Big)$ and $w_h^\star = \bar{G}(G(A_h, D_h, F_h))$.

By Lemma 2,

$$(\widehat{A}_{T,h}, \widehat{D}_{T,h}, \widehat{F}_{T,h}) \xrightarrow{p} (A_h, D_h, F_h). \tag{B.1}$$

Assumption 5 implies $A_h + D_h - 2F_h \geq c > 0$. Therefore, G is continuous in a neighborhood of (A_h, D_h, F_h). Since the limiting solution is interior, clipping is asymptotically inactive. The continuous mapping theorem gives $\widehat{w}_h \xrightarrow{p} w_h^\star$.

For the rate statement, suppose that (3.6) holds. Because G is continuously differentiable in a neighborhood of (A_h, D_h, F_h) and its denominator is bounded away from zero, a mean-value expansion yields

$$\begin{aligned} |\widehat{w}_h - w_h^\star| &\leq C \left\| (\widehat{A}_{T,h}, \widehat{D}_{T,h}, \widehat{F}_{T,h}) - (A_h, D_h, F_h) \right\| \\ &= O_p(r_T) + O_p(B^{-1/2}). \end{aligned} \tag{B.2}$$

This proves the theorem. □

Proof of Theorem 2. Define $\widehat{\theta}_h(w) = w\widehat{\theta}_{LP,h} + (1-w)\widehat{\theta}_{VAR,h}$. Then

$$\widehat{\theta}_h(\widehat{w}_h) - \theta_h = \big(\widehat{\theta}_h(w_h^\star) - \theta_h\big) + (\widehat{w}_h - w_h^\star)(\widehat{\theta}_{LP,h} - \widehat{\theta}_{VAR,h}). \tag{B.3}$$

For consistency, Lemma 1 gives $(\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})' \xrightarrow{p} (\theta_h, \theta_h)'$. Together with $\widehat{w}_h \xrightarrow{p} w_h^\star$ from Theorem 1, this implies $\widehat{\theta}_h(\widehat{w}_h) \xrightarrow{p} \theta_h$.

For asymptotic normality, multiply (B.3) by $\sqrt{T}$:

$$\sqrt{T}\big(\widehat{\theta}_h(\widehat{w}_h) - \theta_h\big) = \sqrt{T}\big(\widehat{\theta}_h(w_h^\star) - \theta_h\big) + (\widehat{w}_h - w_h^\star)\sqrt{T}(\widehat{\theta}_{LP,h} - \widehat{\theta}_{VAR,h}). \tag{B.4}$$

By Lemma 1, $\sqrt{T}(\widehat{\theta}_{LP,h} - \widehat{\theta}_{VAR,h}) = O_p(1)$, and by Theorem 1, $\widehat{w}_h - w_h^\star = o_p(1)$. Thus the second term in (B.4) is $o_p(1)$.

For the first term,

$$\sqrt{T}\big(\widehat{\theta}_h(w_h^\star) - \theta_h\big) = e(w_h^\star)'\sqrt{T}\Big((\widehat{\theta}_{LP,h}, \widehat{\theta}_{VAR,h})' - (\theta_h, \theta_h)'\Big), \tag{B.5}$$

where $e(w_h^\star) = (w_h^\star, 1 - w_h^\star)'$. By Lemma 1 and the continuous mapping theorem,

$$\sqrt{T}\big(\widehat{\theta}_h(w_h^\star) - \theta_h\big) \Rightarrow \mathcal{N}\Big(0,\, e(w_h^\star)'\Omega_h^{(2)} e(w_h^\star)\Big). \tag{B.6}$$

Since $e(w_h^\star)'\Omega_h^{(2)} e(w_h^\star) = V_h(w_h^\star, \Omega_h^{(2)})$, Slutsky's theorem gives the stated limit for $\widehat{\theta}_h(\widehat{w}_h)$. □

Proof of Theorem 3. The map $(w, \Omega) \mapsto V_h(w, \Omega)$ is a polynomial in w and the entries of Ω, hence continuous. By Theorem 1, $\widehat{w}_h \xrightarrow{p} w_h^\star$, and by assumption, $\widehat{\Omega}_h^{(2)} \xrightarrow{p} \Omega_h^{(2)}$. The continuous mapping theorem gives

$$\widehat{V}_h \xrightarrow{p} V_h(w_h^\star, \Omega_h^{(2)}). \tag{B.7}$$

For the rate statement, suppose that $|\widehat{w}_h - w_h^\star| = O_p(r_T) + O_p(B^{-1/2})$ and $\|\widehat{\Omega}_h^{(2)} - \Omega_h^{(2)}\| = O_p(s_T) + O_p(B^{-1/2})$ for some $s_T \to 0$. A mean-value expansion of $V_h(w, \Omega)$ gives

$$\begin{aligned} |\widehat{V}_h - V_h(w_h^\star, \Omega_h^{(2)})| &\le C_1|\widehat{w}_h - w_h^\star| + C_2\|\widehat{\Omega}_h^{(2)} - \Omega_h^{(2)}\| \\ &= O_p(r_T + s_T) + O_p(B^{-1/2}). \end{aligned} \tag{B.8}$$

□

Appendix C Multivariate DGP Details

The coefficient matrices for the multivariate simulations in Section 4.2 are defined as follows. For both specifications, the dimension is $n = 3$, the autoregressive order is $p = 4$, and the structural impact matrix M_0 is defined below.

SVAR(4) Specification

The autoregressive coefficients $A_1, \ldots, A_4$ are:

$$A_1 = \begin{pmatrix} 1.31 & 0.75 & 0.25 \\ -0.12 & 2.08 & 0.23 \\ -0.23 & 0.56 & 1.75 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} -0.52 & -1.06 & -0.35 \\ 0.16 & -1.59 & -0.33 \\ 0.32 & -0.78 & -1.12 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 0.04 & 0.48 & 0.16 \\ -0.08 & 0.53 & 0.15 \\ -0.14 & 0.35 & 0.31 \end{pmatrix}, \qquad A_4 = \begin{pmatrix} 0.01 & -0.07 & -0.02 \\ 0.01 & -0.06 & -0.02 \\ 0.02 & -0.05 & -0.03 \end{pmatrix}.$$

The structural impact matrix M_0 is:

$$M_0 = \begin{pmatrix} 2.0 & -1.5 & 0.2 \\ 1.7 & 1.3 & 0.7 \\ 0.6 & -0.6 & 1.7 \end{pmatrix}$$

SVARMA(4,1) Specification

The autoregressive coefficients $A_1, \ldots, A_4$ are:

$$A_1 = \begin{pmatrix} 1.24 & -0.04 & -0.03 \\ -0.58 & 1.77 & 0.32 \\ -0.78 & 0.76 & 1.63 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} -0.52 & 0.02 & 0.06 \\ 0.74 & -1.23 & -0.39 \\ 1.04 & -0.98 & -1.02 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 0.08 & 0.00 & -0.03 \\ -0.30 & 0.39 & 0.16 \\ -0.44 & 0.41 & 0.29 \end{pmatrix}, \qquad A_4 = \begin{pmatrix} -0.01 & 0.00 & 0.00 \\ 0.04 & -0.04 & -0.02 \\ 0.06 & -0.05 & -0.03 \end{pmatrix}.$$

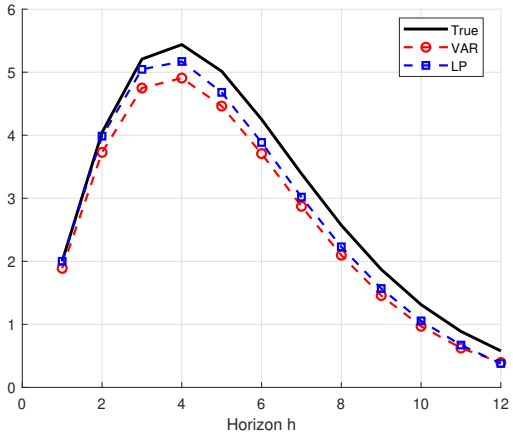
The moving average matrix M_1 is:

$$M_1 = \begin{pmatrix} -0.30 & 0.10 & -0.40 \\ -0.20 & 0.20 & -1.00 \\ -0.30 & 0.07 & 0.20 \end{pmatrix}$$

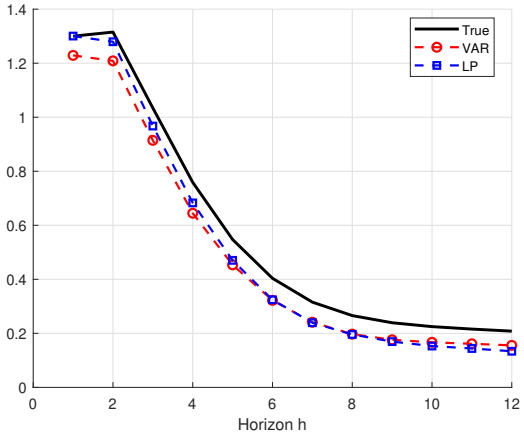
The structural impact matrix M_0 is:

$$M_0 = \begin{pmatrix} 1.30 & 0.40 & 0.10 \\ -0.02 & 0.05 & 2.00 \\ -0.08 & -1.70 & 0.80 \end{pmatrix}$$

Figure 6 displays the true impulse response functions alongside the average estimates from the VAR and LP estimators ($T = 200$).



(a) SVAR(4)



(b) SVARMA(4,1)

Figure 6: True impulse responses and average VAR and LP estimates for two multivariate data-generating processes. Results are based on a sample size of $T = 200$, using 1,000 Monte Carlo replications and 500 bootstrap iterations. VAR lag order is selected via AIC and is also applied to the LP estimations.

Appendix D Empirical Impulse Response Functions with Confidence Band

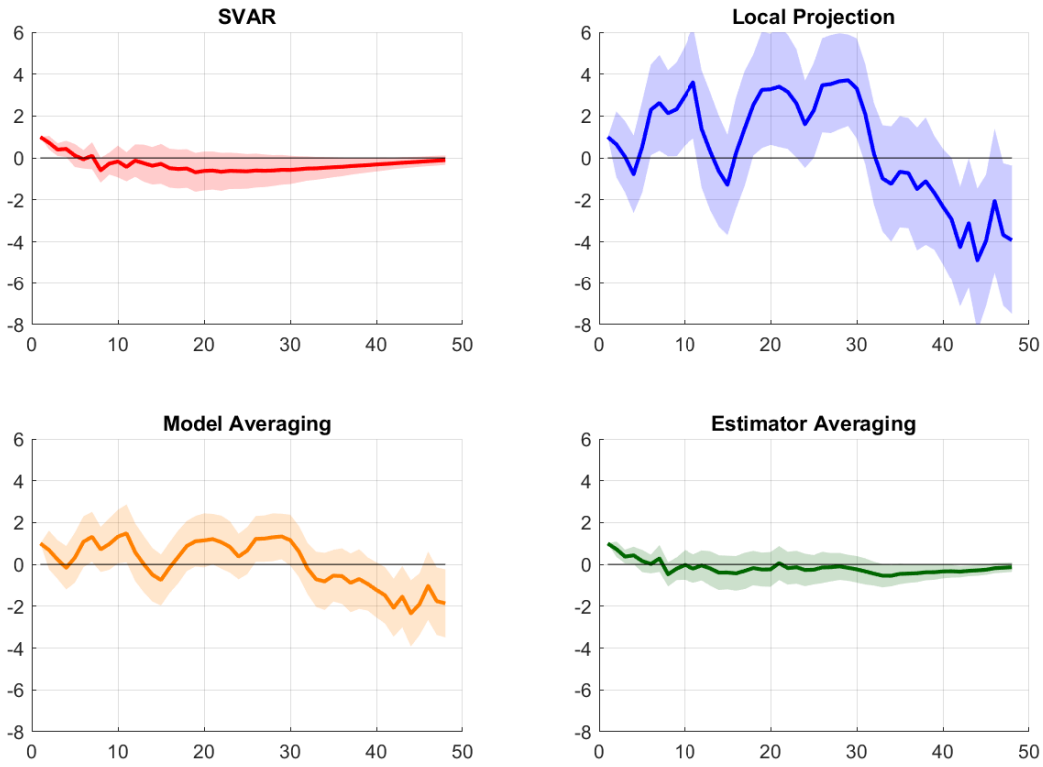


Figure 7: Estimated responses of the two-year Treasury yield to a 25-basis-point contractionary monetary policy shock. The four panels display the estimates from the IV-SVAR, IV-LP, model-averaging, and estimator-averaging approaches. The 68 percent confidence bands are calculated using a nested VAR-sieve wild bootstrap procedure (500 inner and outer iterations) and are centered around the point estimates.

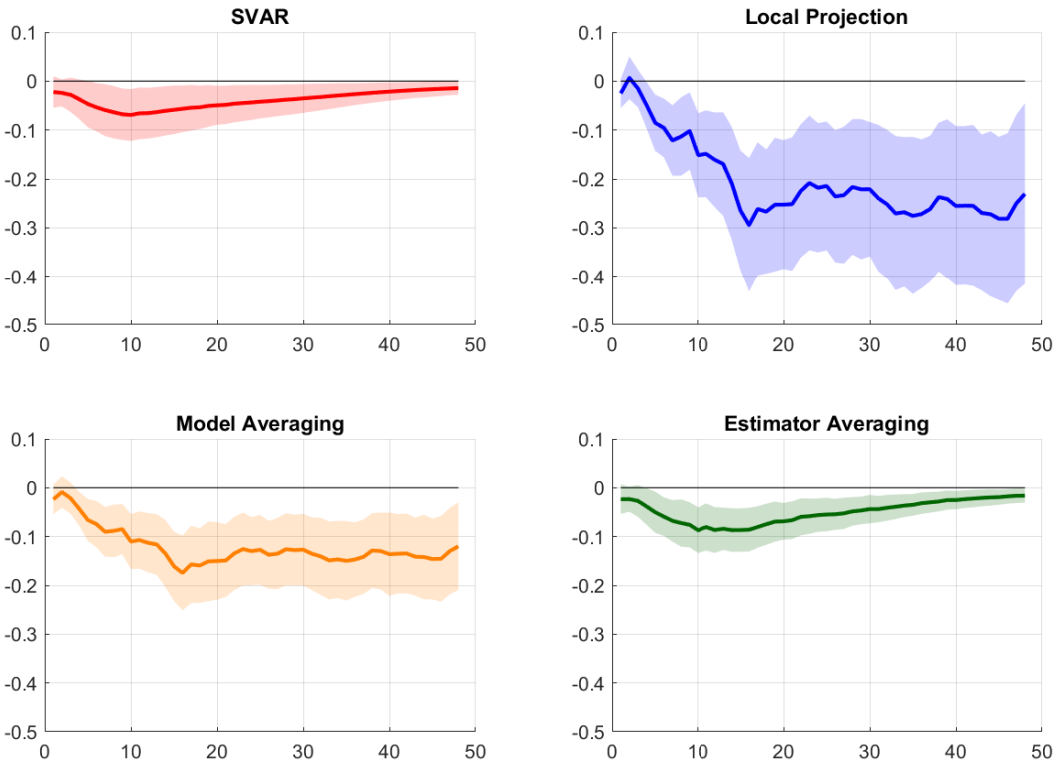


Figure 8: Estimated responses of industrial production to a 25-basis-point contractionary monetary policy shock. The four panels display the estimates from the IV-SVAR, IV-LP, model-averaging, and estimator-averaging approaches. The 68 percent confidence bands are calculated using a nested VAR-sieve wild bootstrap procedure (500 inner and outer iterations) and are centered around the point estimates.

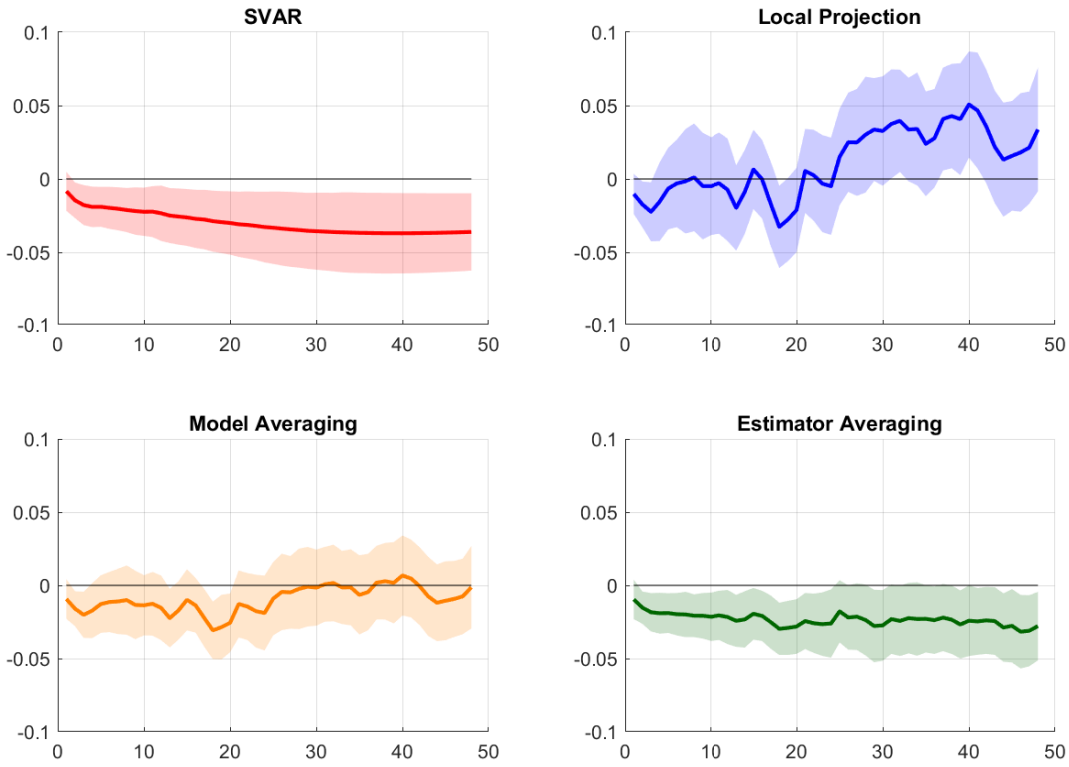


Figure 9: Estimated responses of consumer prices to a 25-basis-point contractionary monetary policy shock. The four panels display the estimates from the IV-SVAR, IV-LP, model-averaging, and estimator-averaging approaches. The 68 percent confidence bands are calculated using a nested VAR-sieve wild bootstrap procedure (500 inner and outer iterations) and are centered around the point estimates.

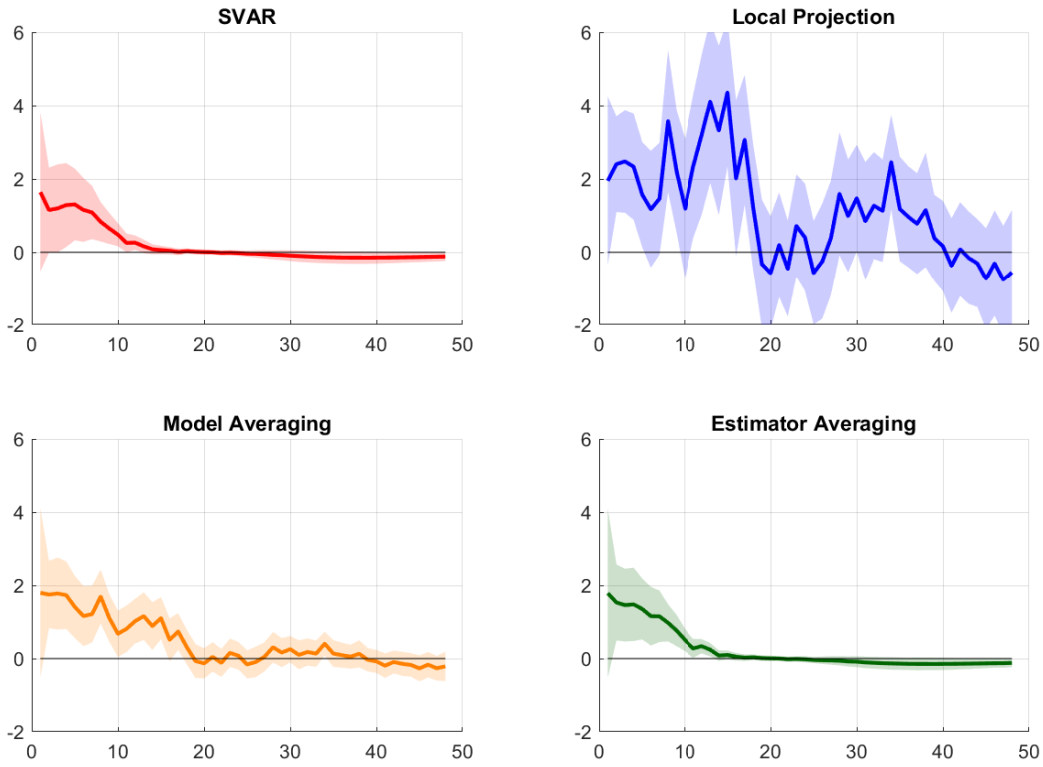


Figure 10: Estimated responses of the excess bond premium to a 25-basis-point contractionary monetary policy shock. The four panels display the estimates from the IV-SVAR, IV-LP, model-averaging, and estimator-averaging approaches. The 68 percent confidence bands are calculated using a nested VAR-sieve wild bootstrap procedure (500 inner and outer iterations) and are centered around the point estimates.