Scanning for Significance: False Discovery Control for Impulse Responses*

Giorgi Nikolaishvili
Wake Forest University
nikolag@wfu.edu

Noah D. Gade
Wake Forest University
gaden@wfu.edu

This Version: March 11, 2026
[Preliminary and Incomplete]

Abstract

Empirical impulse response analysis builds economic narratives by scanning dozens or hundreds of coefficients for significant effects. Standard pointwise inference ignores this multiplicity, so the false rejection rate grows unbounded with the response family. Simultaneous inference methods correct for multiplicity by bounding the probability of even a single false rejection, which yields increasingly uninformative results as the family expands. Researchers are left to choose between overstating their evidence and understating it. To address this limitation, we propose false discovery rate (FDR) and false coverage rate (FCR) control as a more appropriate target: bounding the expected share of false rejections among responses declared significant, with calibrated post-selection confidence intervals. Neither guarantee deteriorates as the response family grows, so researchers are rewarded for investigating transmission mechanisms thoroughly rather than penalized for it. The procedure integrates into standard VAR and local projection bootstrap workflows. Applications to monetary policy and oil supply shocks show that FDR/FCR-controlled inference recovers effects lost under simultaneous bands while discarding fragile pointwise findings, in some cases materially altering the economic narrative.

JEL Codes: C12; C22; C32; E00
Keywords: impulse response; multiple testing; false discovery; VAR; local projection

*We thank Jeremy Piger, Kurt Lunsford, Aeimit Lakdawala, Cynthia Wu, Tatevik Sekhposyan, and David Evans for helpful comments. We also thank Jimmy Ren for excellent research assistance. Computations were performed using the Wake Forest University High Performance Computing Facility. All errors are our own.

1 Introduction

Impulse response functions (IRFs) are among the most widely reported empirical objects in macroeconomics and finance, used to build narratives about how structural shocks propagate through the economy. A typical application estimates responses across dozens or hundreds of horizon–variable–shock–regime combinations, and researchers scan the resulting grid for significant effects, identifying which responses are nonzero, when they peak, and how long they persist. The resulting economic narrative often rests on pointwise confidence bands that offer no multiplicity control: each band is valid in isolation, but as the number of reported responses grows, so does the expected number of false rejections.[1] Even a modestly sized exercise produces a large hypothesis family: a five-variable vector autoregression (VAR) with a single identified shock examined over thirty-six horizons yields 185 coefficients, and state-dependent or multi-shock extensions easily involve several hundred. A growing body of work addresses this problem with simultaneous inference methods that control the family-wise error rate (FWER), the probability of even a single false rejection anywhere in the response family. However, this target is too strict for a setting in which researchers can tolerate a small proportion of false inclusions: FWER-controlling bands widen as the response family grows, often to the point where they fail to detect nearly any true dynamic effects, discouraging researchers from investigating transmission mechanisms thoroughly. The result is an uncomfortable choice between overstating evidence and understating it — one that most researchers resolve in favor of pointwise bands, leaving the resulting economic narratives exposed to an unknown share of false rejections.

This paper argues that false discovery rate (FDR) and false coverage rate (FCR) control provide the appropriate inferential target for impulse response analysis. FDR control bounds the expected proportion of false rejections among those responses declared significant: if the procedure selects twenty significant responses at a 5% FDR level, roughly one of those twenty is expected to be a false discovery. FCR control complements this guarantee by delivering post-selection confidence intervals for the selected responses: if twenty intervals are reported, roughly nineteen are expected to cover their true parameter values. Together, they discipline the economic narrative against false discoveries while preserving the power to detect true dynamic effects, resolving the tradeoff that has

[1] To name a few familiar applications: Christiano et al. (2005) and Romer and Romer (2004) build their monetary policy transmission narratives by reading significance patterns across variables and horizons; Gertler and Karadi (2015) identify a persistent credit-spread response by scanning both dimensions of a proxy SVAR grid; Ramey (2011) and Ramey and Zubairy (2018) adjudicate the fiscal multiplier debate by inspecting significance across horizons and variables, with Auerbach and Gorodnichenko (2012, 2013) and Ramey and Zubairy (2018) further expanding the hypothesis family through state dependence.

constrained what researchers can credibly learn from impulse response analysis. Crucially, neither guarantee becomes more conservative as the dimension of the response family grows, so a researcher who examines six variables over fifty horizons faces the same reliability standard for their reported findings as one who examines two variables over twenty horizons. Adopting this framework changes the empirical workflow in a specific way: the researcher still estimates the same high-dimensional IRF grid, but no longer overlays pointwise bands and informally selects salient features. Instead, they apply an FDR-controlling selection rule to the full collection of test statistics and obtain a formal rejection set — the subset of responses that can be declared significantly different from zero after accounting for the scan across horizons, variables, shocks, and states. These selected responses are then reported with confidence intervals calibrated for false coverage rate control.

The existing literature on impulse response inference has approached multiplicity exclusively through simultaneous bands that target FWER control. Analytic approaches based on Scheffé-type and Bonferroni adjustments have been proposed for VARs (Jordà, 2009; Lütkepohl et al., 2015, 2020), and resampling-based methods using sup-t critical values have been developed for both VAR and LP estimators (Inoue and Kilian, 2016; Bruder and Wolf, 2018; Montiel Olea and Plagborg-Møller, 2019; Jordà, 2023; Jordà and Taylor, 2025; Inoue et al., 2026). This body of work has sharpened our understanding of the pointwise-versus-simultaneous tradeoff considerably, but it has also treated that tradeoff as exhaustive.[2] The possibility of targeting an intermediate error rate that accounts for multiplicity without requiring joint coverage of the entire response family, and that directly governs the reliability of the selected findings rather than the full grid, has not been explored in the IRF setting. Yet precisely such a framework exists in the statistics literature on multiple testing, where FDR control (Benjamini and Hochberg, 1995) and FCR control (Benjamini and Yekutieli, 2005) have become standard tools for high-dimensional inference problems with similar structure.

The contribution of this paper is first and foremost conceptual: we argue that the construction of economic narratives by scanning high-dimensional IRF grids is a multiple testing and selective reporting problem, and that the pointwise-versus-simultaneous

[2]IRF applications have a different structure from settings where FWER control is natural. A researcher testing a small pre-specified set of coefficients needs the guarantee that none of those rejections is a false positive. A researcher who scans a large grid to assemble a transmission narrative can tolerate a small number of incorrect inclusions provided their proportion is controlled, because some responses anchor the conclusions while others provide corroborating detail. From this perspective, the widening of simultaneous bands with family size is not merely a power problem but a symptom of target mismatch, and it creates a perverse incentive against thorough analysis: expanding the set of horizons, variables, shocks, or states mechanically reduces the chance of detecting any effect.

dichotomy the literature has treated as exhaustive overlooks the inferential framework best suited to it. FDR and FCR control fill this gap by directly governing the reliability of the reported findings rather than the full test family, providing a middle ground with formal multiplicity control that does not sacrifice power as the response family grows.

Building on this reframing, the paper makes three specific contributions. First, we adapt an FDR- and FCR-controlling procedure to impulse response inference. Because IRF test statistics exhibit dependence across horizons that need not satisfy the positive dependence conditions assumed by standard FDR-control procedures, we implement our approach using the resampling-based stepwise methodology of Romano et al. (2008), which accommodates arbitrary dependence by estimating the joint distribution of test statistics directly. We establish asymptotic validity of the procedure for both VAR and LP estimators under weak regularity conditions, showing that FDR and FCR are controlled at the nominal level in large samples. Second, we provide Monte Carlo evidence on finite-sample performance in designs calibrated to empirically relevant macroeconometric settings. Across both estimators, a range of sample sizes, and varying horizon ranges and sparsity configurations, the procedure delivers systematic power gains over simultaneous inference while maintaining FDR and FCR control at the target level, with the gains largest when the response family is high-dimensional. Third, we apply our methods to two prominent empirical settings: monetary policy shocks (Bu et al., 2021) and oil supply news shocks (Känzig, 2021). In both applications, FWER-controlling bands are largely uninformative, detecting few if any significant effects. FDR/FCR-controlled inference recovers substantial power: in the monetary policy application, it preserves the qualitative transmission narrative implied by pointwise inference while placing it on formally reliable footing; in the oil shock application, it retains the core findings but discards the pointwise evidence that oil supply shocks affect world industrial production, altering one dimension of the economic narrative.

Related literature. Our paper connects the statistics and econometrics literatures on multiple testing to the macroeconometrics literature on joint IRF inference. The foundational references on FDR control are Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001), with FCR control for post-selection confidence intervals introduced by Benjamini and Yekutieli (2005). Our baseline implementation uses the bootstrap stepdown procedure of Romano et al. (2008), which controls FDR under general dependence by estimating the joint distribution of the test statistic vector directly; this procedure builds on the broader resampling-based multiple-testing framework of Westfall and Young (1993), White (2000), Romano and Wolf (2005a,b), and Romano and Wolf (2007),

surveyed for econometric practice by Romano et al. (2010).

In economics and finance, FDR-style reasoning has been applied to mutual fund performance evaluation (Barras et al., 2010) and asset pricing anomaly discovery (Harvey and Liu, 2020), where strong cross-test dependence and selective reporting create concerns analogous to those arising in IRF families. More broadly, the applied microeconomics and program evaluation literatures have increasingly recognized that scanning across many outcomes, subgroups, or time periods requires multiplicity adjustment (Anderson, 2008; Lee and Shaikh, 2014; List et al., 2019), and Freyaldenhoven et al. (2019) address a closely related problem in panel event-study designs. The macroeconometrics literature has not had the same reckoning, even though the structure of the problem is directly analogous.

The IRF inference literature has approached multiplicity exclusively through FWER-controlling simultaneous bands, with the methodological arc moving from analytic to resampling-based approaches. Early work constructed Scheffé-type regions (Jordà, 2009) and Bonferroni or projection-based bands (Lütkepohl et al., 2015, 2020), establishing that pointwise bands can substantially undercover IRF paths. Bootstrap sup-t methods then emerged as a more powerful alternative for both structurally identified VARs (Inoue and Kilian, 2016; Bruder and Wolf, 2018) and local projections (Montiel Olea and Plagborg-Møller, 2019; Jordà, 2023; Jordà and Taylor, 2025; Inoue et al., 2026), and the literature has largely converged on this implementation. From a Bayesian perspective, Sims and Zha (1999) and Inoue and Kilian (2022) develop posterior-based joint credible regions that provide a posterior analog of FWER-style coverage, arriving at the same inferential target from different foundations. Frequentist and Bayesian approaches alike thus share a single inferential target: zero tolerance for false rejections anywhere in the response family. The question our paper raises is whether a different target, one that bounds the share of false rejections among the selected findings rather than prohibiting them entirely, is better matched to how researchers actually use IRF grids.

Outline. Section 2 uses simulations and an empirical monetary policy application to show that the pointwise-versus-simultaneous tradeoff is severe in practice: pointwise bands produce high rates of false rejection across the response grid, while simultaneous sup-t bands lose the ability to detect effects that are plausibly present in the data. Section 3 develops our FDR- and FCR-adjusted procedure, establishes its asymptotic properties for VAR and LP estimators, and shows through Monte Carlo experiments that it recovers a substantial share of pointwise power while maintaining false discovery control at the target level. Section 4 applies the methods to oil supply news shocks identified through OPEC announcement surprises, illustrating how FDR/FCR-controlled inference can alter

which transmission patterns the data support. Section 5 provides practical guidance for applied researchers. Section 6 concludes.

2 The Multiple Testing Problem with IRFs

This section documents the severity of the pointwise-versus-simultaneous tradeoff in impulse response inference. We first establish the relevant error-rate definitions and then turn to an empirical application to monetary policy shocks identified using the Bu et al. (2021) series, where pointwise bands support a rich narrative about the transmission of contractionary policy but simultaneous bands render nearly all of these findings insignificant. A controlled simulation design confirms that this tradeoff reflects a systematic pattern: pointwise confidence intervals produce high rates of false discovery and post-selection miscoverage that worsen as the response family grows, while simultaneous sup-t bands that control the FWER eliminate false discoveries at the cost of substantial power loss.

2.1 Preliminaries: Pointwise and Simultaneous Inference

Consider a family of m scalar impulse response coefficients $\theta_1, \ldots, \theta_m$, indexed across horizons, variables, and potentially shocks or states. Let $\widehat{\theta}_j$ denote the estimator of θ_j with standard error $\widehat{s}_j$. Let $\mathcal{H}_0 \subseteq \{1, \ldots, m\}$ denote the subset of true null hypotheses (those j for which $\theta_j = 0$) and let $\mathcal{H}_1 = \{1, \ldots, m\} \setminus \mathcal{H}_0$.

Pointwise inference. Pointwise inference targets *marginal* (one-dimensional) coverage and testing statements. A pointwise $(1 - \alpha_{\text{pt}})$ confidence interval CI_j^{pt} satisfies

$$\mathbb{P}\left(\theta_j \in CI_j^{\text{pt}}\right) \geq 1 - \alpha_{\text{pt}} \qquad \text{for each fixed } j, \tag{1}$$

and the associated two-sided pointwise test of

$$H_j : \ \theta_j = 0 \qquad \text{rejects when} \qquad 0 \notin CI_j^{\text{pt}}. \tag{2}$$

The multiple-testing problem arises because researchers rarely interpret a single j in isolation; rather, they scan across the full family of m coefficients and highlight those for

which H_j is rejected. Define the number of (pointwise) rejections

$$R \equiv \sum_{j=1}^{m} \mathbb{I}\{0 \notin CI_j^{\text{pt}}\}, \qquad V \equiv \sum_{j \in \mathcal{H}_0} \mathbb{I}\{0 \notin CI_j^{\text{pt}}\}, \tag{3}$$

where V counts *false* rejections among the true nulls. Even if each individual test has size α_{pt}, the FWER (probability of at least one false rejection)

$$\text{FWER} \equiv \mathbb{P}(V > 0), \tag{4}$$

can be large when $|\mathcal{H}_0|$ is large. Under independence across true null tests, FWER $= 1-(1-\alpha_{\text{pt}})^{|\mathcal{H}_0|}$, which approaches one quickly as $|\mathcal{H}_0|$ grows. More generally, without any independence, a union bound yields FWER $\leq |\mathcal{H}_0|\alpha_{\text{pt}}$, which is uninformative when $|\mathcal{H}_0|$ is large.

Thus, pointwise validity does not translate into reliable joint statements about the response path once the analysis implicitly ranges over a high-dimensional IRF family. A parallel issue arises for coverage: while $\mathbb{P}(\theta_j \in CI_j^{\text{pt}}) \approx 1-\alpha_{\text{pt}}$ for each j, the probability that *all* intervals cover simultaneously, $\mathbb{P}(\forall j \in \{1,\dots,m\} : \theta_j \in CI_j^{\text{pt}})$, is typically far below $1-\alpha_{\text{pt}}$ when m is large. This is why pointwise bands tend to undercover IRF trajectories.

Simultaneous inference. Simultaneous inference targets a *uniform* probability statement over a pre-specified family: a collection of intervals $\{CI_j^{\text{sim}}\}_{j=1}^{m}$ is a simultaneous $(1-\alpha_{\text{FWER}})$ band if

$$\mathbb{P}\Big(\theta_j \in CI_j^{\text{sim}} \text{ for all } j = 1,\dots,m\Big) \geq 1-\alpha_{\text{FWER}}. \tag{5}$$

Equivalently, if we test $H_j : \theta_j = 0$ by rejecting whenever $0 \notin CI_j^{\text{sim}}$, then

$$\mathbb{P}\Big(\exists j \in \mathcal{H}_0 :\ 0 \notin CI_j^{\text{sim}}\Big) \leq \alpha_{\text{FWER}}, \tag{6}$$

so the probability of making *any* false discovery in the entire IRF family is controlled.

In the IRF context, a convenient implementation is the sup-t approach. Let $T_j \equiv |\widehat{\theta}_j|/\widehat{s}_j$ and let $T_j^{*(b)}$ denote the corresponding studentized statistic computed on bootstrap draw b (with appropriate recentering). Define the bootstrap distribution of the maximum statistic

$$M^{*(b)} \equiv \max_{1 \leq j \leq m} T_j^{*(b)}. \tag{7}$$

Let $c_{1-\alpha_{\text{FWER}}}$ be the empirical $(1-\alpha_{\text{FWER}})$ quantile of $\{M^{*(b)}\}_{b=1}^{B}$. Then the two-sided sup-t

band takes the form

$$CI_j^{\text{sim}} = \left[\widehat{\theta}_j - c_{1-\alpha_{\text{FWER}}}\widehat{s}_j,\ \widehat{\theta}_j + c_{1-\alpha_{\text{FWER}}}\widehat{s}_j\right], \qquad j = 1, \ldots, m. \tag{8}$$

Intuitively, $c_{1-\alpha_{\text{FWER}}}$ inflates the usual pointwise critical value to account for searching over m coefficients and for dependence across horizons, variables, shocks, and states. Because it is tied to an extreme-value functional (a maximum), the critical value (and hence band width) typically increases with the effective dimension of the response family, generating power loss in large IRFs.

Error rates for evaluating inference procedures. To compare pointwise and simultaneous inference, it is useful to define error rates that measure the reliability of the set of "discoveries" (rejected hypotheses) and the confidence intervals attached to them. The false discovery proportion and false discovery rate are

$$\text{FDP} \equiv \begin{cases} V/R, & R > 0, \\ 0, & R = 0, \end{cases} \qquad \text{and} \qquad \text{FDR} \equiv \mathbb{E}[\text{FDP}]. \tag{9}$$

To assess post-selection coverage, let V^{CI} count the number of selected coefficients whose reported confidence interval fails to cover the true θ_j. The false coverage proportion and false coverage rate are

$$\text{FCP} \equiv \begin{cases} V^{CI}/R, & R > 0, \\ 0, & R = 0, \end{cases} \qquad \text{and} \qquad \text{FCR} \equiv \mathbb{E}[\text{FCP}]. \tag{10}$$

Finally, we report average power over $\mathcal{H}_1$ (the fraction of true non-null coefficients correctly rejected) and compare typical interval widths across methods.

2.2 Empirical Example

We carry out a common empirical macroeconomic exercise: we estimate the dynamic effects of an externally identified monetary policy shock on output, prices, and financial conditions. The application illustrates how the choice between pointwise and simultaneous inference can determine whether the data appear to support or reject standard transmission channels, even in a modestly sized response family. Our shock measure is the Bu–Rogers–Wu (BRW) series of Bu et al. (2021). We follow Bu et al. (2021) and interpret the impulse as a contractionary monetary policy shock, normalized to a

100bp increase in the 2-year Treasury yield on impact.

Data. The response variables of interest are log Industrial Production (IP), log CPI, and the excess bond premium (EBP) (Gilchrist and Zakrajšek, 2012). Following Bu et al. (2021), we also treat the *cumulative* BRW series as an outcome of interest, which serves as a diagnostic for the persistence of the identified policy stance after the shock.[3]

For illustrative purposes, we restrict the sample to 2008m1–2021m12 (post–global financial crisis). This period is particularly relevant for BRW because the series is constructed to bridge the effective-lower-bound era.[4]

VAR and LP specification. We estimate a monthly reduced-form VAR with $p = 12$ lags on the four-dimensional vector $(\text{cumulated BRW}_t, \log \text{IP}_t, \log \text{CPI}_t, \text{EBP}_t)'$, including an intercept. Following Bu et al. (2021), we order the cumulated BRW series first and identify the monetary policy shock as the one-month innovation to this first equation, so the shock is allowed to affect all other variables contemporaneously. We report both 90% pointwise intervals and 0.1-FWER sup-t bands, using the bootstrap-based studentized procedure of Montiel Olea and Plagborg-Møller (2019) applied over the full response family (all outcomes and horizons).

We also estimate IRFs using LPs. For each horizon $h = 0, \ldots, 24$ and each outcome, we regress the h-step-ahead value on the contemporaneous BRW policy shock and $p = 13$ lags of the full state vector, including an intercept, and we treat the coefficient on the shock as the LP estimate of the IRF at horizon h.[5]

Multiplicity. Figure 1 reports responses over horizons $h = 0, \ldots, 24$ months using both a VAR-based estimator and local projections. Even with only four variables, each panel contains $m = 4 \times 25 = 100$ scalar IRF coefficients. This is already large enough that pointwise inference can be misleading. To calibrate magnitudes, under independence the probability of at least one false rejection at a 10% pointwise level is $1-(1-0.10)^{100} \approx 0.99997$, even if *all* coefficients were truly zero. In practice, IRF estimates are dependent across horizons, so the independence calculation may overstate the error rate; nevertheless, it

[3] Unlike Bu et al. (2021), we do not include a commodity price index anywhere in our empirical design.

[4] Bu et al. (2021) show that this sample produces a large set of nonzero pointwise 90% confidence intervals (rejections) for the estimated IRFs, using both VAR- and LP-based estimation approaches.

[5] The lag length follows the lag augmentation recommended by Montiel Olea and Plagborg-Møller (2021). To make the pointwise and simultaneous comparisons comparable across estimators, we construct LP pointwise intervals and sup-t bands using the VAR-based bootstrap developed by Montiel Olea and Plagborg-Møller (2019, 2021), which preserves the joint dependence across horizons and outcomes while delivering studentized critical values for both pointwise and simultaneous inference.

illustrates why pointwise bands are not a reliable basis for joint statements about the response path once the implicit hypothesis family is moderately large.

Pointwise vs simultaneous inference. The pointwise 90% confidence intervals in Figure 1 suggest many nonzero effects. Interpreted literally, the contractionary shock has a persistent effect on the policy stance (the cumulative BRW series), a delayed negative effect on output around one to two years after impact (most visible in the VAR specification), and a persistently negative effect on prices over much of the two-year horizon. This qualitative narrative lines up with the conventional transmission emphasized by Bu et al. (2021) in closely related VAR/LP designs.

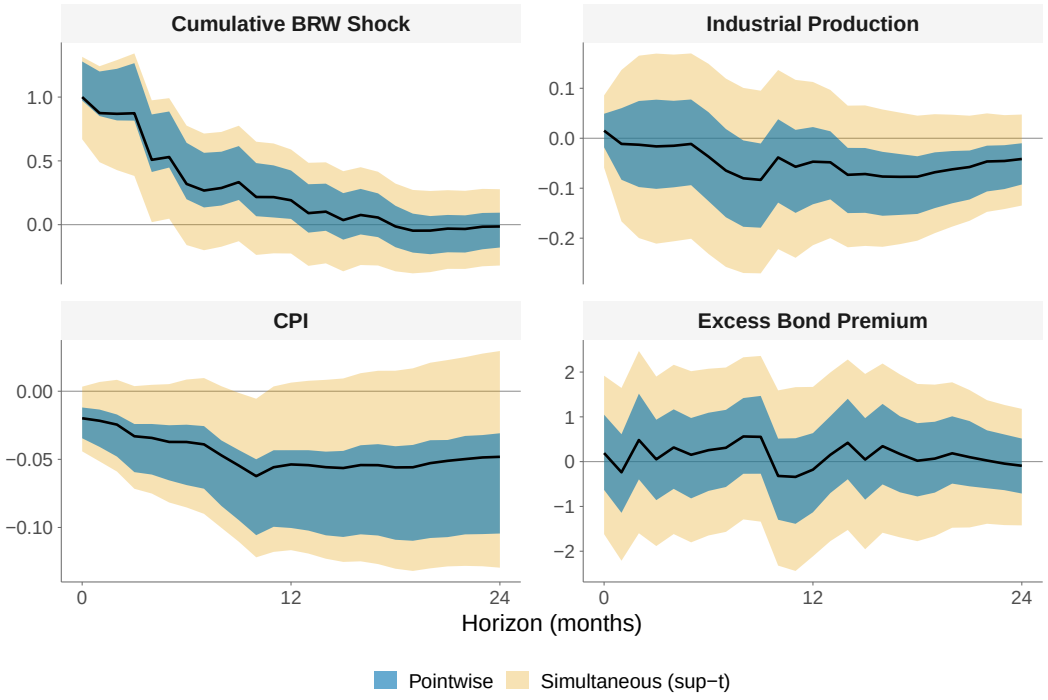
However, when we replace pointwise bands with simultaneous sup-t bands that control the FWER at 10% over the full response family, nearly all of these "discoveries" disappear. The only robust conclusion that remains is that the shock generates a short-lived movement in the policy stance itself. In other words, once we require joint error control across the full grid of horizons and outcomes, the simultaneous confidence bands become almost completely uninformative on the effects of the shock on output, inflation, and financial conditions. This result is uniform across both the VAR and LP estimation approaches.

Takeaway. This empirical example illustrates why simultaneous inference that targets "no false positives anywhere" can be too conservative for IRF work once the response family is moderately high-dimensional, even in settings where the underlying economic signal is plausibly present. At the same time, the richness of the pointwise plot reflects the reality that applied IRF analysis rarely commits ex ante to a single horizon or outcome. To understand the forces driving this tradeoff, we now examine a controlled simulation environment where the true impulse responses are known, so that false discovery rates, coverage, and power can be measured exactly.

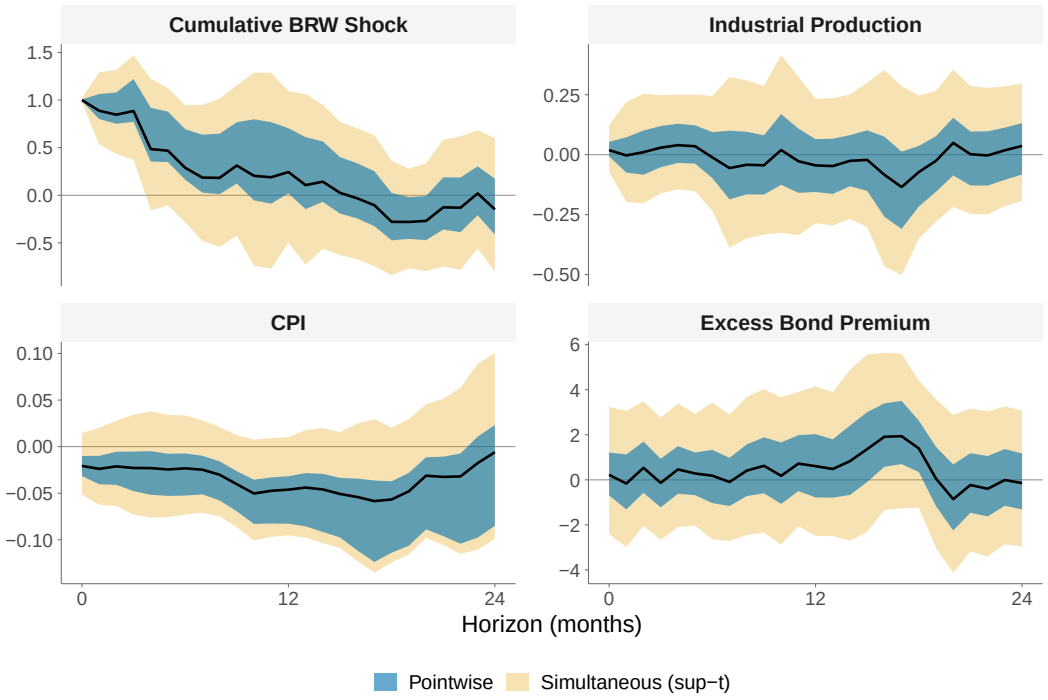
2.3 Simulation Evidence

Illustrative DGP. Consider the following six-variable VAR(4) data-generating process (DGP):

$$y_t = \sum_{\ell=1}^{4} A_\ell y_{t-\ell} + u_t, \qquad u_t = C\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I_6). \tag{11}$$



(a) VAR



(b) LP

Figure 1: Responses to a contractionary monetary policy shock.

Partition the endogenous vector into two three-variable blocks, $y_t = (y_{1t}', y_{2t}')'$. For each lag $\ell = 1, \ldots, 4$, the DGP is block-diagonal both in the autoregressive coefficients and in the contemporaneous impact matrix:

$$A_\ell = \begin{bmatrix} A_{\ell,1} & 0 \\ 0 & A_{\ell,2} \end{bmatrix}, \qquad C = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix}. \tag{12}$$

This structure eliminates any contemporaneous or dynamic transmission across blocks. As a result, all cross-block impulse responses are identically zero by construction, corresponding to null effects. See Appendix B, Figures B.1 and B.2, for the true IRF grids used in the simulation designs.

Objective. Consider a researcher who has correctly identified the structural shocks, in the sense that the contemporaneous impact matrix C is known, but who does not know the full DGP.[6] Their goal is to summarize dynamic effects by estimating a collection of structural impulse responses over horizons $h \in \{0, 1, \ldots, H\}$ for a subset of the structural shocks $\mathcal{S} \subseteq \{1, \ldots, 3\}$. They estimate an unrestricted six-dimensional VAR(4) and compute IRFs for each variable–shock pair and each horizon. The key inferential choice is between (i) *pointwise* procedures that guarantee coverage/testing validity coefficient-by-coefficient and (ii) *simultaneous* procedures that guarantee joint validity over a pre-specified family of coefficients.

Hypothesis family. Let the reduced-form moving-average coefficients implied by the VAR be $\{\Phi_h\}_{h\geq 0}$ with $\Phi_0 = I_6$ and $\Phi_h = \sum_{\ell=1}^{4} A_\ell \Phi_{h-\ell}$ for $h \geq 1$. The structural impulse response matrices are

$$\Psi_h = \Phi_h C, \qquad h = 0, 1, \ldots, H, \tag{13}$$

and the scalar IRF coefficient for variable $k \in \{1, \ldots, 6\}$, shock $s \in \mathcal{S}$, and horizon h is

$$\theta_{k,s}(h) \equiv e_k^\top \Psi_h e_s. \tag{14}$$

Collect the full set of reported coefficients into an index set

$$\mathcal{J} \equiv \{(k, s, h) : k \in \{1, \ldots, 6\},\ s \in \mathcal{S},\ h \in \{0, \ldots, H\}\}, \qquad m \equiv |\mathcal{J}| = 6 \cdot |\mathcal{S}| \cdot (H + 1). \tag{15}$$

[6] The purpose of this setup is to abstract away from identification issues, and instead focus purely on inference.

Write θ_j for a generic element of this family (via any fixed one-to-one mapping $j \leftrightarrow (k, s, h)$), and let $\widehat{\theta}_j$ denote its estimator with standard error $\widehat{s}_j$.

The block-diagonal structure in Eq. (11) implies a large subset of *exact* null coefficients: all cross-block responses satisfy $\theta_{k,s}(h) = 0$ for every h whenever k is in one block and s is in the other. Let

$$\mathcal{H}_0 \equiv \{j \in \{1, \ldots, m\} : \theta_j = 0\}, \qquad \mathcal{H}_1 \equiv \{1, \ldots, m\} \setminus \mathcal{H}_0, \tag{16}$$

so that $\mathcal{H}_0$ is sizeable in this design by construction. This makes the DGP a clean environment to study false rejections and miscoverage induced purely by multiplicity.

Simulation design. We evaluate pointwise and sup-t inference in the DGP specified in Eqs. (11)–(12) by repeatedly simulating samples of size T, estimating the unrestricted VAR(4), and constructing (i) pointwise confidence intervals CI_j^{pt} and (ii) sup-t simultaneous bands CI_j^{sim} using bootstrap-based studentization. In the baseline implementation, pointwise intervals use a bias-corrected (Pope, 1990) residual bootstrap, while simultaneous bands use the Montiel Olea and Plagborg-Møller (2019) sup-t methodology; Appendix A provides full algorithmic details. We summarize performance using the FDR, FCR, power, and interval width measures defined in Section 2.1.

Insights. Figure 2 makes the core size–power tension transparent in a setting with many exact zeros: pointwise intervals deliver narrow bands and therefore many apparent "discoveries," but the false discovery rate and post-selection miscoverage rise quickly when inference is conducted over a large IRF grid.[7] By contrast, sup-t simultaneous bands control errors uniformly over the entire family by calibrating to an extreme-value statistic, which sharply reduces false positives but widens bands and depresses power.

The practical implication is that the inferential choice determines what kinds of empirical claims can be made responsibly from IRF plots. For instance, if the objective is to learn which responses are likely to be nonzero and economically meaningful while keeping plots informative, strict FWER control can be too blunt: it trades away detection of true dynamics to insure against *any* false rejection, even when a small number of false positives would not alter the economic conclusion. In applied work this maps directly into mechanism and policy interpretation risk: pointwise procedures can overstate evidence for channels that are in fact absent, while overly conservative simultaneous bands can understate evidence for true dynamic effects. The next section develops an inferential

[7] Figure B.3 in Appendix B reports the corresponding results for the LP estimator; the patterns are qualitatively similar.

framework that bounds the expected share of false discoveries among the selected findings, and shows that this intermediate guarantee resolves the tradeoff in both the empirical and simulation settings examined here.

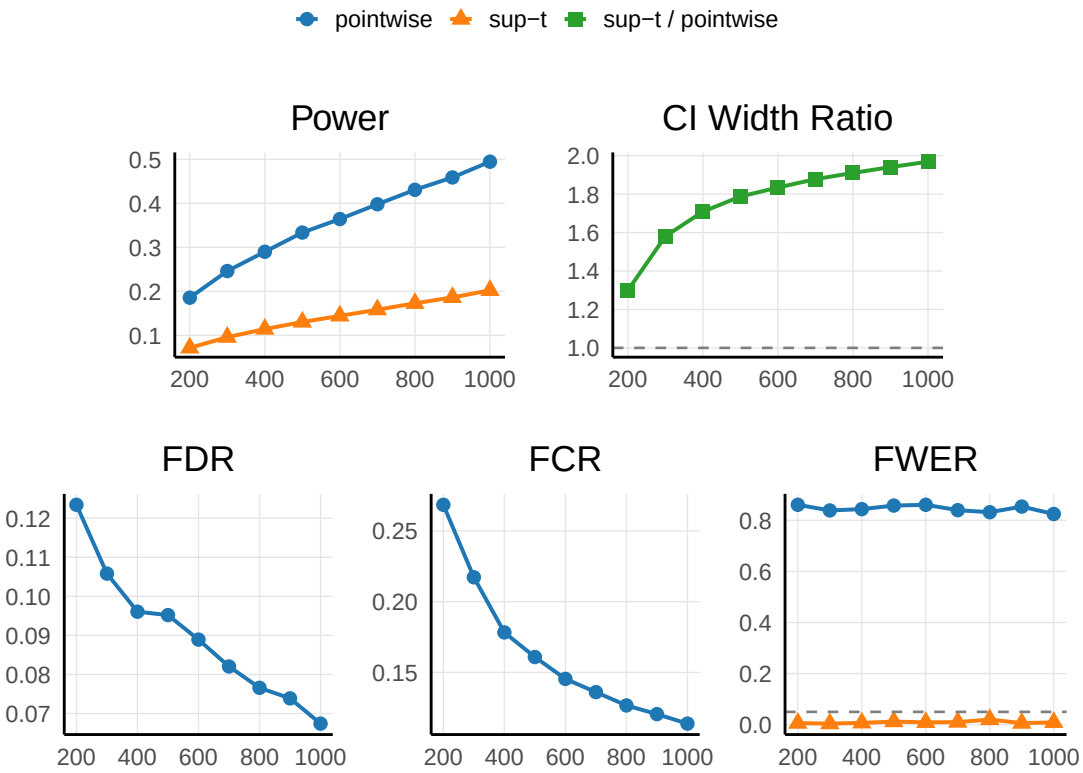


Figure 2: Monte Carlo performance (bootstrapped VAR IRFs, baseline DGP, $H = 20$): 95% pointwise bands versus sup-t bands controlling FWER at 5%. Panels report power, average confidence-interval width inflation (sup-t / pointwise), and the realized FDR, FCR, and FWER.

3 FDR- and FCR-Adjusted IRF Inference

This section presents the FDR- and FCR-adjusted inference procedure. The main idea is a two-step reporting rule: first, select the subset of IRF coefficients that can be declared nonzero while controlling the expected share of false discoveries among those declarations at a user-chosen level q; second, report confidence intervals only for the selected coefficients, constructed so that the expected fraction of noncovering reported intervals is also bounded by q.

Section 3.1 presents the methodology. We describe the general selection-and-inference pipeline, detail its implementation for VAR and LP estimators using the same joint bootstrap machinery already used for pointwise and sup-t bands, and discuss the choice of selection rule. Our baseline is the RSW stepdown procedure (Romano et al., 2008), which leverages the joint bootstrap distribution to deliver FDR control under the general dependence structures that arise in IRF applications; we also present the Benjamini–Hochberg (Benjamini and Hochberg, 1995) and Benjamini–Yekutieli (Benjamini and Yekutieli, 2001) procedures as alternatives. Section 3.2 revisits the monetary policy application to show how FDR/FCR-adjusted bands preserve economically interpretable dynamics in a setting where sup-t bands are largely uninformative, and Section 3.3 uses Monte Carlo evidence to quantify the power, width, and error-rate tradeoffs.

3.1 Methodology

We adopt the notation and error rate definitions introduced in Section 2.1. The family of m IRF coefficients $\theta_1, \ldots, \theta_m$ is tested against the null $H_j : \theta_j = 0$ for $j = 1, \ldots, m$, with performance measured by the FDR (expected share of false rejections among all rejections) and the FCR (expected share of noncovering intervals among those reported). The goal is to choose a procedure that controls both FDR and FCR at a user-specified level q.[8]

Empirical objects and test statistics. The procedure requires a point estimate $\hat{\theta}_j$ and standard error $\hat{\sigma}_j$ for each coefficient. Because IRF estimates are dependent across horizons, variables, and shocks, the procedure cannot treat coefficients in isolation. Instead, it uses a joint bootstrap that produces B draws of the full IRF vector, $\hat{\theta}^{*(b)} = (\hat{\theta}_1^{*(b)}, \ldots, \hat{\theta}_m^{*(b)})'$ for $b = 1, \ldots, B$. The bootstrap may also deliver studentized statistics directly, $t^{*(b)} = (t_1^{*(b)}, \ldots, t_m^{*(b)})'$, which is convenient for local projections with percentile-t inference. Drawing the full vector jointly in each bootstrap replication is essential: it preserves the cross-coefficient dependence structure so that the subsequent selection and inference steps can exploit the fact that many of the m tests are near-redundant rather than treating them as independent sources of multiplicity.

To put all coefficients on a comparable scale, each is summarized by a two-sided studentized statistic $T_j = |\hat{\theta}_j / \hat{\sigma}_j|$. When the bootstrap delivers draws on the coefficient scale, the corresponding bootstrap statistic is $T_j^{*(b)} = |(\hat{\theta}_j^{*(b)} - \hat{\theta}_j)/\hat{\sigma}_j|$; when it delivers

[8] Let m_0 denote the number of true nulls in the IRF array, i.e., $\theta_j = 0$. In practice, methods implementing FDR control satisfy the asymptotic behavior FDR $\leq \frac{m_0}{m} q \leq q$, including the case where $m_0 = m$. Adaptive procedures exist to tighten these bounds and estimate m_0/m (Storey, 2002), but these lie outside the initial scope of this work.

studentized draws directly, $T_j^{*(b)} = |t_j^{*(b)}|$. A bootstrap p-value for each hypothesis is

$$\hat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left\{T_j^{*(b)} \geq T_j\right\}}{B+1}. \tag{17}$$

General FDR/FCR procedure. The selection-and-inference pipeline can be described independently of the particular IRF estimator. The core logic has two parts: first, select which coefficients to report using an FDR-controlling rule; second, widen the confidence intervals for the selected set to account for the fact that these coefficients were chosen precisely because they appeared significant. Concretely, a multiple-testing decision rule that controls FDR at level q is applied to the family of m hypotheses to obtain a rejection set $\widehat{\mathcal{R}}$, and then confidence intervals are reported only for the selected coefficients, with coverage calibrated to the size of the rejection set. The steps are as follows:

1. Estimate the full IRF vector and compute a standard error for each coefficient.
2. Generate bootstrap draws of the full IRF vector and compute the corresponding studentized statistics.
3. Apply the chosen FDR-controlling rule to the vector of statistics or p-values to obtain the rejection set $\widehat{\mathcal{R}}$.
4. Let $\widehat{R} = |\widehat{\mathcal{R}}|$ and set $\alpha^* = q\,\widehat{R}/m$.
5. For each rejected coefficient, compute a marginal confidence interval with miscoverage probability α^* using the same bootstrap distribution.
6. Report only the intervals corresponding to rejected coefficients, and treat all nonrejected coefficients as not selected.

The remainder of this subsection describes the two key components in detail: the FDR-controlling selection rule (step 3) and the FCR-adjusted confidence intervals (steps 4–5). We then discuss how both components are implemented for VAR and LP estimators and present alternative selection rules.

RSW FDR control. The RSW method (Romano et al., 2008) is a bootstrap stepdown procedure designed to control FDR under general dependence. Its key advantage in the IRF context is that it calibrates critical values using the joint bootstrap distribution of the full vector of test statistics, rather than reducing each hypothesis to a marginal p-value

and discarding information about how test statistics co-move. In IRF applications, this co-movement is strong: test statistics at adjacent horizons are highly correlated because they are nonlinear functions of the same estimated parameters, and the effective number of independent tests is typically much smaller than m. By working with the joint bootstrap distribution, RSW exploits this redundancy automatically, producing less conservative critical values than procedures that treat each test in isolation or that impose worst-case dependence corrections.

The procedure begins by ordering hypotheses from least to most significant, $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)}$, with π denoting the permutation that maps the order-statistic index to the original hypothesis index, so that $T_{(j)} = T_{\pi(j)}$. The bootstrap statistics are reordered using the same permutation, $T^{*(b)}_{(j)} = T^{*(b)}_{\pi(j)}$, to preserve the dependence structure in the relevant ordering. For any bootstrap draw, let $T^{*(b)}_{(r):t}$ denote the r-th smallest element of $\{T^{*(b)}_{(1)}, \ldots, T^{*(b)}_{(t)}\}$, and let $\widehat{P}^*(\cdot)$ denote the empirical bootstrap probability operator,

$$\widehat{P}^*(A) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{A \text{ holds in draw } b\}. \tag{18}$$

The procedure then computes a sequence of critical values $\hat{c}_1, \ldots, \hat{c}_m$ by recursion, working from the least significant hypothesis ($j = 1$) upward. At each step j, the critical value $\hat{c}_j$ is chosen as the smallest threshold such that the expected false discovery proportion among the j most significant hypotheses, evaluated under the bootstrap distribution, does not exceed q. Formally:

$$\hat{c}_j = \inf \left\{ c \in \mathbb{R} : \sum_{r=m-j+1}^{m} \frac{r-m+j}{r} \widehat{P}^* \left(\{T^*_{(j):j} \geq c\} \cap \bigcap_{\ell=m-r+1}^{j-1} \{T^*_{(\ell):j} \geq \hat{c}_\ell\} \cap \{T^*_{(m-r):j} < \hat{c}_{m-r}\} \right) \leq q \right\}. \tag{19}$$

The final set in the intersection is interpreted as automatically satisfied when its subscript would be zero. The event inside the bootstrap probability corresponds to a particular configuration of how many hypotheses among the first j would be rejected by the stepdown rule in the bootstrap draw. The stepdown aspect is what gives the procedure its power advantage over single-step methods: once a hypothesis is "cleared" (not rejected) at an earlier step, the procedure conditions on that outcome and recalibrates the remaining critical values, progressively sharpening the thresholds for the more significant hypotheses. The critical values are computed sequentially from $j = 1$ up to $j = m$ and are enforced to be weakly increasing to ensure a coherent stepdown decision rule.

Given the ordered statistics and critical values, the rejection set is a tail set in the ordering:

$$\hat{j}_0 = \min\left\{j \in \{1, \ldots, m+1\} : T_{(r)} \geq \hat{c}_r \text{ for all } r = j, \ldots, m\right\}, \tag{20}$$

$$\widehat{\mathcal{R}} = \{\pi(r) : r = \hat{j}_0, \ldots, m\}. \tag{21}$$

All hypotheses indexed by $\widehat{\mathcal{R}}$ are rejected; the remainder are not. By construction, the procedure targets the guarantee that the expected fraction of false rejections among all rejections is at most q.

FCR-adjusted confidence intervals. After selection, confidence intervals are reported only for coefficients in $\widehat{\mathcal{R}}$. The FCR adjustment replaces the fixed marginal miscoverage level with a data-dependent level,

$$\alpha^* = q\,\frac{\widehat{R}}{m}. \tag{22}$$

To see the logic, contrast this with two benchmarks. A pointwise interval uses level α^{pt} for every coefficient, ignoring that m intervals are examined. A Bonferroni correction would set $\alpha^{\text{Bonf}} = q/m$, controlling the probability that any reported interval fails to cover. The FCR adjustment lies between these extremes because it targets a proportion, not an event: among the $\widehat{R}$ intervals actually reported, we require only that the expected share of noncovering intervals is at most q, not that every single one covers. This means the error budget grows with the size of the reported set. When many coefficients survive selection ($\widehat{R}$ large relative to m), the procedure can tolerate a larger absolute number of noncovering intervals while still keeping their share below q, so each interval receives a more generous miscoverage allowance and can therefore be narrower. When few coefficients survive ($\widehat{R}$ small), fewer intervals share the budget and each must be wider. Formally, $\alpha^* = q\,\widehat{R}/m$ can be motivated as follows: if the FDR selection step delivers roughly $\widehat{R}$ discoveries of which at most a fraction q are expected to be false, then allocating each of the $\widehat{R}$ reported intervals a miscoverage level of $q\,\widehat{R}/m$ keeps the expected fraction of noncovering reported intervals at the same level q.[9] This scaling property is the coverage-side analog of the principle that makes FDR control non-conservative as m grows: in both cases the guarantee attaches to

[9] As a concrete example, suppose that in a six-variable VAR over fifty horizons the FDR selection retains $\widehat{R} = 20$ of $m = 300$ coefficients. The FCR-adjusted level is $\alpha^* = 0.05 \times 20/300 \approx 0.003$, which lies between the pointwise level (0.05) and the Bonferroni level ($0.05/300 \approx 0.0002$). The resulting intervals are wider than pointwise but substantially narrower than Bonferroni, and the width adjusts automatically to $\widehat{R}$ rather than being fixed at the most conservative case.

the selected set, not the full family.

In the IRF plots that follow, this means that horizons surviving a stringent selection (few red bars) carry wider uncertainty bands than horizons surviving a lenient selection (many red bars).

When bootstrap draws on the coefficient scale are available, let $\hat{q}_{j,\tau}$ denote the τ quantile of the centered bootstrap distribution $\{\hat{\theta}_j^{*(b)} - \hat{\theta}_j\}_{b=1}^{B}$. The FCR-adjusted basic bootstrap interval is

$$\widehat{C}_j^{\mathrm{FCR}} = \left[\hat{\theta}_j - \hat{q}_{j,1-\alpha^*/2},\ \hat{\theta}_j - \hat{q}_{j,\alpha^*/2}\right]. \tag{23}$$

When bootstrap t-statistics are available, let $\hat{t}_{j,\tau}$ denote the τ quantile of $\{t_j^{*(b)}\}_{b=1}^{B}$. The FCR-adjusted percentile-t interval is

$$\widehat{C}_j^{\mathrm{FCR}} = \left[\hat{\theta}_j - \hat{\sigma}_j\hat{t}_{j,1-\alpha^*/2},\ \hat{\theta}_j - \hat{\sigma}_j\hat{t}_{j,\alpha^*/2}\right]. \tag{24}$$

The reported interval for coefficient j is $\widehat{C}_j^{\mathrm{FCR}}$ if $j \in \widehat{\mathcal{R}}$ and is not reported otherwise. Under the usual conditions for FCR adjustment, this construction controls the expected proportion of noncovering reported intervals at level q.

VAR implementation. The general procedure above is agnostic to the IRF estimator, but the bootstrap construction differs between VARs and local projections. This distinction determines which version of the bootstrap test statistic ($T_j^{*(b)}$) and which version of the FCR-adjusted interval (basic bootstrap or percentile-t) is used.

For the VAR, consider a stable reduced-form VAR(p) for the K-dimensional vector y_t,

$$y_t = c + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \tag{25}$$

with structural shocks introduced via an impact matrix C satisfying $u_t = C\varepsilon_t$ and $\mathbb{E}[\varepsilon_t\varepsilon_t'] = I_K$. The reduced-form moving-average representation $y_t = \sum_{h=0}^{\infty}\Phi_h u_{t-h}$ yields structural impulse response matrices $\Theta(h) = \Phi_h C$, with scalar coefficients $\theta_{k,s}(h) = e_k'\Theta(h)e_s$.

Estimation proceeds by OLS equation by equation. A residual bootstrap generates bootstrap samples by resampling the estimated reduced-form residuals and simulating a bootstrap path recursively from the fitted VAR,

$$y_t^{*(b)} = \hat{c} + \hat{A}_1 y_{t-1}^{*(b)} + \cdots + \hat{A}_p y_{t-p}^{*(b)} + u_t^{*(b)}. \tag{26}$$

Each bootstrap sample is re-estimated to obtain bootstrap IRF draws for the entire

coefficient vector. In simulation designs where the impact matrix is treated as known, C is held fixed across bootstrap draws so that only sampling variation in the reduced-form dynamics drives uncertainty; in empirical designs where C is estimated from the data, the bootstrap re-estimates the impact matrix within each draw so that identification uncertainty is propagated into the IRF distribution.

The standard error for each coefficient is taken as the standard deviation of the bootstrap IRF draws, $\hat{\sigma}_j = \mathrm{sd}(\{\hat{\theta}_j^{*(b)}\}_{b=1}^{B})$, and the studentized statistics are

$$T_j = \left|\frac{\hat{\theta}_j}{\hat{\sigma}_j}\right|, \qquad T_j^{*(b)} = \left|\frac{\hat{\theta}_j^{*(b)} - \hat{\theta}_j}{\hat{\sigma}_j}\right|. \tag{27}$$

The full vector of statistics is passed to the RSW stepdown algorithm to obtain the rejection set, and the same bootstrap draws are used to form FCR-adjusted basic bootstrap confidence intervals at the α^* level implied by the size of the rejection set. If small-sample bias in estimated VAR dynamics is a concern, the bootstrap data-generating process can be based on a bias-corrected VAR (Pope, 1990) while leaving the reported point estimate unchanged; this alters the bootstrap distribution used for p-values and intervals but does not alter the RSW and FCR logic.

LP implementation. For local projections, each horizon-specific response is estimated by a regression of future outcomes on the shock and lagged controls. A lag-augmented LP takes the generic form

$$y_{t+h} = a_h + \Theta(h)\varepsilon_t + B_{h,1}y_{t-1} + \cdots + B_{h,p+1}y_{t-p-1} + e_{t,h}, \tag{28}$$

where overlapping outcomes imply that the projection errors $e_{t,h}$ are serially correlated for $h > 0$, so robust inference requires either a long-run variance estimator or a bootstrap that reproduces the dependence.

A convenient approach is a VAR-based wild bootstrap in which a fitted VAR serves as an auxiliary model to generate bootstrap paths. The wild bootstrap multiplies the estimated residuals by i.i.d. mean-zero, unit-variance scalars $\xi_t^{*(b)}$ (Gaussian or Rademacher) to preserve conditional heteroskedasticity:

$$u_t^{*(b)} = \xi_t^{*(b)}\hat{u}_t, \qquad y_t^{*(b)} = \hat{c} + \hat{A}_1 y_{t-1}^{*(b)} + \cdots + \hat{A}_p y_{t-p}^{*(b)} + u_t^{*(b)}. \tag{29}$$

To maintain the same shock regressor as in the original sample, the bootstrap reduced-form innovations are mapped into bootstrap structural shocks, $\varepsilon_t^{*(b)} = C^{-1}u_t^{*(b)}$. For

each bootstrap draw, the LP regressions are re-estimated on the bootstrap data to obtain bootstrap IRF coefficients $\hat{\theta}_j^{*(b)}$ and robust standard errors $\hat{\sigma}_j^{*(b)}$. A bootstrap-t statistic is formed by centering at the pseudo-true value implied by the auxiliary VAR and dividing by the bootstrap standard error,

$$t_j^{*(b)} = \frac{\hat{\theta}_j^{*(b)} - \theta_j^{\text{aux}}}{\hat{\sigma}_j^{*(b)}}, \tag{30}$$

where θ_j^{aux} is the impulse response implied by the fitted VAR, which serves as the population LP coefficient under the auxiliary model. The observed test statistic and bootstrap counterpart are $T_j = |\hat{\theta}_j/\hat{\sigma}_j|$ and $T_j^{*(b)} = |t_j^{*(b)}|$. The RSW stepdown algorithm is applied to the full vector of studentized statistics so that selection is calibrated to the dependence structure across horizons and variables, and FCR-adjusted intervals are obtained by reusing the percentile-t inversion with α^* in place of the pointwise α level.

Alternative selection rules. The RSW method can be replaced by classical step-up procedures that operate only on the vector of marginal p-values. The Benjamini–Hochberg (BH) procedure orders p-values from smallest to largest and compares them to a linear critical line:

$$\hat{p}_{(1)} \le \hat{p}_{(2)} \le \cdots \le \hat{p}_{(m)}, \tag{31}$$

$$\hat{k}_{\text{BH}} = \max\left\{k \in \{0, 1, \ldots, m\} : \hat{p}_{(k)} \le \frac{k}{m} q\right\}, \tag{32}$$

$$\widehat{\mathcal{R}}_{\text{BH}} = \{j : \hat{p}_j \le \hat{p}_{(\hat{k}_{\text{BH}})}\}. \tag{33}$$

Under independence or positive regression dependence on the subset of true nulls (PRDS), BH controls FDR at level q (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). The Benjamini–Yekutieli (BY) procedure modifies the BH critical line by a harmonic factor $c_m = \sum_{\ell=1}^{m} 1/\ell$ to obtain a dependence-robust bound:

$$\hat{k}_{\text{BY}} = \max\left\{k \in \{0, 1, \ldots, m\} : \hat{p}_{(k)} \le \frac{k}{m\, c_m} q\right\}, \tag{34}$$

$$\widehat{\mathcal{R}}_{\text{BY}} = \{j : \hat{p}_j \le \hat{p}_{(\hat{k}_{\text{BY}})}\}. \tag{35}$$

BY controls FDR at level q under arbitrary dependence but is typically much more conservative than both BH and RSW, especially when m is large. Once a rejection set is obtained by BH or BY, the same FCR-adjustment step applies without modification: the

number of selected coefficients determines α^*, and the reported intervals are constructed at the α^* level using the same bootstrap machinery.

Choice of selection rule. The theoretical case for RSW over BH in the IRF setting rests on the dependence structure of IRF test statistics. BH's FDR guarantee requires either independence or PRDS across the null test statistics, but neither condition is generically satisfied in IRF applications. IRF t-statistics are dependent across horizons, variables, and shocks because all coefficients are functions of common estimated objects, and this dependence can involve both positive and negative correlations. To see why concretely, consider a bivariate VAR(1) with coefficient matrix A and impact vector $b = (b_1, b_2)'$ for a single identified shock. The horizon-1 IRF for variable 1 is $\rho(b_1 + b_2)$ and for variable 2 is $\rho(-b_1 + b_2)$, where the two responses load on the same estimated object $\hat{b}_1$ with opposite signs. Under the null $b_1 = b_2 = 0$, the studentized t-statistics for these two responses are negatively correlated: a positive estimation error in $\hat{b}_1$ inflates one statistic while deflating the other. PRDS is a global monotonicity restriction on conditional distributions that is not compatible with such mixed-sign dependence. Appendix C formalizes this intuition and constructs explicit counterexamples showing that PRDS fails for simple VAR and LP configurations. BY resolves this by guaranteeing FDR control under arbitrary dependence, but the harmonic penalty c_m can substantially reduce the number of discoveries relative to BH or RSW. RSW avoids both problems by calibrating critical values directly from the joint bootstrap distribution of the full test statistic vector, incorporating the actual dependence structure without requiring PRDS or paying a worst-case penalty.

That said, our Monte Carlo evidence suggests that BH performs well in practice across a range of empirically relevant IRF designs. Across both DGPs, both estimators, and multiple horizon ranges, BH delivers FDR control at or near the target level and achieves power comparable to RSW, while BY is noticeably more conservative throughout (see Appendix B, Figures B.13–B.14 and accompanying figures for other configurations). This is consistent with a body of evidence in the statistics literature suggesting that BH tends to control FDR even under moderate departures from its sufficient conditions. For applied researchers who prefer a simpler implementation, BH+FCR is therefore a reasonable practical alternative. However, the fact that BH performs well in the configurations we examine does not constitute a theoretical guarantee that it will do so in others, and the dependence patterns that violate PRDS are not pathological edge cases but generic features of how IRF statistics are constructed. The situation is analogous to other settings in econometrics where a procedure that is valid under stronger assumptions happens to perform adequately in simulations, but the principled default is the one whose formal

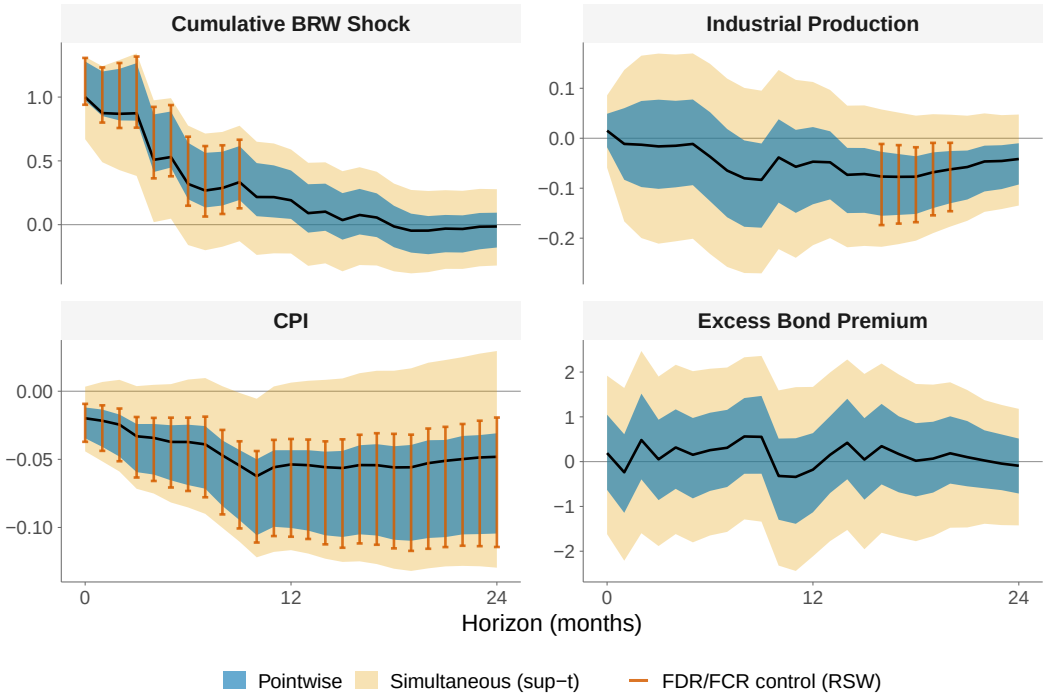
guarantee matches the data structure at hand. We adopt RSW as our baseline for this reason: it is the only procedure among the three with a formal FDR guarantee under the dependence structures that arise generically in IRF applications. The cost of this choice is computational rather than statistical, and it is modest in practice, because the joint bootstrap required by RSW is the same machinery already needed for sup-t bands in standard VAR and LP workflows.

3.2 Empirical Example

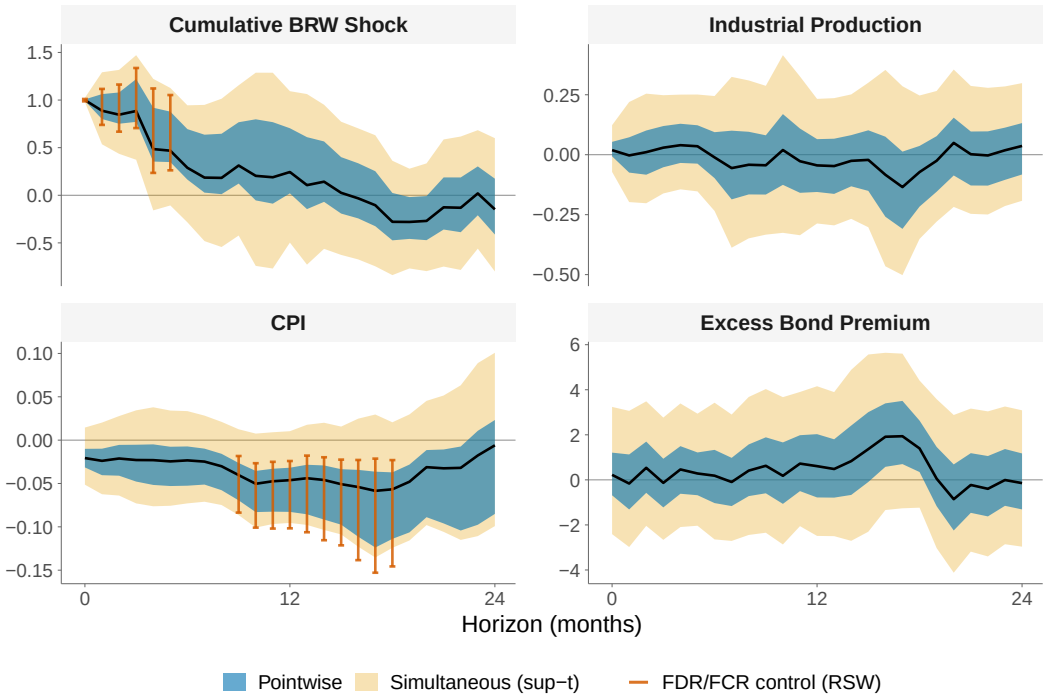
We return to the monetary policy application from Section 2.2 and use it to illustrate how FDR/FCR-adjusted reporting sits between naive pointwise inference and fully simultaneous (FWER-controlled) inference. As before, the reported response family consists of the four outcomes (cumulative BRW, IP, CPI, and the excess bond premium) over horizons $h = 0, \ldots, 24$, so $m = 4 \times 25 = 100$ horizon-by-outcome coefficients are implicitly scanned when interpreting the IRF plot. Figure 3 overlays three inferential objects for this same response family: conventional 90% pointwise confidence bands (blue shading), 90% simultaneous sup-t bands controlling the FWER at 0.1 over the entire grid (tan shading), and our dependence-robust RSW procedure controlling the FDR at $q = 0.1$ with post-selection intervals that control the FCR at the same q (red intervals).

The key visual feature of the FDR/FCR approach is that it reports uncertainty only for coefficients that survive FDR-controlled selection. Concretely, the RSW stepdown algorithm is applied to the full vector of studentized IRF statistics to determine a rejection (selection) set. For the selected coefficients, we then report marginal confidence intervals computed at the FCR-adjusted level $\alpha^\star = q\,\widehat{R}/m$, where $\widehat{R}$ is the realized number of selections. These FCR-adjusted intervals are plotted as red bands (error-bar style) around the corresponding IRF point estimates; horizons/outcomes that are not selected simply have no red band, indicating that they are not reported as nonzero after multiplicity control.

Figure 3 shows that this intermediate error criterion preserves much of the qualitative narrative suggested by pointwise bands, while avoiding the near-complete loss of informativeness induced by uniform FWER control. First, in both the VAR and LP specifications, the cumulative BRW response indicates that the contractionary policy stance persists for several months after impact before decaying back toward zero; the FDR/FCR selections concentrate on these early horizons, consistent with a short- to medium-lived policy tightening. Second, the CPI response displays a persistent disinflationary effect: in the VAR, the selected CPI coefficients are negative across a large



(a) VAR



(b) LP

Figure 3: Responses to a contractionary monetary policy shock.

portion of the horizon range, and in the LP the selected negatives concentrate in the middle horizons, but in both cases the post-selection red intervals support a sustained decline in prices relative to the baseline. Third, the output response differs across estimators in a way that is transparent under FDR/FCR control: the VAR results show a delayed contraction in industrial production (selected negative coefficients roughly around the one-year horizon, fading again before two years), whereas the LP results do not yield selected nonzero IP responses over the same horizon range. Finally, neither estimator produces FDR-selected effects for the excess bond premium, so the method does not support a robust financial-spread response once the horizon-by-outcome multiplicity is accounted for.

Overall, the figure illustrates the practical role of FDR/FCR control for IRF reporting: it delivers an explicit guarantee that, among the horizons/outcomes highlighted as nonzero, the expected share of false discoveries is bounded by 0.1, and the expected share of noncovering reported intervals is also bounded by 0.1, while still allowing economically interpretable dynamics to emerge in settings where simultaneous sup-t bands are too conservative to be informative.

3.3 FDR/FCR Control vs. Pointwise and Simultaneous Inference

This subsection benchmarks the proposed RSW-based FDR/FCR pipeline against the two procedures that dominate applied IRF reporting: naive pointwise inference and FWER-controlled simultaneous (sup-t) inference. The comparison uses the block-diagonal simulation design introduced in Section 2, with full algorithmic details in Appendix A: a six-variable VAR(4) with three reported shocks, $H = 19$ horizons, and $m = 360$ IRF coefficients per replication, of which a large share are exact zeros by construction. In each Monte Carlo replication we estimate the full IRF array and apply: (i) pointwise 95% confidence intervals (equivalently, per-coefficient tests at $\alpha^{\text{pt}} = 0.05$), (ii) a sup-$t$ band controlling the family-wise error rate at $\alpha^{\text{FWER}} = 0.05$ over the full IRF family, and (iii) the dependence-robust RSW stepdown procedure controlling the FDR at $q = 0.05$, followed by post-selection intervals constructed at the FCR-adjusted level $\alpha^{\star} = q\,\widehat{R}/m$. We summarize performance using average power over true nonzero coefficients, the realized FDR and FCR, and the average reported interval width (shown as a ratio relative to pointwise width).

Figure 4 reports these metrics as a function of sample size for the baseline DGP with the VAR estimator. Appendix B reports the complete set of results across both DGPs (baseline and persistent), both estimators (VAR and LP), and all five hypothesis-family specifications

described in Appendix A: the baseline family ($H = 20$, Figures B.3–B.4), fewer horizons ($H = 10$, Figures B.5–B.6), more horizons ($H = 40$, Figures B.7–B.8), lag misspecification (Figures B.9–B.10), and a sparser true-null configuration (Figures B.11–B.12). The patterns described below for the baseline case hold uniformly across all of these configurations.

The top-left panel shows the expected power ordering. Pointwise inference is the most powerful, but this power comes from treating each horizon-by-variable coefficient in isolation. Simultaneous sup-t inference is the least powerful because it calibrates to an extreme-value statistic over the entire IRF family, producing wide bands and few rejections. The RSW procedure sits cleanly between these extremes: it recovers a substantial fraction of the pointwise discoveries while still imposing an explicit multiplicity constraint. In other words, replacing the "no false positives anywhere" criterion of FWER control with the "few false positives among those flagged" criterion of FDR control delivers a material power gain in IRF grids of empirically relevant dimension. The power ordering of pointwise above RSW above sup-t is maintained without exception across both DGPs, both estimators, all horizon ranges, and the lag-misspecification and sparsity robustness checks in Appendix B.

The top-left panel shows the expected power ordering: pointwise inference is most powerful but unregulated, sup-t inference is least powerful, and RSW sits cleanly between the two, recovering a substantial fraction of pointwise discoveries while imposing an explicit multiplicity constraint. This ordering is maintained without exception across both DGPs, both estimators, all horizon ranges, and the lag-misspecification and sparsity robustness checks in Appendix B.

The top-right panel shows the mechanism behind this power gain: interval width. The RSW/FCR-adjusted intervals are only modestly wider than pointwise intervals and remain markedly tighter than sup-t bands. The width advantage of RSW over sup-t is amplified as the hypothesis family grows: the sup-t/pointwise width ratio ranges from roughly 1.3–1.6 at $H = 10$ ($m = 180$) to approximately 2.0–2.2 at $H = 40$ ($m = 720$), because the extreme-value critical value must account for a larger maximum, while the RSW/pointwise ratio remains in the range 1.0–1.3 throughout (Figures B.5–B.7).

The bottom panels confirm that these gains do not come at the expense of reliability. RSW delivers FDR and FCR control at or below the target $q = 0.05$ across the full sample-size grid, while pointwise inference exhibits uniformly elevated error rates on both measures. This error-rate control is robust across all configurations in Appendix B, including the persistent DGP, the LP estimator, the expanded horizon range $H = 40$, and the lag-misspecification scenario in which the estimator uses a VAR(2) while the true DGP is a VAR(4) (Figures B.9–B.10).

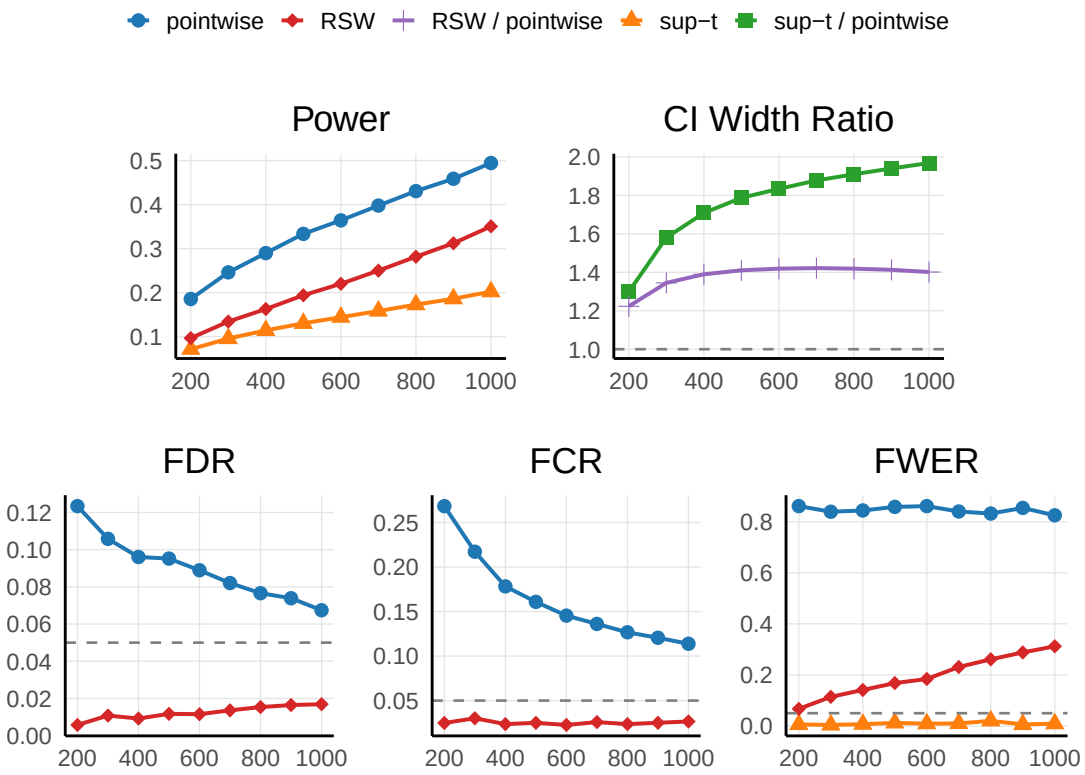


Figure 4: Monte Carlo comparison of IRF inference methods (baseline DGP, $H = 20$): 95% pointwise bands, sup-t bands controlling FWER at 5%, and RSW bands controlling FDR/FCR at $q = 0.05$; panels report power, average interval width (relative to pointwise), and realized FDR, FCR, and FWER.

4 Empirical Application: Oil Supply Shocks

We now apply the proposed FDR/FCR-adjusted reporting rule to a prominent empirical IRF design: oil supply news shocks identified by high-frequency OPEC announcement surprises in an external-instrument VAR (Känzig, 2021). A single identified shock traced across six variables and fifty-one horizons generates a 306-dimensional response surface, so the implicit hypothesis family is large even though the underlying design is standard. As in Section 3.2, we keep estimation close to the original study and report three inferential summaries of the same response family: pointwise bands, simultaneous sup-t bands controlling the FWER, and dependence-robust RSW selections with post-selection intervals controlling the FDR and FCR.

Empirical design. Känzig (2021) studies the macroeconomic effects of oil supply news, i.e., shocks to expectations about future oil supply rather than unanticipated contemporaneous supply disruptions. The identification strategy exploits the institutional timing of OPEC production announcements and measures the high-frequency change in oil futures prices in a tight window around the announcements to isolate revisions in supply expectations. The announcement-window surprises are aggregated to the monthly frequency and used as an external instrument in an oil-market VAR; the baseline surprise series is a composite measure spanning the first year of the oil futures term structure, designed to capture shifts in expected future supply rather than spot-market noise. The key feature of the identification is a distinction between news shocks and contemporaneous physical disruptions: a news shock raises the oil price on impact and causes inventories to accumulate in anticipation of future scarcity, whereas a physical disruption depletes inventories immediately.[10]

We replicate Känzig's baseline monthly six-variable VAR and identification using the authors' replication data and the same reduced-form choices for ordering, lag length, and deterministic terms. Let y_t denote the 6×1 vector collecting the real oil price, world oil production, world oil inventories, world industrial production, U.S. industrial production, and the U.S. CPI. All variables enter in log levels, so the impulse responses can be interpreted as approximate percent changes, consistent with Känzig (2021). The reduced-form dynamics are summarized by the VAR

$$y_t = c + \sum_{\ell=1}^{p} A_\ell y_{t-\ell} + u_t, \tag{36}$$

$$u_t = C\varepsilon_t, \tag{37}$$

with $p = 12$ lags and a constant. Identification follows the external-instrument (proxy-SVAR) restriction that the monthly oil-supply-surprise proxy z_t is correlated with the oil supply news shock and orthogonal to the remaining structural shocks,

$$\mathbb{E}[z_t \varepsilon_{1,t}] \neq 0, \tag{38}$$

$$\mathbb{E}[z_t \varepsilon_{j,t}] = 0, \qquad j = 2, \ldots, 6. \tag{39}$$

We normalize the identified shock so that the impact response of the real oil price is

[10]The identified shock is also contractionary and inflationary for the United States, with industrial production falling while consumer prices rise, consistent with the experience of a large net oil importer facing higher expected energy costs. We discuss these macroeconomic patterns in detail under "Results: macroeconomic block" below.

a 10% increase, matching the normalization used by Känzig (2021). Figure 5 reports the resulting impulse responses over horizons $h = 0, \ldots, 50$ months. The response family is $\mathcal{F} \equiv \{\theta_k(h) : k = 1, \ldots, 6;\ h = 0, \ldots, 50\}$ with $m = |\mathcal{F}| = 306$, and all three inferential summaries are constructed from the same residual-based bootstrap described in Section 3.2 at nominal level $\alpha^{pt} = \alpha^{FWER} = q = 0.01$.[11]

Results: oil-market block. The point estimates in Figure 5 replicate the qualitative dynamics emphasized by Känzig (2021). Consider how the three inferential layers compare for each variable in the oil-market block.

The real oil price response is significant under all three methods at impact and across nearly the full horizon range, reflecting both the normalization and the strength of the instrument (first-stage $F = 22.67$). Here FDR control and sup-t bands agree, confirming that the oil price result is among the most robust features of the response surface.

The inventory response is where the methods begin to diverge. Pointwise bands show a broad, persistent buildup. The sup-t band, calibrated to prevent any false rejection across 306 tests, struggles to detect the inventory accumulation even at horizons where the point estimates are large. FDR control recovers this finding: whiskers appear from roughly horizon 5 onward and persist through the end of the response window, placing the positive inventory response on formally reliable footing without requiring the extreme conservatism of full FWER control. This matters for the identification argument because the positive inventory response is the feature that distinguishes a news shock from a contemporaneous physical shortfall. That it survives multiplicity adjustment means the data's support for the news interpretation specifically is not an artifact of scanning a large response family.

World oil production shows a similar pattern. Pointwise bands indicate significance over a wide horizon range; sup-t bands are largely uninformative. FDR whiskers appear from roughly horizon 10 onward, where the gradual decline is deepest, but not at the early horizons where the near-zero impact response reflects the 30-day implementation lag of OPEC production decisions. The procedure thus confirms that production falls significantly with a lag, recovering what the sup-t band was too conservative to detect, while trimming horizon ranges where the pointwise evidence is most marginal.

Results: macroeconomic block. The macroeconomic block is where FDR selection departs most sharply from the pointwise narrative. Pointwise bands suggest that both

[11] In Figure 5, FDR/FCR output is displayed as whiskers at the selected horizons. Post-selection intervals use the Benjamini–Yekutieli adjustment $\alpha^\star = q\,\widehat{R}/m$, delivering FCR control at q for the reported set.

world and U.S. industrial production decline significantly and persistently, together forming a story of broad-based contractionary effects. FDR selection refines this by splitting the two activity variables apart.

The most instructive case is world industrial production. Känzig (2021) notes that global activity does not change much over the first year after the shock but then starts to fall significantly and persistently, and interprets this delayed decline as evidence that adverse general-equilibrium effects on oil-exporting countries eventually dominate their short-run gains from higher prices. FDR selection retains the short-run positive response over the first several months, which is consistent with Känzig's observation that oil-exporting countries may benefit from higher prices in the short run. However, the subsequent decline that forms the core of the global contractionary narrative does not survive multiplicity adjustment at any horizon. This does not mean the true response is zero at longer horizons, but it identifies the delayed global contraction as the portion of Känzig's narrative most vulnerable to the false-discovery problem. It is notable that this is also the variable whose longer-run response Känzig himself characterizes as delayed and modest relative to the U.S. counterpart.

In contrast, U.S. industrial production retains FDR whiskers from roughly horizon 10 onward, covering the range where the contraction deepens and persists. This is a result that the sup-t band was too conservative to support. U.S. consumer prices likewise survive from near-impact through about horizon 40, covering the full buildup of the inflationary pass-through. Together, these selections preserve the stagflationary narrative for the United States while refining the broader claim about global activity: the short-run boost to world output survives, but the delayed contraction does not. The FDR exercise thus concentrates the contractionary evidence on the economy for which the transmission channel is clearest: a historically large net oil importer whose consumers and firms face the most direct exposure to higher energy costs.

Takeaways. The FDR/FCR analysis of Känzig's oil supply news shock provides a dimension of robustness that is complementary to the extensive sensitivity analysis in the original study. Känzig (2021) demonstrates robustness along the dimensions of identification design (informationally robust surprises, ordinary versus extraordinary meetings), estimation approach (heteroskedasticity-based identification, local projections), and model specification (alternative activity indicators, oil price measures, lag orders, subsamples). The FDR analysis asks a different question: which of the reported significance patterns are robust to the multiplicity inherent in scanning a 306-dimensional response surface? The answer sharpens some conclusions and qualifies

others. The core oil-market signatures, the news-shock identification argument, and the U.S. stagflationary pattern all rest on firmer ground than pointwise bands alone can establish, precisely because they survive formal accounting for the search across variables and horizons. The delayed decline in world industrial production does not survive, identifying the part of the original narrative that is most fragile to the false-discovery problem.

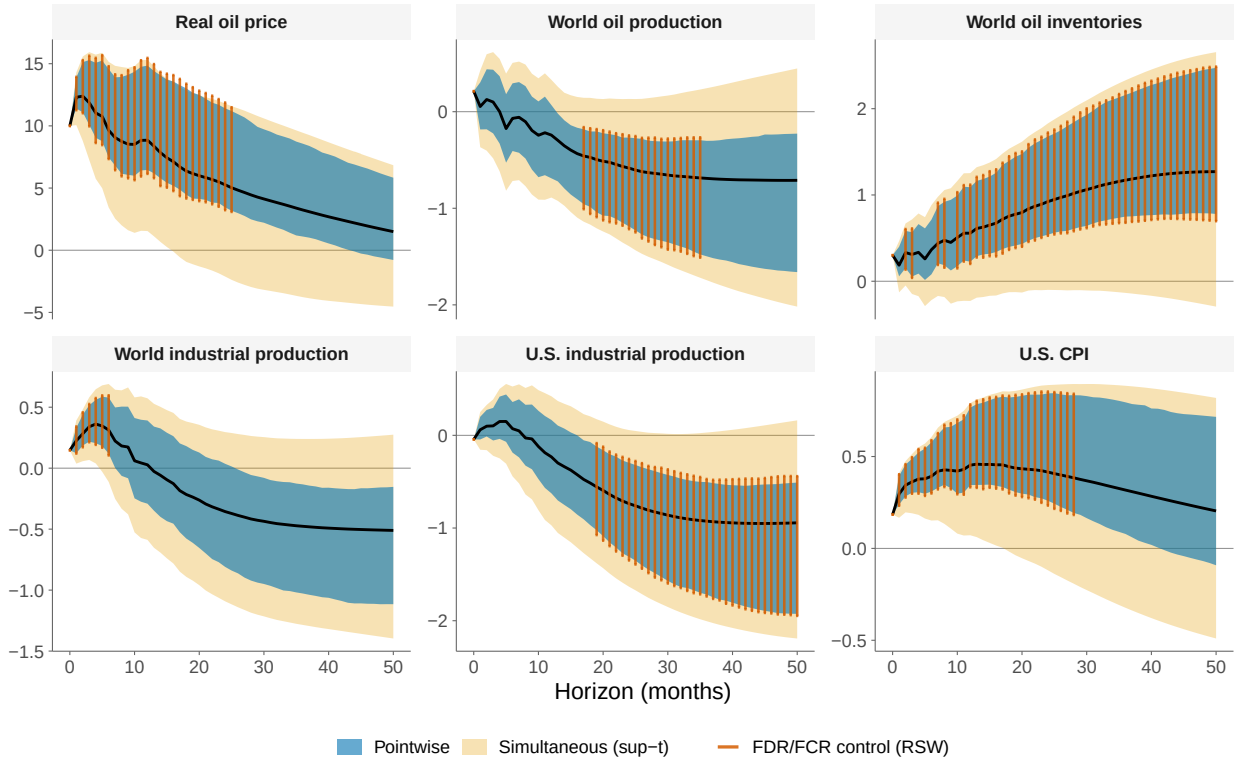


Figure 5: Impulse responses to an oil supply news shock. Pointwise bands are 99% bootstrap confidence intervals ($\alpha^{pt} = 0.01$); simultaneous sup-t bands control FWER at $\alpha^{FWER} = 0.01$ over the full response family; RSW whiskers display coefficients selected under FDR control at $q = 0.01$ together with post-selection intervals constructed at $\alpha^{\star} = q\,\widehat{R}/m$ (FCR control at $q = 0.01$).

5 Practical Guidance

Here we distill the main implementation decisions into practical guidance for applied researchers who want to adopt FDR/FCR-adjusted reporting in their own IRF analysis.

Specifying the test family. The most consequential design choice is the definition of the response family $\mathcal{F}$, because FDR and FCR guarantees apply strictly to that set of tests.

The guiding principle is that the family should include every coefficient the researcher will inspect when drawing conclusions from the IRF plot. If the researcher plans to scan responses of six variables over fifty horizons, the family is $m = 6 \times 51 = 306$. If the analysis also distinguishes two regimes, the family doubles to $m = 612$. Omitting coefficients from the family that will nonetheless be examined amounts to conducting uncontrolled pointwise inference on the omitted set, which defeats the purpose of the adjustment.

In practice, three common configurations arise. First, a single identified shock traced across multiple outcome variables and horizons, as in the oil supply application (Section 4), where $\mathcal{F}$ collects all variable-by-horizon pairs for one shock. Second, a single outcome variable examined across horizons only, where the family is the horizon vector $\{\beta_h\}_{h=1}^{H}$ within each outcome equation. Third, a state-dependent design in which each regime contributes its own set of horizon-specific coefficients, where the family includes both $\{\beta_h^A\}$ and $\{\beta_h^B\}$ and potentially also the difference $\{\beta_h^A - \beta_h^B\}$ if the researcher intends to test for state dependence directly.

Two boundary cases deserve comment. When the researcher examines multiple shocks, the family can either pool all shocks into a single test family (so that the FDR guarantee applies to the union of all reported findings) or treat each shock as a separate family (so that FDR is controlled within each shock but not across shocks). Pooling is more conservative but more honest when the researcher's narrative draws on cross-shock comparisons. At the other extreme, if only a single pre-specified coefficient is of interest, such as the output response at a particular horizon, there is no multiplicity problem and pointwise inference is appropriate. The procedure is designed for the intermediate case that dominates applied practice: the researcher has not pre-committed to a single coefficient but will scan a grid and highlight significant features.

Choosing the target level q. The target q controls the expected share of false discoveries among the selected responses and, through the FCR adjustment, the expected share of noncovering intervals among those reported. A smaller q produces fewer but more reliable selections; a larger q produces more selections at the cost of a higher expected false discovery proportion.

There is no single correct choice, and q should reflect the tolerance for false positives in the specific application. A useful calibration is to ask: if the procedure selects $\widehat{R}$ responses as significantly different from zero, how many of those would I tolerate being wrong? At $q = 0.05$, roughly one in twenty selected responses is expected to be a false discovery. At $q = 0.10$, roughly one in ten. At $q = 0.01$, roughly one in a hundred. For exploratory analyses where the goal is to map out a broad transmission mechanism and false positives

will be checked in subsequent work, $q = 0.10$ is reasonable. For confirmatory analyses where each selected response will be interpreted as established evidence, $q = 0.01$ or $q = 0.05$ may be more appropriate.

One practical consideration is that q also enters the FCR adjustment through $\alpha^* = q\,\widehat{R}/m$. When q is small and $\widehat{R}$ is moderate, α^* can become very small, producing wide post-selection intervals. If the resulting intervals are uninformatively wide, this is a signal that the data do not support confident statements about those coefficients at the chosen q, not a reason to raise q after the fact. We recommend reporting results at a single pre-committed q in the main text, with robustness to alternative levels available elsewhere.

Choosing the selection rule. Section 3 describes three selection rules: RSW, BH, and BY. Table 1 summarizes the tradeoffs.

Table 1: Selection rule comparison

	RSW	**BH**	**BY**
FDR guarantee	Under arbitrary dependence	Under independence or PRDS	Under arbitrary dependence
Inputs required	Joint bootstrap draws of full test statistic vector	Marginal p-values only	Marginal p-values only
Power	Highest among the three	Close to RSW in simulations	Noticeably lower
Computation	Requires joint bootstrap (same as sup-t)	Minimal	Minimal

Our baseline recommendation is RSW, because it is the only procedure with a formal FDR guarantee under the dependence structures that arise generically in IRF applications. Researchers who already compute sup-t bands have the joint bootstrap machinery in hand and can implement RSW at negligible additional cost. BH is a practical alternative when the joint bootstrap is unavailable or computationally prohibitive, with the caveat that its theoretical guarantee does not formally cover the IRF dependence setting, even though it performs well in our simulations. BY is the most conservative option and is appropriate when the researcher wants a dependence-robust guarantee without the computational cost of the bootstrap, at the expense of fewer discoveries.

Regardless of which selection rule is used, the FCR adjustment step is identical: once the rejection set $\widehat{\mathcal{R}}$ is obtained, post-selection intervals are constructed at level $\alpha^* = q\,\widehat{R}/m$ using the same bootstrap (or asymptotic) machinery that produced the pointwise intervals.

Bootstrap implementation. The procedure does not require a new bootstrap; it reuses the same joint bootstrap draws that standard VAR or LP workflows already produce for pointwise intervals and sup-t bands. Two implementation details are worth noting. First, the number of bootstrap draws B should be large enough that the bootstrap distribution of the test statistic vector is well approximated in the tails, because the RSW stepdown recursion and the FCR-adjusted quantiles both depend on tail behavior. We use $B = 2{,}000$ throughout and find this sufficient; $B = 1{,}000$ is a reasonable lower bound for moderate m, but larger B may be needed when m exceeds several hundred. Second, for VARs, bias correction of the bootstrap DGP (Pope, 1990) can improve the centering of the bootstrap distribution in small samples without affecting the RSW or FCR logic. For LPs, the VAR-based wild bootstrap of Montiel Olea and Plagborg-Møller (2019) preserves cross-horizon dependence and delivers studentized statistics directly.

Reading and presenting FDR/FCR-adjusted IRF plots. The visual output of the procedure differs from standard IRF reporting in two ways: not all coefficients receive confidence intervals, and the intervals that are reported may vary in width across horizons.

We recommend the following plotting conventions. Plot the point estimates for all horizons and variables as a connected line, exactly as in a standard IRF figure, so that the full dynamic profile is visible. Overlay the FDR-selected coefficients with error bars (whiskers) at the FCR-adjusted level, and leave non-selected horizons without any interval. The absence of a whisker at a given horizon means that the coefficient was not selected as significantly different from zero after FDR control; it does not mean the true response is zero. If space permits, pointwise bands and sup-t bands can be displayed in the background (e.g., as shaded regions) to provide a visual benchmark, as in Figures 3 and 5.

When describing the results in text, we suggest three elements. First, state the family size m, the target level q, and the number of selections $\widehat{R}$, so the reader can assess the severity of the multiplicity adjustment. Second, describe which variables and horizon ranges survive selection, because this is the primary output of the procedure. Third, note where the FDR-selected set differs from the pointwise set, because these discrepancies identify the features of the pointwise narrative that are most fragile to multiplicity. Conversely, features that survive FDR selection can be described with greater confidence as robust dynamic effects.

When to use FDR/FCR control versus simultaneous bands. FDR/FCR-adjusted reporting and simultaneous (FWER-controlling) bands answer different inferential questions, and the appropriate choice depends on the research objective. If the goal

is to make a uniform statement about the entire response path (e.g. "the impulse response is everywhere contained within this band"), then simultaneous bands are the correct tool and FDR/FCR control is not a substitute. If the goal is to identify which specific horizons, variables, or states exhibit significant responses while controlling the reliability of this identification, then FDR/FCR control is the appropriate target.

In most applied IRF work, the latter objective is closer to what researchers actually do: they scan the response grid, highlight features that appear nonzero, and build a narrative from the selected set. For this common use case, FDR/FCR-adjusted reporting provides an explicit guarantee on the quality of the selected set that neither pointwise bands (which ignore multiplicity) nor simultaneous bands (which target a different error rate) deliver.

6 Conclusion

This paper argues that impulse response analysis should be treated as a multiple-testing problem. In typical applications, researchers interpret significance patterns after scanning a high-dimensional grid of responses across horizons, variables, shocks, and sometimes states. For that use case, pointwise inference is too permissive and simultaneous inference is often too conservative. We propose instead an FDR/FCR-based reporting framework that controls the expected share of false discoveries among reported responses and the expected share of noncovering post-selection intervals.

The results show that this change in inferential target matters in practice. In simulations, the proposed procedure delivers substantial power gains relative to simultaneous bands while maintaining control of false discoveries and false coverage across a range of empirically relevant designs. In the monetary policy application, it largely preserves the qualitative transmission narrative that pointwise inference suggests, but places that narrative on firmer statistical footing. In the oil supply news application, it preserves the core oil-market and U.S. stagflationary responses while showing that some features of the pointwise narrative, especially the delayed decline in world industrial production, are not robust to multiplicity adjustment.

More broadly, the paper's message is practical as well as methodological. Researchers do not usually use IRFs to make a uniform statement about an entire response path; they use them to identify and interpret selected dynamic effects. When that is the objective, inference should be calibrated to the reliability of the selected set rather than to the probability of avoiding any false rejection anywhere in the response family. Because the procedure reuses the same joint bootstrap machinery already standard in VAR and local projection workflows, the barrier to adoption is not computational but conceptual:

recognizing that when researchers scan IRF grids to build economic narratives, inference should govern the reliability of the narrative they report, not the family they searched.

References

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Auerbach, A. J. and Gorodnichenko, Y. (2012). Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy*, 4(2):1–27.

Auerbach, A. J. and Gorodnichenko, Y. (2013). Fiscal multipliers in recession and expansion. *NBER Chapters*, pages 63–98.

Barras, L., Scaillet, O., and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The Journal of Finance*, 65(1):179–216.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81.

Bruder, S. and Wolf, M. (2018). Balanced bootstrap joint confidence bands for structural impulse response functions. *Journal of Time Series Analysis*, 39(5):641–664.

Bu, C., Rogers, J., and Wu, W. (2021). A unified measure of fed monetary policy shocks. *Journal of Monetary Economics*, 118:331–349.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy*, 113(1):1–45.

Freyaldenhoven, S., Hansen, C., and Shapiro, J. M. (2019). Pre-event trends in the panel event-study design. *American Economic Review*, 109(9):3307–3338.

Gertler, M. and Karadi, P. (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, 7(1):44–76.

Gilchrist, S. and Zakrajšek, E. (2012). Credit spreads and business cycle fluctuations. *American Economic Review*, 102(4):1692–1720.

Harvey, C. R. and Liu, Y. (2020). False (and missed) discoveries in financial economics. *The Journal of Finance*, 75(5):2503–2553.

Information Systems and Wake Forest University (2021). WFU High Performance Computing Facility.

Inoue, A., Jordà, Ò., and Kuersteiner, G. M. (2026). Inference for local projections. *The Econometrics Journal*, 29(1):2–26.

Inoue, A. and Kilian, L. (2016). Joint confidence sets for structural impulse responses. *Journal of Econometrics*, 192(2):421–433.

Inoue, A. and Kilian, L. (2022). Joint Bayesian inference about impulse responses in VAR models. *Journal of Econometrics*, 231(2):457–476.

Jordà, Ò. (2009). Simultaneous confidence regions for impulse responses. *Review of Economics and Statistics*, 91(3):629–647.

Jordà, Ò. (2023). Local projections for applied economics. *Annual Review of Economics*, 15:607–631.

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

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics*, 80(2):218–230.

Känzig, D. R. (2021). The macroeconomic effects of oil supply news: Evidence from opec announcements. *American Economic Review*, 111(4):1092–1125.

Lee, S. and Shaikh, A. M. (2014). Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of PROGRESA on school enrollment. *Journal of Applied Econometrics*, 29(4):612–626.

List, J. A., Shaikh, A. M., and Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4):773–793.

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

Lütkepohl, H., Staszewska-Bystrova, A., and Winker, P. (2015). Confidence bands for impulse responses: Bonferroni vs. wald. *Oxford Bulletin of Economics and Statistics*, 77(5):783–799.

Lütkepohl, H., Staszewska-Bystrova, A., and Winker, P. (2020). Constructing joint confidence bands for impulse response functions of var models—a review. *Econometrics and Statistics*, 13:69–83.

Montiel Olea, J. L. and Plagborg-Møller, M. (2019). Simultaneous confidence bands: Theory, implementation, and an application to svars. *Journal of Applied Econometrics*, 34(1):1–17.

Montiel Olea, J. L. and Plagborg-Møller, M. (2021). Local projection inference is simpler and more robust than you think. *Econometrica*, 89(4):1789–1823.

Pope, A. L. (1990). Biases of estimators in multivariate non-gaussian autoregressions. *Journal of Time Series Analysis*, 11(3):249–258.

Ramey, V. A. (2011). Identifying government spending shocks: It's all in the timing. *Quarterly Journal of Economics*, 126(1):1–50.

Ramey, V. A. and Zubairy, S. (2018). Government spending multipliers in good times and in bad: Evidence from US historical data. *Journal of Political Economy*, 126(2):850–901.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17(3):417–442.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2010). Hypothesis testing in econometrics. *Annual Review of Economics*, 2:75–104.

Romano, J. P. and Wolf, M. (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Romano, J. P. and Wolf, M. (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4):1378–1408.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4):1055–1084.

Sims, C. A. and Zha, T. (1999). Error bands for impulse responses. *Econometrica*, 67(5):1113–1155.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, 64(3):479–498.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Appendices

A Simulation Design

A.1 Overview and indexing

This appendix provides the complete algorithmic specification of the simulation design summarized in Section 3. Sections A.1–A.2 describe the DGPs and design parameters. Sections A.3–A.4 give the full bootstrap and inference algorithms, including implementation details not covered in Section 3.1.

A *scenario* is defined by a triple (d, e, T) consisting of a data-generating process (DGP) index d, an estimator label $e \in \{\text{VAR}, \text{LP}\}$, and a sample size T. For each scenario, the simulation performs R Monte Carlo replications. In each replication it:

(i) simulates a multivariate time series $\{y_t\}_{t=1}^{T}$ from the DGP (after burn-in), together with the structural shock series $\{\varepsilon_t\}_{t=1}^{T}$ and reduced-form innovations $\{u_t\}_{t=1}^{T}$;

(ii) computes an estimated impulse response function (IRF) vector $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^m$ using estimator e;

(iii) runs a bootstrap with B draws to construct coefficientwise p-values, pointwise confidence intervals, and a sup-t simultaneous band;

(iv) applies the Romano–Shaikh–Wolf (RSW) stepdown procedure for multiple testing and constructs post-selection intervals;

(v) computes replication-level performance metrics by comparing inference outcomes to the DGP-implied "true" IRF vector $\boldsymbol{\theta}^0$.

Indexing conventions: sample size $T \in \mathcal{T}$; replication $r \in \{1, \ldots, R\}$; bootstrap draw $b \in \{1, \ldots, B\}$; horizon $h \in \{0, \ldots, H\}$; variable index $k \in \{1, \ldots, K\}$; shock index $s \in \mathcal{S} = \{1, \ldots, S\}$; coefficient index $j \in \{1, \ldots, m\}$.

A.2 Design parameters

Table A.1 collects baseline scalar design choices and indicates which parameters vary across scenario families.

Parameters held fixed across scenarios are K, S, $\mathcal{S}$, R, B, T_{burn}, α_{pt}, α_{FWER}, q, and $\mathcal{T}$.

For each replication, data are generated for $T_{\text{burn}} + T$ periods and the first T_{burn} observations are discarded, so the effective estimation sample is exactly T.

Table A.1: Simulation design parameters

Symbol	Value	Description
K	6	Number of endogenous variables
p	4 (DGP); $p_{\text{est}} = 2$ in misspec	VAR lag order (varies only in lag-misspecification estimation)
H	19 (baseline); also 9, 39	Maximum IRF horizon (varies across family specifications)
S	3	Number of shocks reported/tested
$\mathcal{S}$	$\{1, \dots, S\}$	Shock index set
m	360 (baseline); also 180, 720, 240	Total number of scalar IRF coefficients (varies with family)
R	1000	Monte Carlo replications
B	2000	Bootstrap draws
T_{burn}	1000	Burn-in length
α_{pt}	0.05	Pointwise test/interval level
α_{FWER}	0.05	FWER target for the sup-t band
q	0.05	FDR and FCR target for RSW reporting
$\mathcal{T}$	$\{200, 300, \dots, 1000\}$	Sample-size grid

A.3 Data generation step: generic DGP interface

Generic interface. Each DGP specifies (i) a list of p coefficient matrices $(A_1, \dots, A_p)$ with $A_\ell \in \mathbb{R}^{K\times K}$ and (ii) an impact matrix $C \in \mathbb{R}^{K\times K}$ mapping structural shocks to reduced-form innovations. Given $(A_1, \dots, A_p, C)$, a sample size T, and a burn-in length T_{burn}, the design returns

$$\{y_t\}_{t=1}^{T} \in \mathbb{R}^{T\times K}, \qquad \{\varepsilon_t\}_{t=1}^{T} \in \mathbb{R}^{T\times K}, \qquad \{u_t\}_{t=1}^{T} \in \mathbb{R}^{T\times K},$$

where y_t is the observed series, ε_t is the structural shock vector, and u_t is the reduced-form innovation.

Law of motion, initialization, and burn-in. Data are generated from the stable VAR(p) recursion

$$y_t = \sum_{\ell=1}^{p} A_\ell y_{t-\ell} + u_t, \tag{A.1}$$

$$u_t = C\,\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I_K) \text{ i.i.d. across } t. \tag{A.2}$$

The process is initialized at zero for the first p observations,

$$y_1 = y_2 = \cdots = y_p = 0, \tag{A.3}$$

then iterated to time $T_{\text{burn}}+T$. The first T_{burn} observations are discarded, and the remaining T observations are retained. No deterministic terms (intercept or trend) are included in (A.1).

True IRFs and vectorization. Define reduced-form moving-average (MA) coefficient matrices $\{\Phi_h\}_{h=0}^{H}$ implied by $(A_1, \ldots, A_p)$:

$$\Phi_0 = I_K, \tag{A.4}$$

$$\Phi_h = \sum_{\ell=1}^{p} A_\ell \Phi_{h-\ell}, \qquad h \geq 1, \tag{A.5}$$

with the convention $\Phi_h = 0$ for $h < 0$. Structural IRF matrices are

$$\Psi_h = \Phi_h C, \qquad h = 0, 1, \ldots, H. \tag{A.6}$$

For variable $k \in \{1, \ldots, K\}$, shock $s \in \mathcal{S}$, and horizon $h \in \{0, \ldots, H\}$, the scalar IRF coefficient is

$$\theta^0_{k,s}(h) = e_k^\top \Psi_h e_s, \tag{A.7}$$

where e_k and e_s are canonical basis vectors. IRFs are stored as a vector $\boldsymbol{\theta} \in \mathbb{R}^m$ using the index map

$$j(k, s, h) = \big((k-1)S + (s-1)\big)(H+1) + (h+1), \tag{A.8}$$

so that $\theta_{j(k,s,h)} = \theta_{k,s}(h)$; ordering is k major, then s, then h (with h fastest).

Specification of the baseline and persistent DGPs. The simulation uses two DGPs. In both, the lag matrices and impact matrix are block-diagonal as in Section 2 (equation (12)):

$$A_\ell = \begin{bmatrix} A_{\ell,1} & 0 \\ 0 & A_{\ell,2} \end{bmatrix}, \qquad \ell = 1, \ldots, 4, \qquad C = 1.25 \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix}, \tag{A.9}$$

with two 3×3 blocks. The baseline DGP chooses $(A_1, \ldots, A_4)$ to be stable and non-oscillatory, with companion-form spectral radius strictly below one. The persistent DGP keeps the same block structure and the same C, and rescales the lag matrices by a common factor so that the companion spectral radius is approximately 0.98. This block-diagonal design implies that all cross-block responses are exact zeros; with $K = 6$, $S = 3$, and

$H = 19$, a known large fraction of the $m = 360$ coefficients is therefore null by construction, providing clean ground truth for evaluating FDR control.

A.4 Estimation methods

VAR estimation. VAR estimation follows Section 3 ("VAR and LP implementation"): the VAR is fit by OLS with lag order p fixed, and identification is treated as known through a fixed impact matrix C. For notation in the simulation design, estimated VAR coefficients are mapped into IRFs through (A.4)–(A.6).

Local projections (LP). LP estimation follows Section 3 ("VAR and LP implementation"): horizon-by-horizon regressions use the structural shock as the impulse regressor and include $p + 1$ lags of y as controls. For horizon h, the sample range is

$$t \in \{p + 2, \ldots, T - h\}, \tag{A.10}$$

so the number of usable observations is $T - h - (p + 1)$. Eicker–Huber–White (HC0) standard errors are used for LP shock coefficients.

A.5 Inference procedures

All inference procedures target coefficientwise null hypotheses

$$H_j : \theta_j^0 = 0, \qquad j = 1, \ldots, m. \tag{A.11}$$

A.5.1 Pointwise inference

Pointwise bootstrap p-values and confidence intervals follow Section 3 ("Empirical objects and test statistics" and "FCR-adjusted confidence intervals"), with estimator-specific implementation details reported here.

VAR bootstrap pseudo-data and bias-corrected bootstrap DGP. Let $\widehat{u}_t$ denote residuals from the OLS VAR fit for $t = p + 1, \ldots, T$, stacked into an $n \times K$ matrix with $n = T - p$. For each bootstrap draw b, residuals are resampled i.i.d. with replacement:

$$\widehat{u}_{1:n}^{*(b)} \text{ are sampled i.i.d. with replacement from } \{\widehat{u}_{1:n}\}. \tag{A.12}$$

Given initial conditions $y_{1:p}^{*(b)} = y_{1:p}$, bootstrap pseudo-data are generated recursively:

$$y_t^{*(b)} = \widetilde{c} + \sum_{\ell=1}^{p} \widetilde{A}_\ell y_{t-\ell}^{*(b)} + \widehat{u}_{t-p}^{*(b)}, \qquad t = p+1, \ldots, T. \tag{A.13}$$

The bootstrap DGP coefficients $(\widetilde{A}_\ell, \widetilde{c})$ are bias-corrected using an internal residual bootstrap of size B_{bias}:

$$B_{\text{bias}} = B = 2000, \tag{A.14}$$

$$A_\ell^{\text{bc}} = 2\widehat{A}_\ell - \frac{1}{B_{\text{bias}}} \sum_{b=1}^{B_{\text{bias}}} \widehat{A}_\ell^{*(b)}, \qquad c^{\text{bc}} = 2\widehat{c} - \frac{1}{B_{\text{bias}}} \sum_{b=1}^{B_{\text{bias}}} \widehat{c}^{*(b)}. \tag{A.15}$$

Stability is enforced through the companion-matrix condition

$$\rho(\text{companion}(A^{\text{bc}})) < 0.999. \tag{A.16}$$

If (A.16) fails, the correction is halved repeatedly (up to 25 steps); if no stable correction is found, the uncorrected OLS coefficients are used.

VAR studentization floor and LP-specific bootstrap details. For VAR studentization, the across-draw standard deviation is

$$\widehat{\sigma}_j = \text{sd}_{b=1,\ldots,B}\left(\theta_j^{*(b)}\right), \tag{A.17}$$

with $\widehat{\sigma}_j$ floored at 10^{-12}, and

$$T_j^{\text{obs}} = \frac{\widehat{\theta}_j}{\widehat{\sigma}_j}, \qquad T_j^{*(b)} = \frac{\theta_j^{*(b)} - \widehat{\theta}_j}{\widehat{\sigma}_j}. \tag{A.18}$$

For LP, the auxiliary VAR-based wild bootstrap uses the same general setup as Section 3. The implementation draws random initial blocks and Gaussian wild multipliers,

$$(y_1^{*(b)}, \ldots, y_p^{*(b)}) = (y_\tau, \ldots, y_{\tau+p-1}), \qquad \tau \sim \text{Unif}\{1, \ldots, T-p+1\}, \tag{A.19}$$

$$U_t \sim \mathcal{N}(0,1) \text{ i.i.d.}, \qquad t = 1, \ldots, T-p, \tag{A.20}$$

and generates

$$y_t^{*(b)} = \widehat{c} + \sum_{\ell=1}^{p} \widehat{A}_\ell y_{t-\ell}^{*(b)} + U_{t-p}\widehat{u}_{t-p}, \qquad t = p+1, \ldots, T, \tag{A.21}$$

with structural shocks recovered via

$$\varepsilon_t^{*(b)} = \left(C^\top\right)^{-1} \left(U_{t-p}\widehat{u}_{t-p}\right)^\top, \qquad t = p+1, \ldots, T. \tag{A.22}$$

LP bootstrap-t statistics are centered at the VAR-implied IRF:

$$T_j^{*(b)} = \frac{\widehat{\theta}_j^{*(b)} - \widehat{\theta}_j^{\text{VAR}}}{\widehat{s}_j^{*(b)}}. \tag{A.23}$$

The recentered LP draws are

$$\theta_j^{*(b)} = \widehat{\theta}_j + \left(\widehat{\theta}_j^{*(b)} - \widehat{\theta}_j^{\text{VAR}}\right), \tag{A.24}$$

and the LP point estimate is bias-corrected as

$$\widehat{\theta}_j^{\text{bc}} = 2\widehat{\theta}_j - \frac{1}{B}\sum_{b=1}^{B} \theta_j^{*(b)}. \tag{A.25}$$

A.5.2 sup-t simultaneous inference

The sup-t simultaneous band follows Section 3; the implementation uses the estimator-specific bootstrap-t statistics already defined in this appendix. For each draw,

$$M^{+(b)} = \max_{1\le j\le m} T_j^{*(b)}, \qquad M^{-(b)} = \max_{1\le j\le m} \left(-T_j^{*(b)}\right), \tag{A.26}$$

and the resulting band is

$$\text{Band}_j^{\text{sup-}t} = \left[\widehat{\theta}_j - c_{\text{lo}}\, s_j,\ \widehat{\theta}_j + c_{\text{hi}}\, s_j\right], \tag{A.27}$$

where $s_j = \widehat{\sigma}_j$ for VAR and $s_j = \widehat{s}_j$ for LP.

A.5.3 RSW stepdown inference and post-selection intervals

The RSW and FCR framework follows Section 3; this subsection records the implementation details used in the simulations.

Statistics and ordering. Let

$$\widehat{\sigma}_j^{\mathrm{RSW}} = \mathrm{sd}_{b=1,\dots,B}\left(\theta_j^{*(b)}\right), \tag{A.28}$$

floored at 10^{-12}, and define

$$T_j^{\mathrm{obs}} = \frac{|\widehat{\theta}_j|}{\widehat{\sigma}_j^{\mathrm{RSW}}}, \qquad T_j^{*(b)} = \frac{|\theta_j^{*(b)} - \widehat{\theta}_j|}{\widehat{\sigma}_j^{\mathrm{RSW}}}. \tag{A.29}$$

Let π satisfy

$$T_{\pi(1)}^{\mathrm{obs}} \le T_{\pi(2)}^{\mathrm{obs}} \le \cdots \le T_{\pi(m)}^{\mathrm{obs}}. \tag{A.30}$$

Prefix-maximum stepdown construction. For each $j = 1, \dots, m$, define

$$M_j^{(b)} = \max_{1 \le t \le j} T_{(t)}^{*(b)}. \tag{A.31}$$

For $j \ge 2$, set

$$t_{\mathrm{fail}}^{(b)}(j) = \max\left\{t \in \{1, \dots, j-1\} : T_{(t)}^{*(b)} < c_t\right\}, \tag{A.32}$$

with convention $t_{\mathrm{fail}}^{(b)}(j) = 0$ if the set is empty, then

$$k^{(b)}(j) = j - t_{\mathrm{fail}}^{(b)}(j), \qquad w^{(b)}(j) = \frac{k^{(b)}(j)}{m - j + k^{(b)}(j)}. \tag{A.33}$$

The objective used to select each critical value is

$$g_j(c) = \frac{1}{B} \sum_{b=1}^{B} w^{(b)}(j)\, \mathbb{I}\{M_j^{(b)} \ge c\}. \tag{A.34}$$

The rejection index is

$$j_0 = \min\left\{j : T_{(i)}^{\mathrm{obs}} \ge c_i \text{ for all } i = j, \dots, m\right\}, \tag{A.35}$$

with no rejections if the set is empty.

FCR-adjusted reported intervals. With $\widehat{r}_j \in \{0,1\}$, $R = \sum_{j=1}^{m} \widehat{r}_j$, and

$$\alpha^\star = \frac{R}{m}\, q, \tag{A.36}$$

reported intervals for selected coefficients use

$$\mathrm{CI}_j^{\mathrm{FCR}} = \left[\widehat{\theta}_j - Q_{1-\alpha^\star/2}\left(\theta_j^{*(b)} - \widehat{\theta}_j\right),\ \widehat{\theta}_j - Q_{\alpha^\star/2}\left(\theta_j^{*(b)} - \widehat{\theta}_j\right)\right], \qquad \text{if } \widehat{r}_j = 1, \tag{A.37}$$

and are left unreported when $\widehat{r}_j = 0$.

A.6 Performance criteria

Performance metrics are computed at the replication level and then averaged across replications within each scenario.

True null classification. A coefficient is treated as a "true zero" if

$$|\theta_j^0| \le 10^{-12}. \tag{A.38}$$

Define

$$\mathcal{H}_0 = \{j : |\theta_j^0| \le 10^{-12}\}, \qquad \mathcal{H}_1 = \{1, \ldots, m\} \setminus \mathcal{H}_0. \tag{A.39}$$

Discovery counts, FDP, and FWER. For a given inference method, let $\widehat{r}_j \in \{0,1\}$ denote the method's rejection/selection indicator. Define the number of discoveries

$$R = \sum_{j=1}^{m} \widehat{r}_j, \tag{A.40}$$

and false discoveries

$$V = \sum_{j \in \mathcal{H}_0} \widehat{r}_j. \tag{A.41}$$

The false discovery proportion is

$$\mathrm{FDP} = \begin{cases} V/R, & R > 0, \\ 0, & R = 0. \end{cases} \tag{A.42}$$

The family-wise error indicator is

$$\mathrm{FWER} = \mathbb{I}\{V > 0\}. \tag{A.43}$$

Power. Let $\mathrm{TP} = \sum_{j \in \mathcal{H}_1} \widehat{r}_j$ and $|\mathcal{H}_1|$ be the number of truly nonzero coefficients. Power is

$$\mathrm{Power} = \begin{cases} \mathrm{TP}/|\mathcal{H}_1|, & |\mathcal{H}_1| > 0, \\ \text{undefined}, & |\mathcal{H}_1| = 0. \end{cases} \tag{A.44}$$

Post-selection coverage: FCP and mean width. Each inference method supplies an interval/band $\mathcal{I}_j = [L_j, U_j]$. Coverage is evaluated by

$$\mathbb{I}\{\theta_j^0 \in \mathcal{I}_j\} = \mathbb{I}\{L_j \le \theta_j^0 \le U_j\}. \tag{A.45}$$

Define the number of selected noncovering intervals

$$V_{\mathrm{CI}} = \sum_{j=1}^{m} \widehat{r}_j \, \mathbb{I}\{\theta_j^0 \notin \mathcal{I}_j\}. \tag{A.46}$$

The false coverage proportion is

$$\mathrm{FCP} = \begin{cases} V_{\mathrm{CI}}/R, & R > 0, \\ 0, & R = 0. \end{cases} \tag{A.47}$$

Mean interval width is computed over selected coefficients:

$$\mathrm{MeanWidth} = \begin{cases} \frac{1}{R} \sum_{j:\widehat{r}_j=1} (U_j - L_j), & R > 0, \\ \text{undefined}, & R = 0. \end{cases} \tag{A.48}$$

For RSW, the value $\alpha^\star$ in (A.36) is recorded (undefined when $R = 0$).

Simultaneous coverage of the sup-t band. For the sup-t method, record whether the entire m-vector is covered:

$$\text{SimulCov} = \mathbb{I}\{\theta_j^0 \in \text{Band}_j^{\text{sup-}t} \text{ for all } j = 1, \ldots, m\}. \tag{A.49}$$

This metric is undefined for other methods.

Monte Carlo aggregation. Let M_r denote any replication-level metric in a fixed scenario. Scenario-level summaries are computed as

$$\overline{M} = \frac{1}{R} \sum_{r=1}^{R} M_r. \tag{A.50}$$

Monte Carlo uncertainty is summarized by the sample standard deviation of $\{M_r\}_{r=1}^{R}$ and the corresponding standard error $\text{sd}(M_r)/\sqrt{R}$.

B Simulation Results

The following figures summarize simulation performance for the scenarios defined in Appendix A. Design dimensions ($K = 6$, $p = 4$, $S = 3$, $\mathcal{T}$, and bootstrap details) are given in Sections A.2–A.5. Simulations were run using the Wake Forest University High Performance Computing Facility (Information Systems and Wake Forest University, 2021).

DGPs. Two data-generating processes are simulated.

- **Baseline DGP**: the block-diagonal VAR(p) described in §A.3—stable, non-oscillatory dynamics; cross-block IRFs are exactly zero.
- **Persistent DGP**: same block structure and impact matrix C as the baseline; lag matrices are scaled so the companion spectral radius is approximately 0.98.

Hypothesis families. For each DGP, five hypothesis-family specifications are run. In each case the coefficient index set is $j \in \{1, \ldots, m\}$ with $m = KS(H + 1)$ unless a subset is specified.

- **Baseline family (H20)**: horizons $h \in \{0, \ldots, H\}$ with $H = 19$ (20 horizons); full coefficient set; $m = 360$. Estimator lag order $p = 4$.
- **Fewer horizons (H10)**: $H = 9$ (10 horizons); full set; $m = 180$; $p = 4$.
- **More horizons (H40)**: $H = 39$ (40 horizons); full set; $m = 720$; $p = 4$.
- **Lag misspecification (misspec)**: $H = 19$, full set; $m = 360$. The estimator uses $p_{\text{est}} = 2$ (VAR(2) and LP with 2 lags) while the DGP remains VAR(4).
- **True-null sparsity (60nulls)**: $H = 19$; hypothesis family restricted to variables $k \in \{1, \ldots, 4\}$, shocks $s \in \{1, 2, 3\}$, horizons $h \in \{0, \ldots, 19\}$; $m = 4 \times 3 \times 20 = 240$. By the block-diagonal structure, 60 of these coefficients are true nulls (cross-block responses).

Estimators and layout. For each DGP and each hypothesis family, results are produced for both the **VAR** estimator (bias-corrected residual bootstrap, §A.5) and the **LP** estimator (VAR-based wild bootstrap, §A.5). Within each figure, the VAR panel appears on top and the LP panel on the bottom. For the H40 specification, LP simulation outputs are still being generated; the LP panel is shown as a placeholder where applicable.

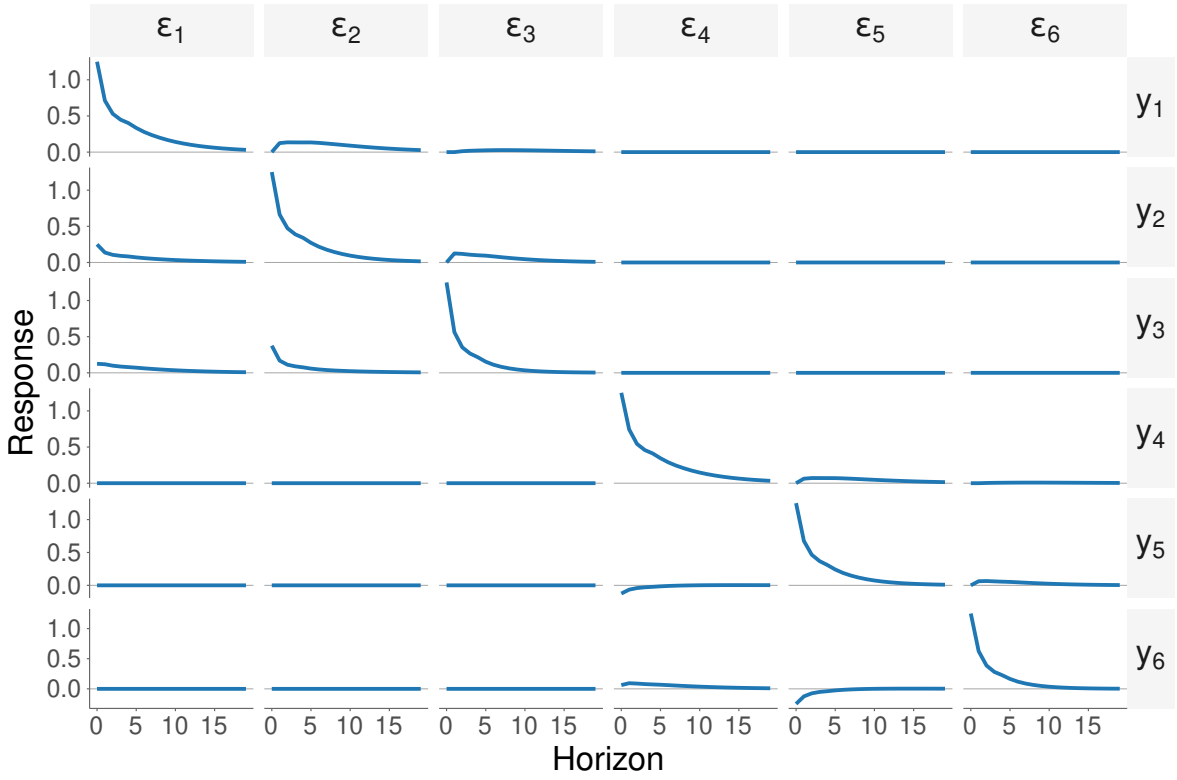


Figure B.1: True IRF grid for the baseline DGP (`dgp_baseline`).

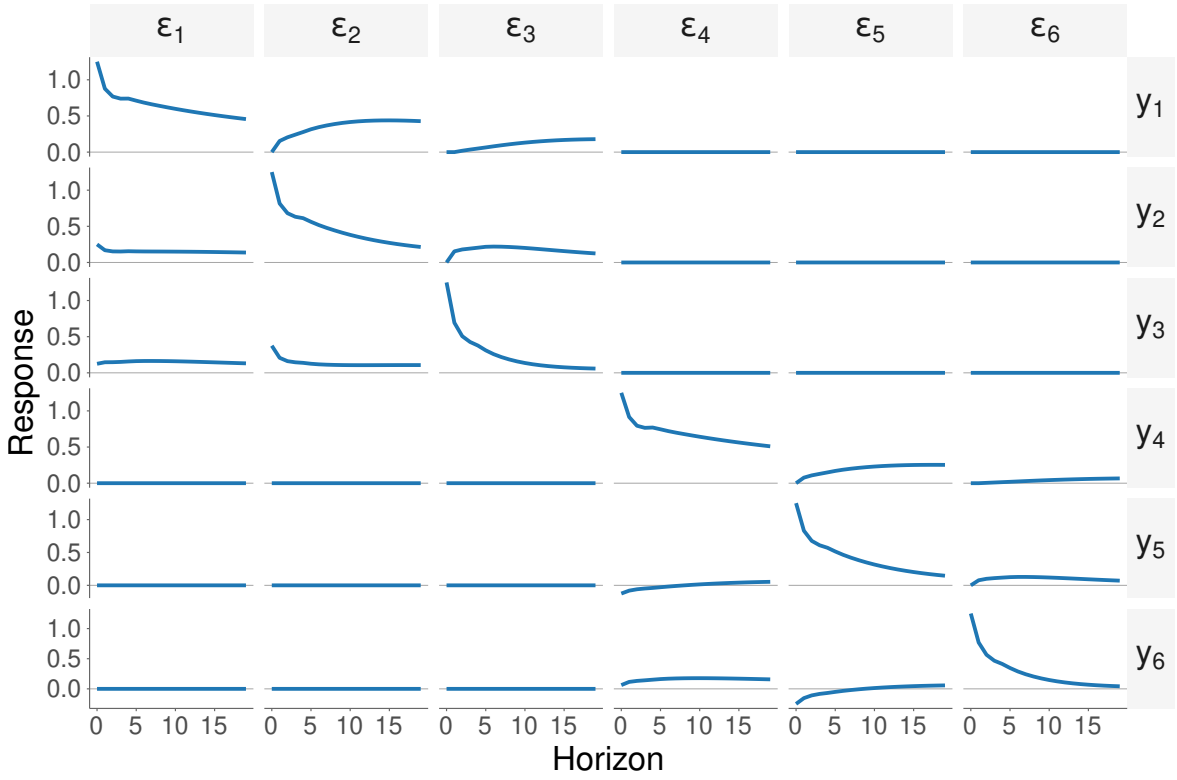
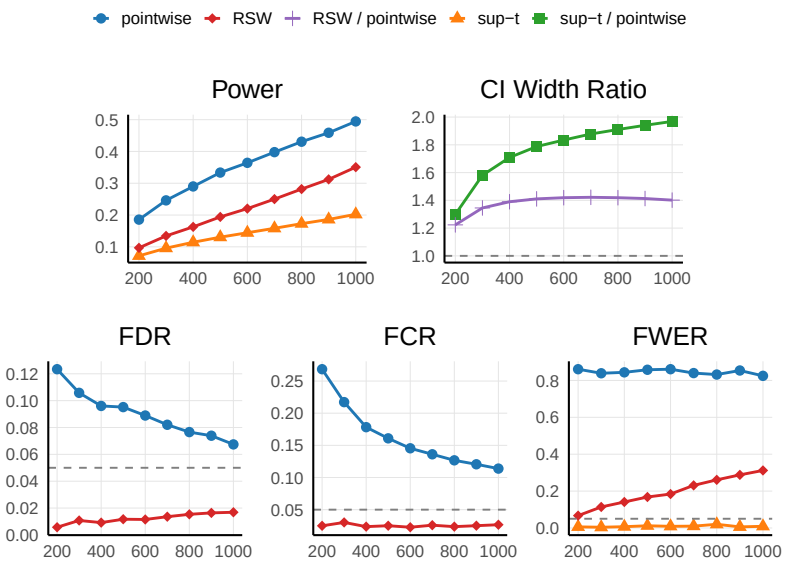


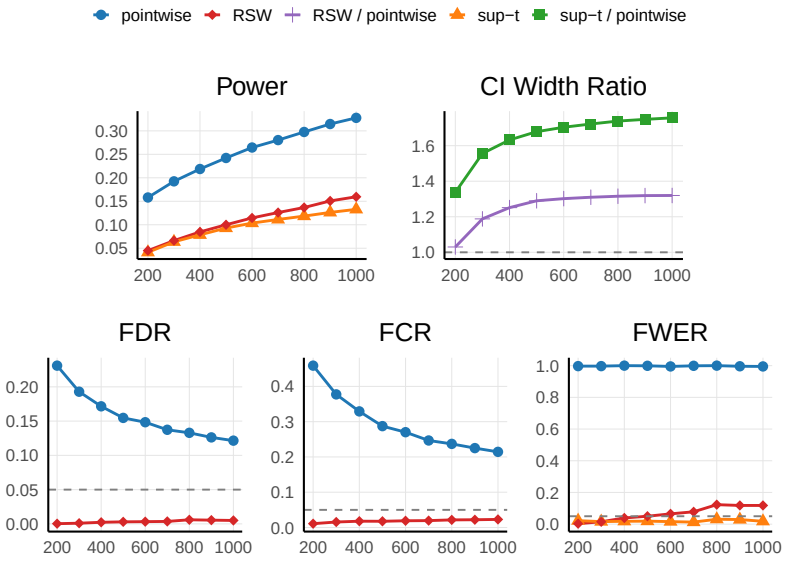
Figure B.2: True IRF grid for the persistent DGP (`dgp_persistent`).

B.1 Pointwise vs sup-t vs RSW

B.1.1 Baseline family ($H = 20$)

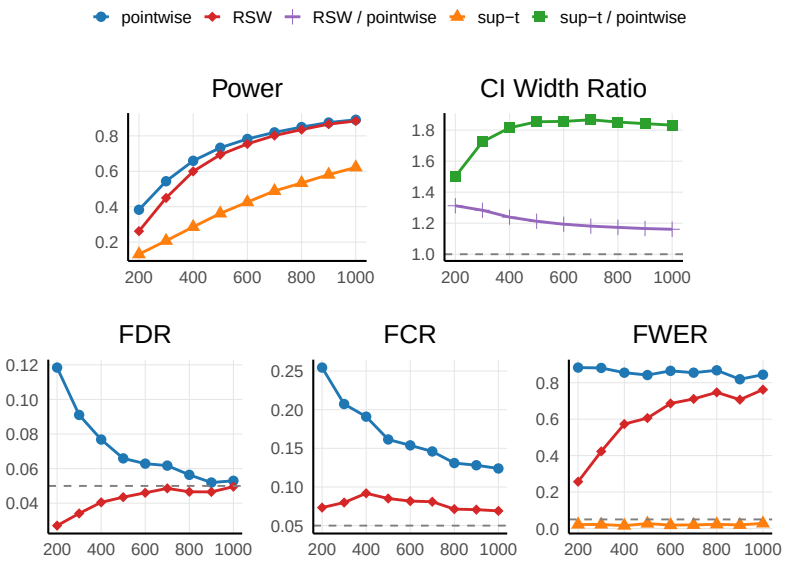


(a) VAR

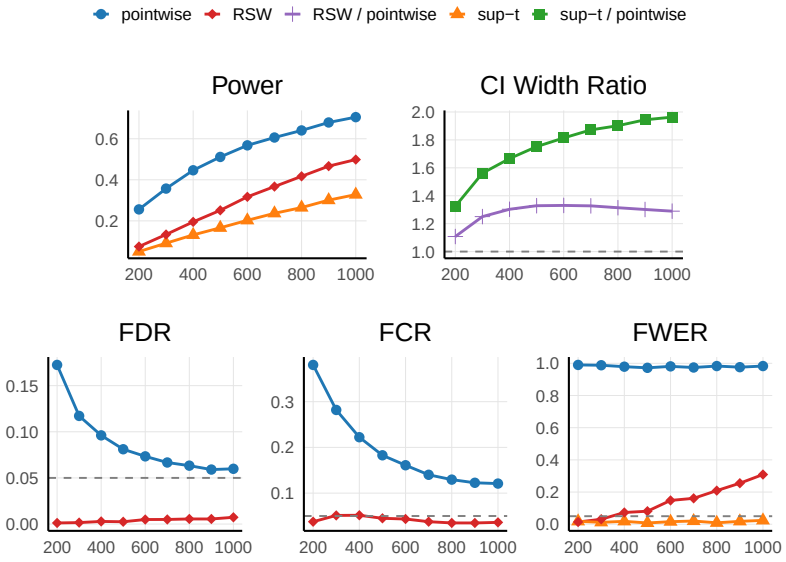


(b) LP

Figure B.3: Pointwise vs. sup-t vs. RSW, baseline DGP, baseline family ($H = 20$).



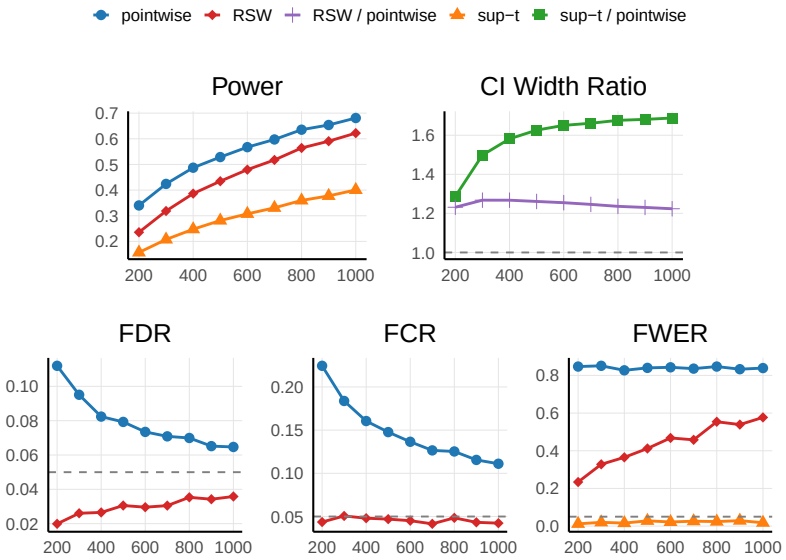
(a) VAR



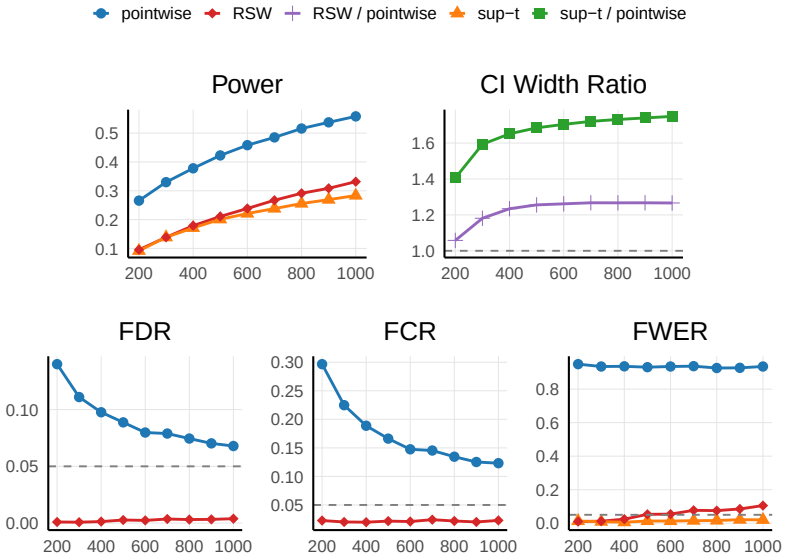
(b) LP

Figure B.4: Pointwise vs. sup-t vs. RSW, persistent DGP, baseline family ($H = 20$).

B.1.2 Fewer horizons ($H = 10$)

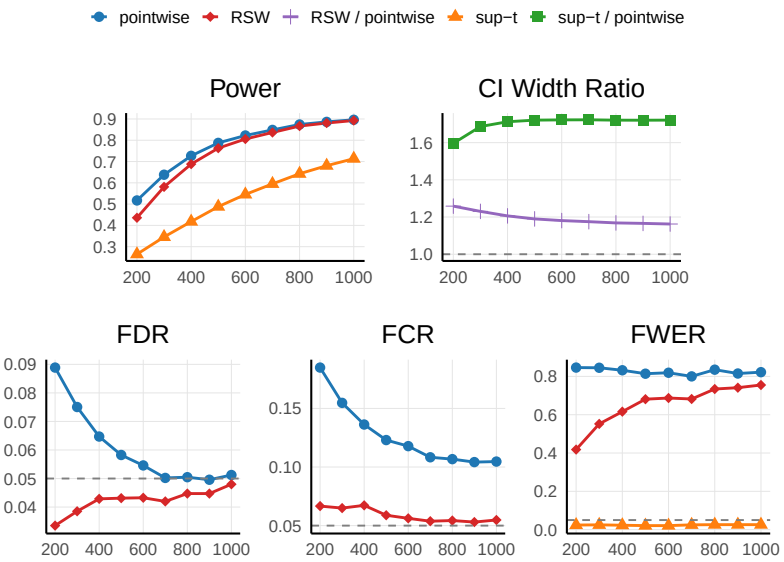


(a) VAR

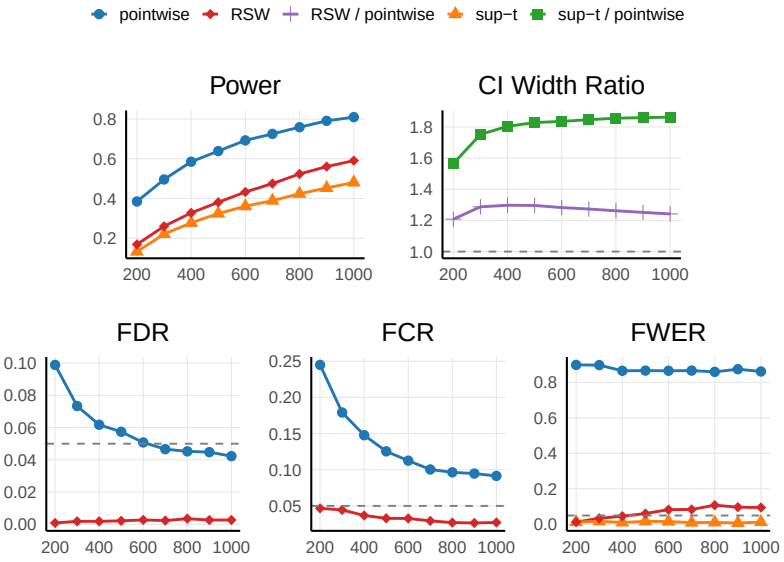


(b) LP

Figure B.5: Pointwise vs. sup-t vs. RSW, baseline DGP, fewer horizons ($H = 10$).



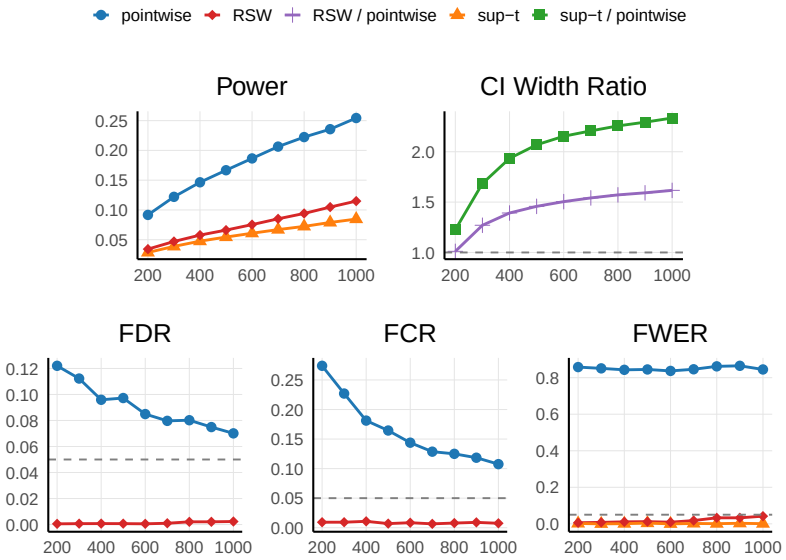
(a) VAR



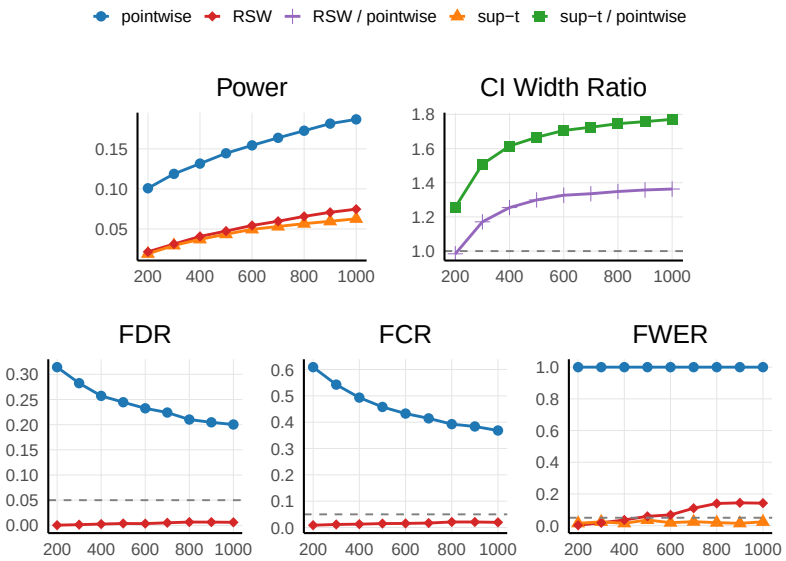
(b) LP

Figure B.6: Pointwise vs. sup-t vs. RSW, persistent DGP, fewer horizons ($H = 10$).

B.1.3 More horizons ($H = 40$)

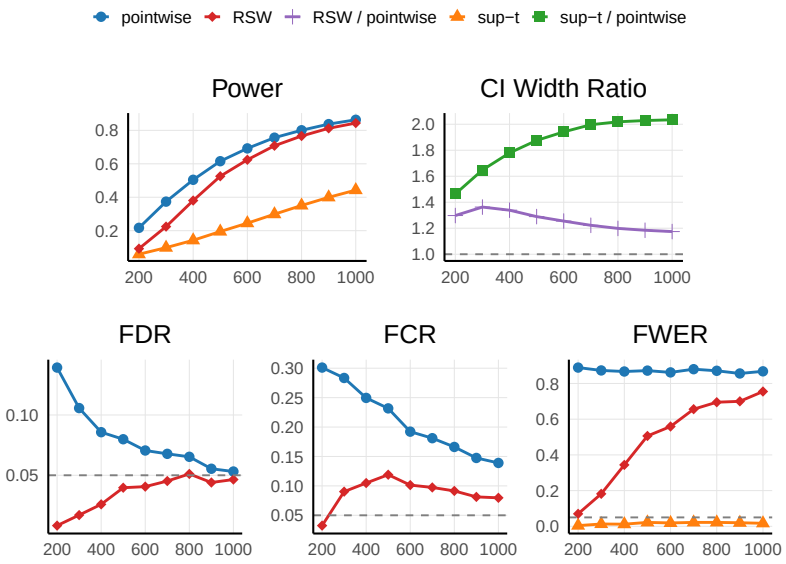


(a) VAR

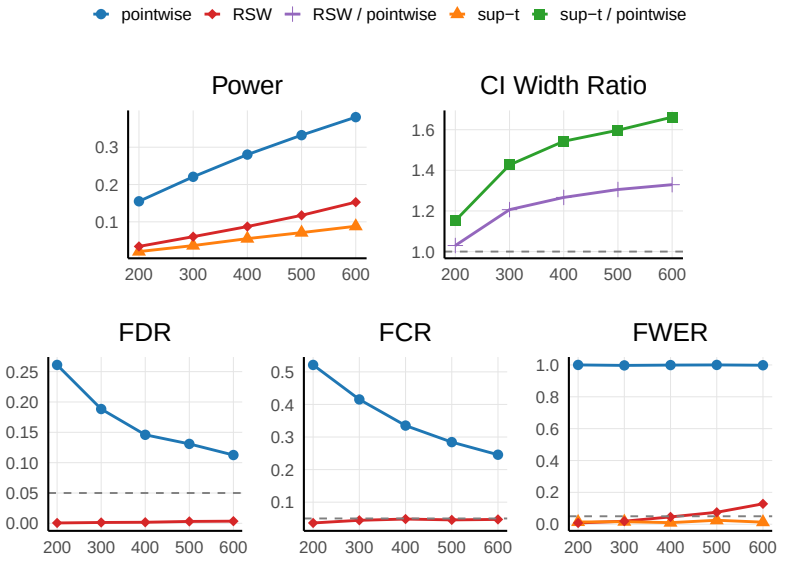


(b) LP

Figure B.7: Pointwise vs. sup-t vs. RSW, baseline DGP, more horizons ($H = 40$).



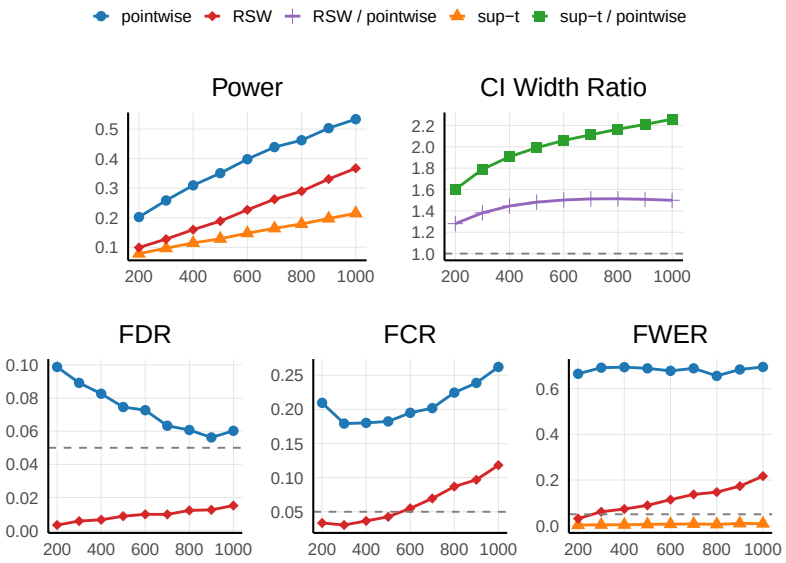
(a) VAR



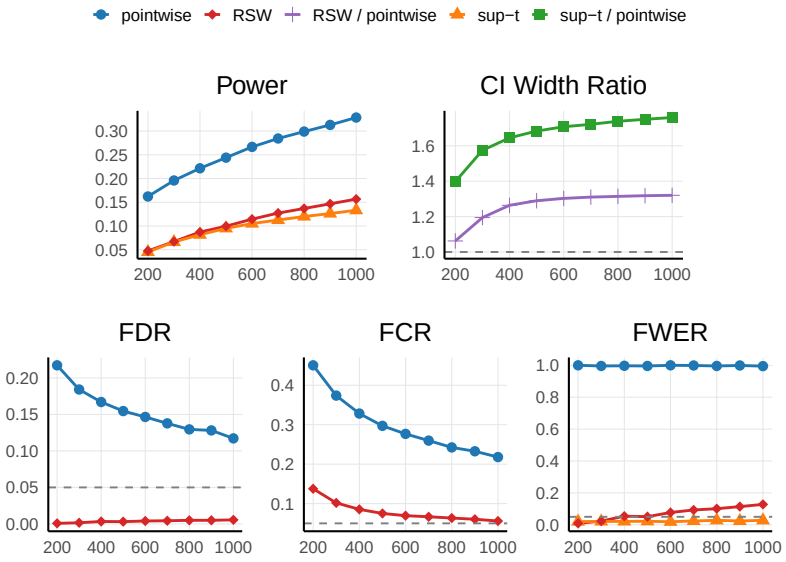
(b) LP

Figure B.8: Pointwise vs. sup-t vs. RSW, persistent DGP, more horizons ($H = 40$).

B.1.4 Lag misspecification

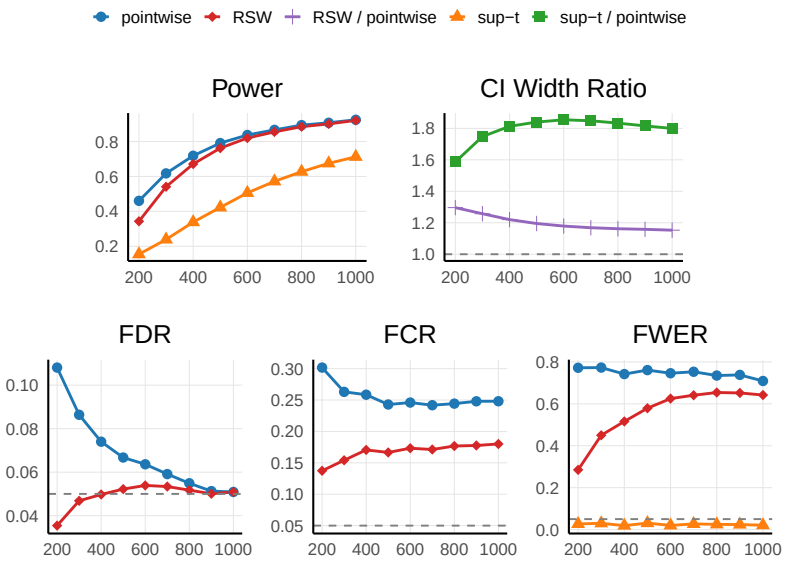


(a) VAR

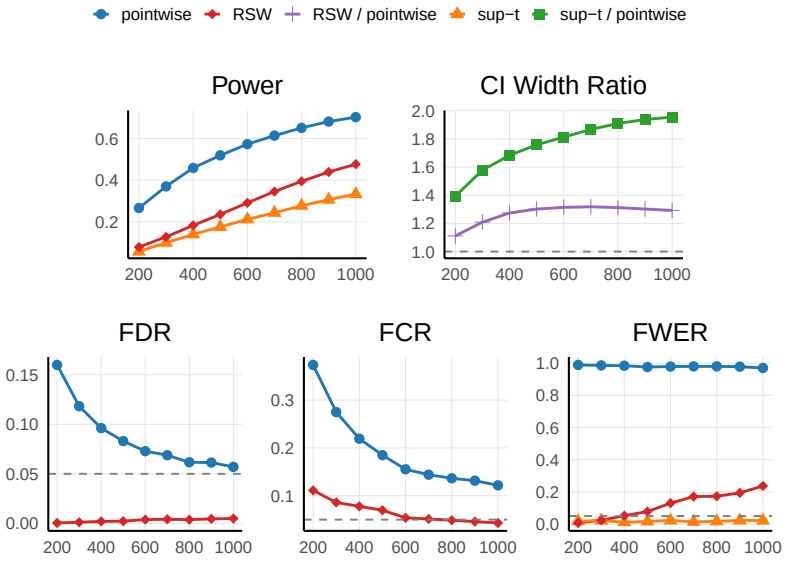


(b) LP

Figure B.9: Pointwise vs. sup-t vs. RSW, baseline DGP, lag misspecification.



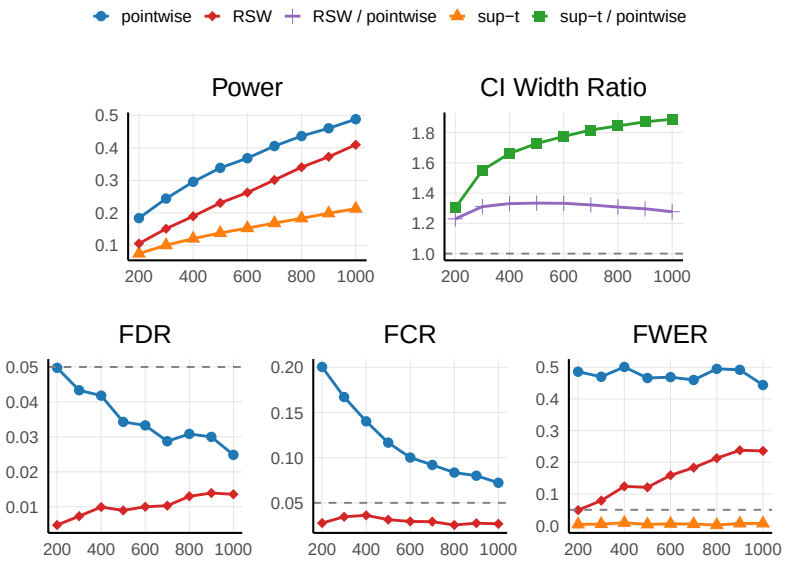
(a) VAR



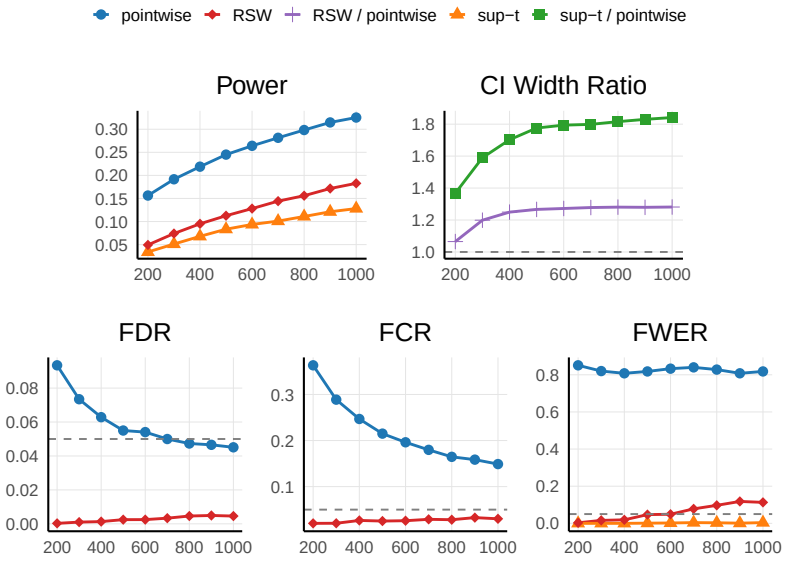
(b) LP

Figure B.10: Pointwise vs. sup-t vs. RSW, persistent DGP, lag misspecification.

B.1.5 True-null sparsity (60nulls)

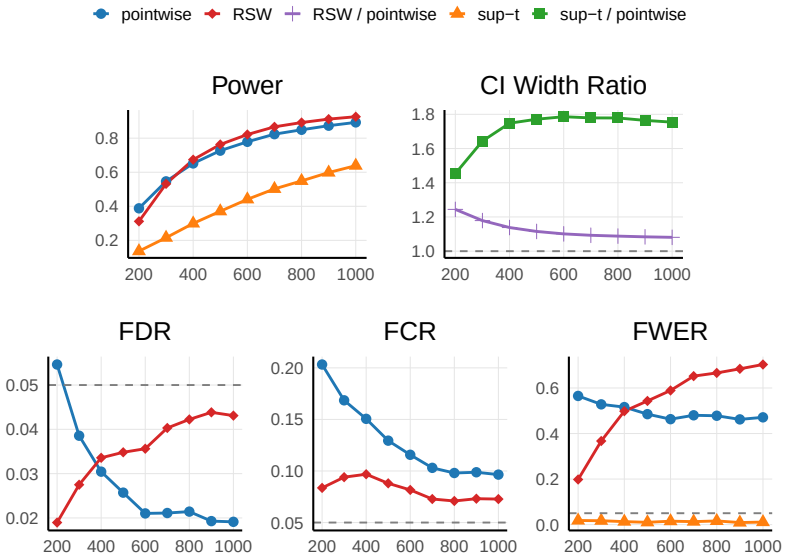


(a) VAR

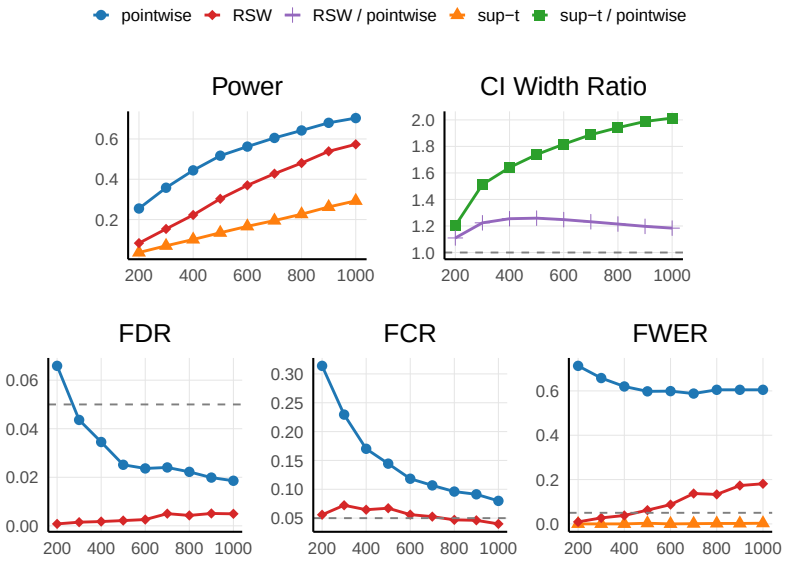


(b) LP

Figure B.11: Pointwise vs. sup-t vs. RSW, baseline DGP, true-null sparsity (60nulls).



(a) VAR



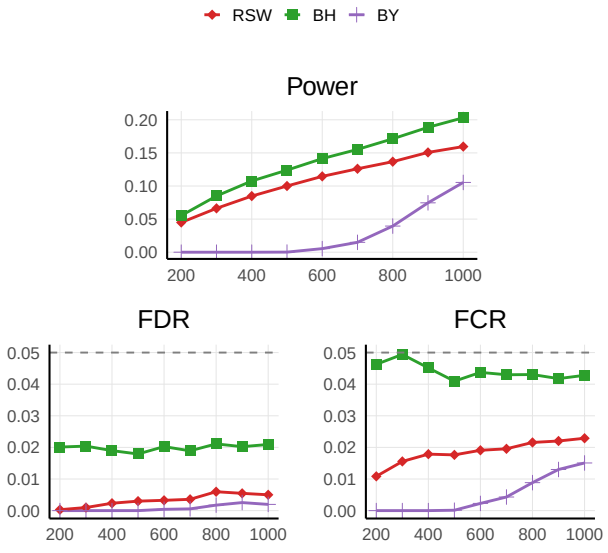
(b) LP

Figure B.12: Pointwise vs. sup-t vs. RSW, persistent DGP, true-null sparsity (60nulls).

B.2 RSW vs BH vs BY

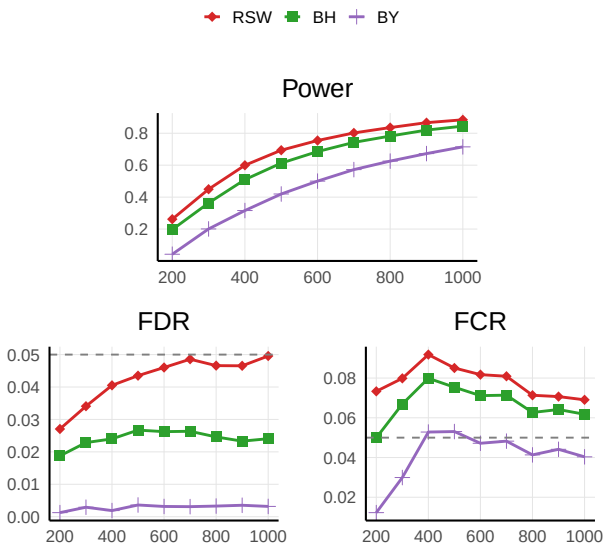
B.2.1 Baseline family ($H = 20$)

(a) VAR

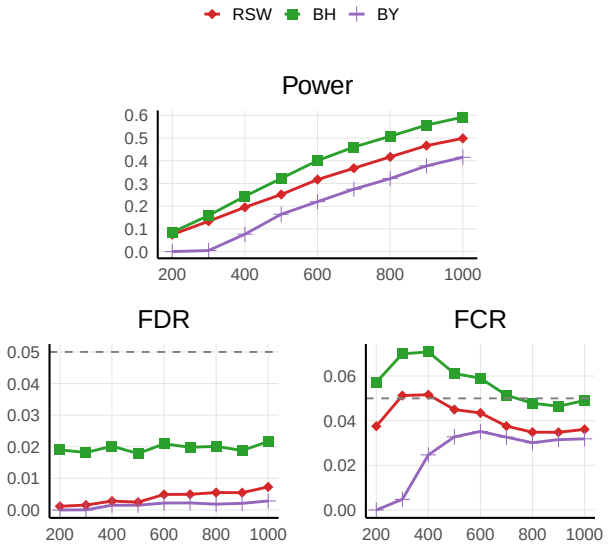


(b) LP

Figure B.13: RSW vs. BH vs. BY, baseline DGP, baseline family ($H = 20$).



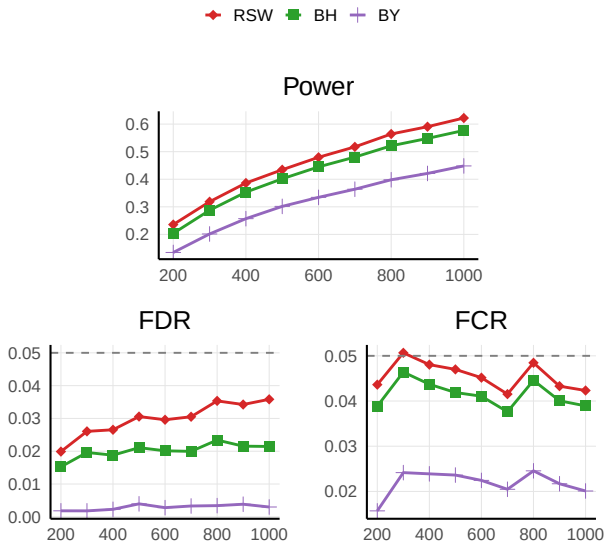
(a) VAR



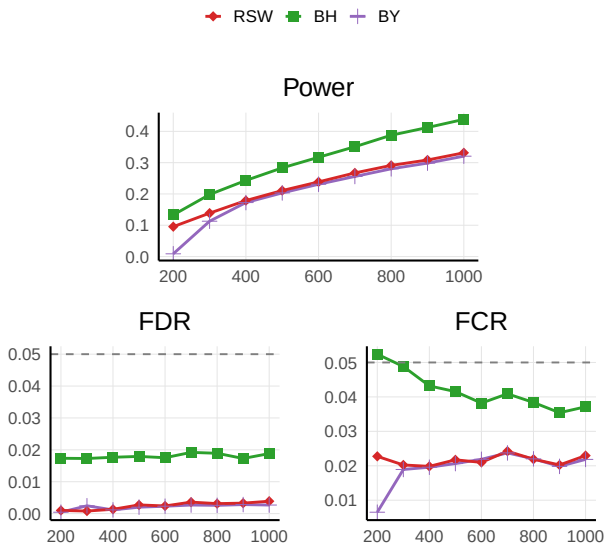
(b) LP

Figure B.14: RSW vs. BH vs. BY, persistent DGP, baseline family ($H = 20$).

B.2.2 Fewer horizons ($H = 10$)

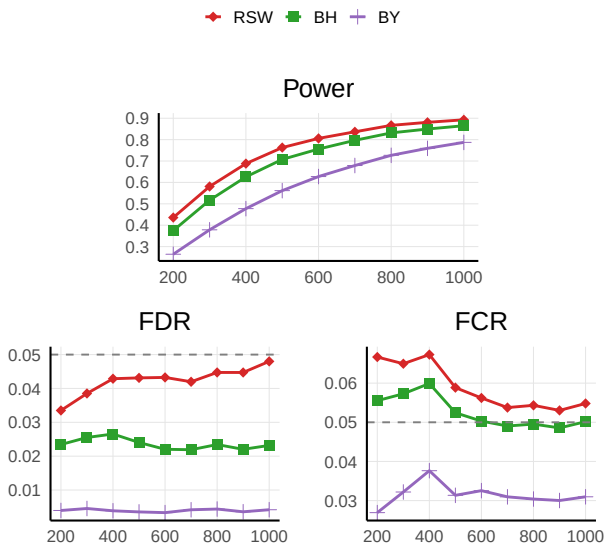


(a) VAR

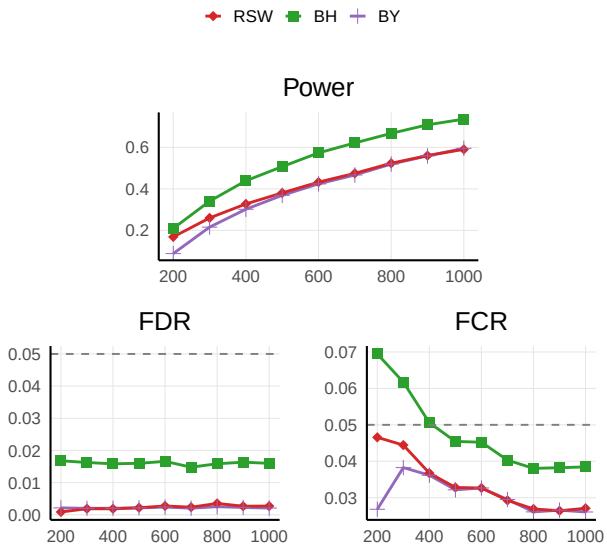


(b) LP

Figure B.15: RSW vs. BH vs. BY, baseline DGP, fewer horizons ($H = 10$).



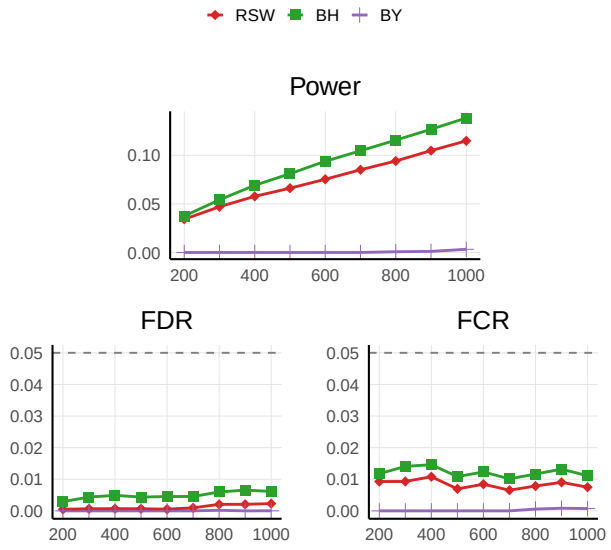
(a) VAR



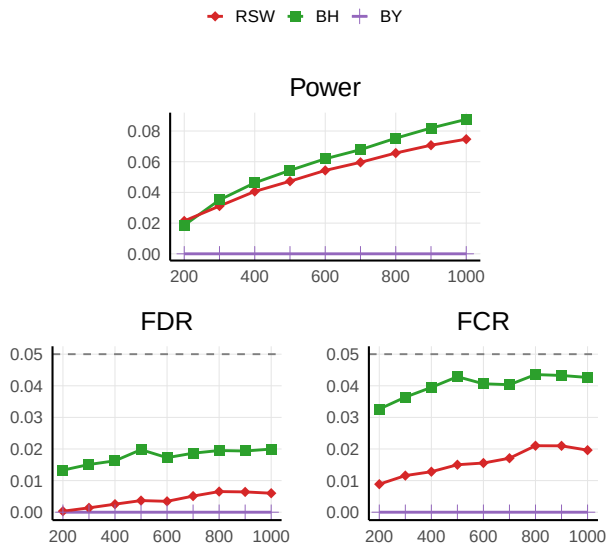
(b) LP

Figure B.16: RSW vs. BH vs. BY, persistent DGP, fewer horizons ($H = 10$).

B.2.3 More horizons ($H = 40$)

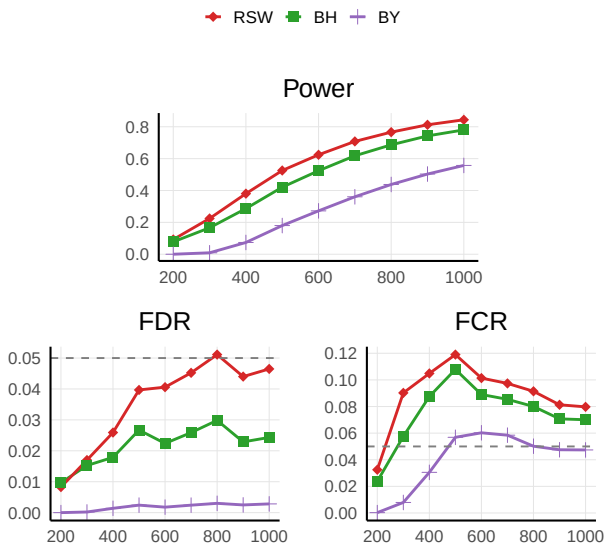


(a) VAR

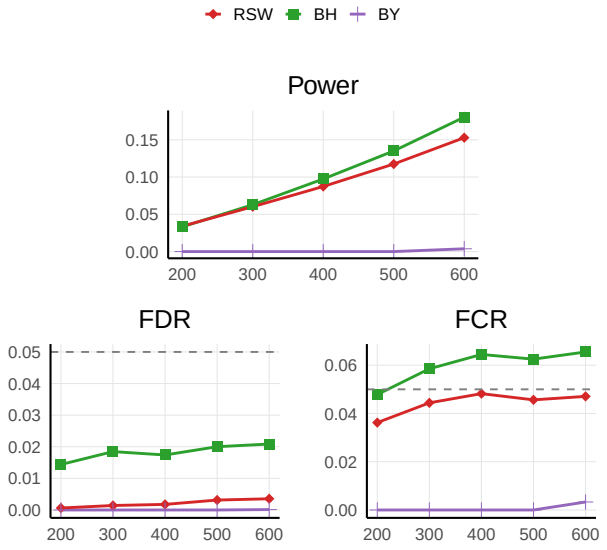


(b) LP

Figure B.17: RSW vs. BH vs. BY, baseline DGP, more horizons ($H = 40$).



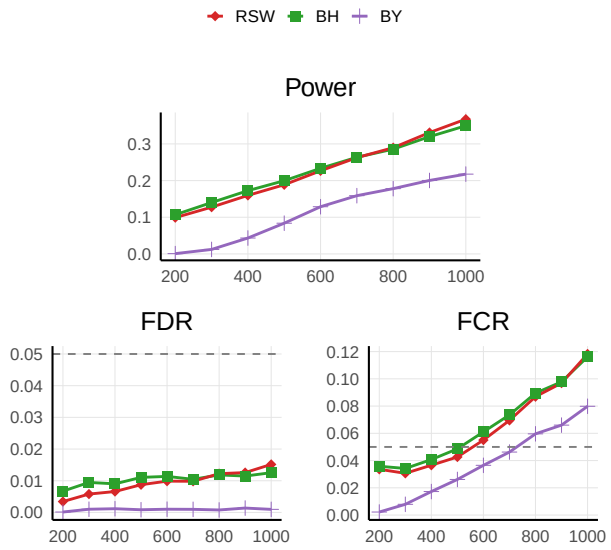
(a) VAR



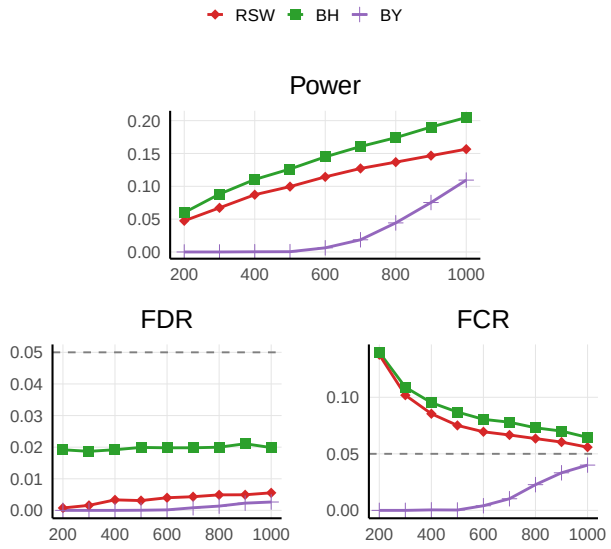
(b) LP

Figure B.18: RSW vs. BH vs. BY, persistent DGP, more horizons ($H = 40$).

B.2.4 Lag misspecification

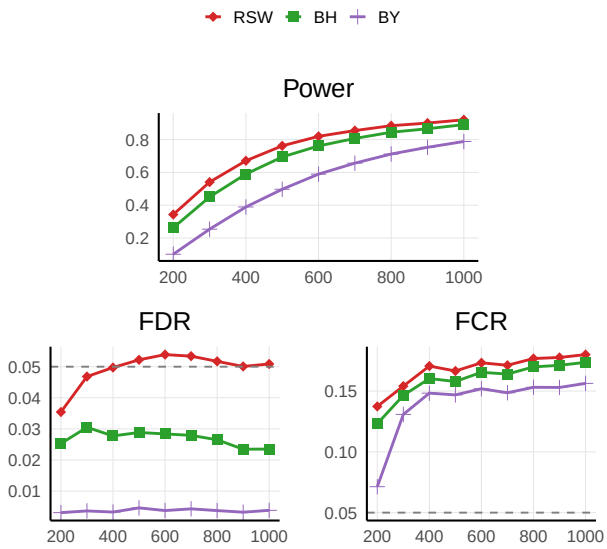


(a) VAR

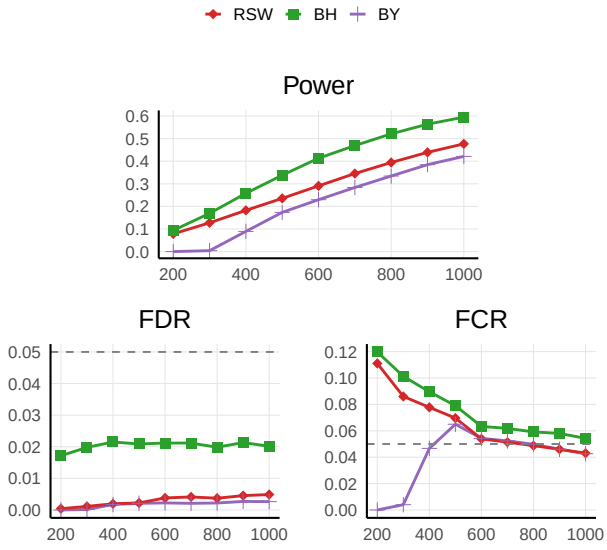


(b) LP

Figure B.19: RSW vs. BH vs. BY, baseline DGP, lag misspecification.



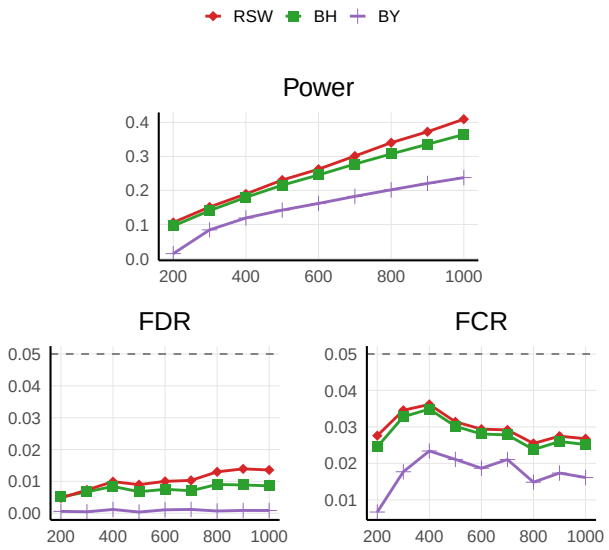
(a) VAR



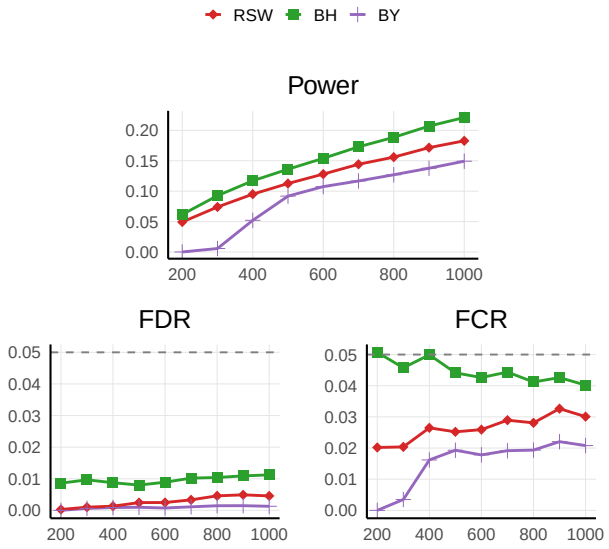
(b) LP

Figure B.20: RSW vs. BH vs. BY, persistent DGP, lag misspecification.

B.2.5 True-null sparsity (60nulls)

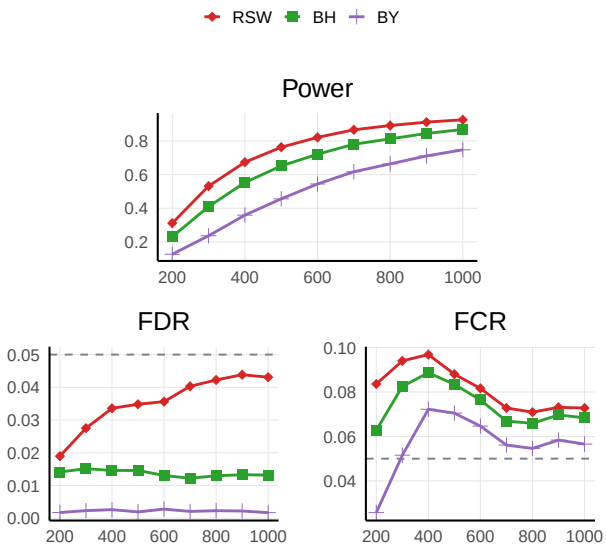


(a) VAR

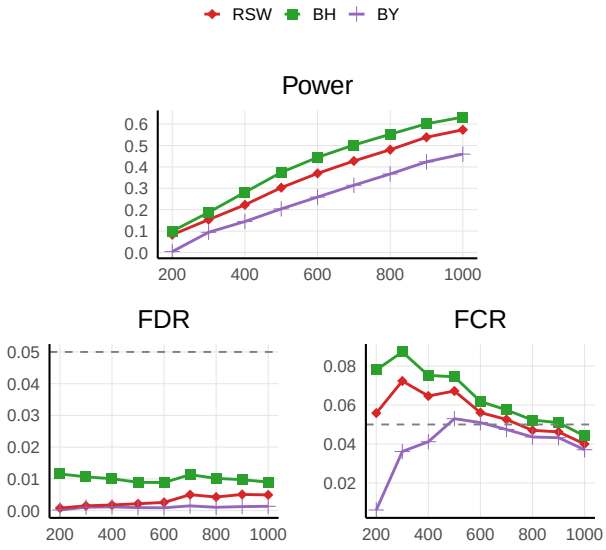


(b) LP

Figure B.21: RSW vs. BH vs. BY, baseline DGP, true-null sparsity (60nulls).



(a) VAR



(b) LP

Figure B.22: RSW vs. BH vs. BY, persistent DGP, true-null sparsity (60nulls).

C BH vs RSW: Dependence in IRF t-statistics

Impulse response function (IRF) t-statistics are typically dependent across horizons, response variables, shocks, and states. Because Benjamini–Hochberg (BH) FDR control relies on independence or on a specific positive dependence condition (PRDS) that is not generically implied by IRF dependence, BH may not be theoretically justified for IRF discovery selection. The Romano–Shaikh–Wolf (RSW) stepdown procedure is designed for dependent studentized statistics, because it calibrates critical values using the empirically estimated joint distribution of IRF t-statistics via bootstrap.

C.1 Setup and notation

We observe a time series sample of length T. All probability statements are under the true data-generating process. Throughout, "large" test statistics mean "more evidence against the null."

Let $\theta_0 \in \mathbb{R}^d$ denote the population parameter vector that is sufficient to define the IRFs of interest. Let $\hat{\theta}$ denote an estimator of θ_0 computed from the T observations. We interpret $\hat{\theta}$ broadly as a stacked estimator containing every estimated object needed to construct the reported IRF panel.

Let $\mathcal{I}$ denote the index set for response variables. Let $\mathcal{H}$ denote the index set for horizons. Let $\mathcal{J}$ denote the index set for structural shocks. Let $\mathcal{K}$ denote the index set for states or regimes. Let $a = (i, h, j, k)$ denote a generic index quadruple with $i \in \mathcal{I}$, $h \in \mathcal{H}$, $j \in \mathcal{J}$, and $k \in \mathcal{K}$. Let $\mathcal{A} \equiv \mathcal{I} \times \mathcal{H} \times \mathcal{J} \times \mathcal{K}$ denote the full index set of reported IRF objects. Let $m \equiv |\mathcal{A}|$ denote the number of simultaneous hypotheses.

For each $a \in \mathcal{A}$, let $\psi_a(\theta_0) \in \mathbb{R}$ denote the corresponding population IRF coefficient. In a state-invariant analysis, $|\mathcal{K}| = 1$ and the index k is redundant. Define the vector-valued IRF mapping $\Psi : \mathbb{R}^d \to \mathbb{R}^m$ by $\Psi(\theta) \equiv (\psi_a(\theta))_{a \in \mathcal{A}}$.

Let $\hat{\psi}_a$ denote the estimator of $\psi_a(\theta_0)$ that is reported in practice. In a VAR, $\hat{\psi}_a$ is a plug-in estimator $\psi_a(\hat{\theta})$. In a local projection (LP), $\hat{\psi}_a$ is an estimated regression coefficient. For uniformity, define the stacked estimator $\hat{\Psi} \equiv (\hat{\psi}_a)_{a \in \mathcal{A}} \in \mathbb{R}^m$. When a plug-in representation is available, $\hat{\Psi} = \Psi(\hat{\theta})$ by construction.

For each $a \in \mathcal{A}$, let $\hat{\sigma}_a$ denote a standard error estimator for $\hat{\psi}_a$. Assume $\hat{\sigma}_a > 0$ with probability tending to one. Define the studentized statistic and its absolute value by

$$T_a \equiv \frac{\hat{\psi}_a - \psi_a(\theta_0)}{\hat{\sigma}_a}, \qquad S_a \equiv |T_a|. \tag{C.1}$$

When the null is $\psi_a(\theta_0) = 0$, the statistic simplifies to $T_a = \hat{\psi}_a/\hat{\sigma}_a$. Define the associated two-sided normal-reference p-value by

$$\hat{p}_a \;\equiv\; 2\Big(1 - \Phi(S_a)\Big), \tag{C.2}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function.

For each $a \in \mathcal{A}$, define the null hypothesis and alternative hypothesis by

$$H_{0,a} : \psi_a(\theta_0) = 0, \qquad H_{1,a} : \psi_a(\theta_0) \neq 0. \tag{C.3}$$

A multiple testing procedure outputs a random empirical rejection set $\widehat{\mathcal{R}} \subseteq \mathcal{A}$. Let $\mathcal{A}_0 \subseteq \mathcal{A}$ denote the (unknown) set of indices corresponding to true null hypotheses. Define the number of rejections and false rejections by

$$R \;\equiv\; |\widehat{\mathcal{R}}|, \qquad F \;\equiv\; |\widehat{\mathcal{R}} \cap \mathcal{A}_0|. \tag{C.4}$$

Define the false discovery proportion (FDP) and false discovery rate (FDR) by

$$\mathrm{FDP} \;\equiv\; \frac{F}{\max\{R, 1\}}, \qquad \mathrm{FDR} \;\equiv\; \mathbb{E}[\mathrm{FDP}]\,. \tag{C.5}$$

A procedure controls the FDR at level $\alpha \in (0, 1)$ if $\mathrm{FDR} \leq \alpha$.

C.2 Joint asymptotics and generic dependence of IRF estimators

Assume $\hat{\theta}$ admits an asymptotic linear representation with a nondegenerate Gaussian limit. In particular, assume

$$\sqrt{T}\,(\hat{\theta} - \theta_0) \;\Rightarrow\; \mathcal{N}(0, V_\theta), \tag{C.6}$$

where $V_\theta \in \mathbb{R}^{d\times d}$ is the asymptotic covariance matrix of $\sqrt{T}(\hat{\theta} - \theta_0)$.

Assume that $\Psi(\theta)$ is differentiable in a neighborhood of θ_0. Define the Jacobian matrix $G \in \mathbb{R}^{m\times d}$ by

$$G \;\equiv\; \left.\frac{\partial \Psi(\theta)}{\partial \theta'}\right|_{\theta=\theta_0}. \tag{C.7}$$

Let $g_a' \in \mathbb{R}^{1\times d}$ denote the ath row of G.

Under the multivariate delta method,

$$\sqrt{T}\left(\Psi(\hat{\theta}) - \Psi(\theta_0)\right) \;\Rightarrow\; \mathcal{N}(0, \Sigma), \tag{C.8}$$

where

$$\Sigma \;\equiv\; G\, V_\theta\, G'. \tag{C.9}$$

Let Σ_{ab} denote the (a, b) entry of Σ for $a, b \in \mathcal{A}$.

For any two indices $a, b \in \mathcal{A}$, the leading covariance approximation implied by the delta method is

$$\mathrm{Cov}\big(\psi_a(\hat{\theta}), \psi_b(\hat{\theta})\big) \;\approx\; \frac{1}{T}\, g_a' V_\theta g_b. \tag{C.10}$$

This covariance is generically nonzero because (i) V_θ is generically non-diagonal and (ii) distinct gradients g_a and g_b load on common components of $\hat{\theta}$. The sign of $g_a' V_\theta g_b$ is not restricted in general. Therefore, the induced dependence across IRF elements is not generically "positive" in any monotone sense.

To translate this to studentized statistics, assume the marginal standard error estimators are consistent in the sense that

$$\hat{\sigma}_a^2 \;\xrightarrow{p}\; \frac{\Sigma_{aa}}{T}, \qquad a \in \mathcal{A}. \tag{C.11}$$

Then the large-sample correlation between the marginal t-statistics satisfies

$$\mathrm{Corr}(T_a, T_b) \;\approx\; \frac{\Sigma_{ab}}{\sqrt{\Sigma_{aa}\Sigma_{bb}}} \;=\; \frac{g_a' V_\theta g_b}{\sqrt{(g_a' V_\theta g_a)(g_b' V_\theta g_b)}}. \tag{C.12}$$

In particular, $(T_a)_{a \in \mathcal{A}}$ is typically a strongly dependent vector.

C.3 The source of IRF dependence in common implementations

This section links the generic covariance formula $\Sigma = GV_\theta G'$ to standard VAR and LP constructions. The goal is not to catalog special cases where dependence disappears. The goal is to emphasize that dependence is the default outcome across horizons, variables, shocks, and states.

C.3.1 VAR IRFs

Consider a stable reduced-form VAR(p) in n_y variables. Let $y_t \in \mathbb{R}^{n_y}$ satisfy

$$y_t \;=\; A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \qquad \mathbb{E}[u_t] = 0, \qquad \mathbb{E}[u_t u_t'] = \Sigma_u. \tag{C.13}$$

Let u_t be linked to structural shocks $\varepsilon_t \in \mathbb{R}^{n_\varepsilon}$ by

$$u_t \;=\; B\,\varepsilon_t, \qquad \mathbb{E}[\varepsilon_t] = 0, \qquad \mathbb{E}[\varepsilon_t \varepsilon_t'] = I_{n_\varepsilon}. \tag{C.14}$$

Let Φ_h denote the reduced-form moving-average coefficient at horizon h. Define $\Phi_0 = I_{n_y}$ and define Φ_h recursively from the VAR coefficients for $h \geq 1$.

The structural IRF of variable i at horizon h to shock j is

$$\psi_{i,h}^{(j)} \;\equiv\; e_i' \,\Phi_h\, B\, e_j, \tag{C.15}$$

where e_i and e_j denote standard basis vectors of appropriate dimensions.

All horizons share the same estimated VAR coefficients $(\hat{A}_1, \ldots, \hat{A}_p)$ through $\hat{\Phi}_h$. All response variables share the same estimated system objects. All shocks share the same estimated impact matrix $\hat{B}$ and any common identification objects used to construct it. Therefore, the full IRF panel is a nonlinear transformation of a common estimated parameter vector. This is exactly the setting in which $\Sigma = GV_\theta G'$ is generically dense, rather than diagonal.

C.3.2 Local projections across horizons

Local projections estimate separate regressions for each horizon, but separate regressions do not imply independent estimators. Fix a scalar shock regressor x_t and controls $w_t \in \mathbb{R}^q$. For each horizon $h \in \mathcal{H}$ and response variable $i \in \mathcal{I}$, consider

$$y_{i,t+h} \;=\; \alpha_{i,h} \;+\; \beta_{i,h}\, x_t \;+\; \gamma_{i,h}' w_t \;+\; \varepsilon_{i,h,t}, \tag{C.16}$$

where $\varepsilon_{i,h,t}$ is the regression disturbance. Define the LP IRF coefficient as $\psi_{i,h}^{(j)} \equiv \beta_{i,h}$ when x_t is the empirical analog of shock j.

Even if each $\hat{\beta}_{i,h}$ is computed by OLS in a horizon-specific regression, the random variables $\{y_{i,t+h}\}_t$ and $\{y_{i,t+h'}\}_t$ are overlapping leads of the same underlying time series. This overlap generally induces correlation between $\varepsilon_{i,h,t}$ and $\varepsilon_{i,h',t}$. This overlap also induces serial correlation in the score contributions used to estimate $(\beta_{i,h})_{h\in\mathcal{H}}$.

A compact representation stacks the horizon-specific score contributions. Let $g_{i,h,t} \equiv x_t \varepsilon_{i,h,t}$ denote the score contribution for $\beta_{i,h}$ at time t. Let $g_{i,t} \equiv (g_{i,h,t})_{h\in\mathcal{H}}$ denote the stacked score vector across horizons for fixed i. Under standard conditions, the joint asymptotic covariance of $(\hat{\beta}_{i,h})_{h\in\mathcal{H}}$ depends on the long-run covariance matrix of $g_{i,t}$, which is typically non-diagonal. Define the long-run covariance matrix $\Omega_i \in \mathbb{R}^{|\mathcal{H}|\times|\mathcal{H}|}$ by

$$\Omega_i \equiv \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(g_{i,t}, g_{i,t-\ell}). \tag{C.17}$$

The off-diagonal elements of Ω_i encode cross-horizon dependence through terms such as the following.

$$\sum_{\ell=-\infty}^{\infty} \mathrm{Cov}\left(x_t \varepsilon_{i,h,t}, x_{t-\ell} \varepsilon_{i,h',t-\ell}\right). \tag{C.18}$$

C.3.3 Dependence across response variables and shocks

Across response variables, dependence arises mechanically because the same shock regressor x_t and the same sample are reused across all response equations. Across response variables, dependence is amplified when the innovation processes of different $y_{i,t}$ are contemporaneously correlated. In a VAR, this enters through Σ_u and through the joint estimation error of the system parameters. In an LP system, this enters through cross-equation covariance in $(\varepsilon_{i,h,t})_{i\in\mathcal{I}}$.

Across shocks, dependence arises whenever the identification of multiple shocks uses common estimated objects. In a VAR, the estimated columns of $\hat{B}$ typically co-move because they are functions of the same reduced-form covariance estimator and any shared identifying restrictions. In an LP setting with external instruments, multiple shock series may share estimated first-stage objects. Therefore, even when shocks are orthonormal in the population, the estimated shock-specific IRFs are not generally independent.

C.3.4 State dependence

Let $s_t \in \{0, 1\}$ denote a state indicator. Consider a state-dependent LP specification with state-specific coefficients:

$$y_{i,t+h} = \alpha_{i,h} + \beta_{i,h,0}\, x_t(1 - s_t) + \beta_{i,h,1}\, x_t s_t + \gamma_{i,h}' w_t + \varepsilon_{i,h,t}. \tag{C.19}$$

The regressors $x_t(1-s_t)$ and $x_t s_t$ are mechanically linked because they partition the same x_t. The two state-specific estimators are computed on the same realized sample path and therefore inherit common randomness. The overlapping-outcome dependence across horizons remains present within each state and across states.

If the state is itself estimated, then $\hat{s}_t$ enters multiple horizons, variables, and shocks simultaneously. This introduces an additional shared estimated object and therefore an additional dependence channel.

C.4 The limitations of BH for FDR-controlled IRF discovery selection

The BH step-up procedure uses only the ordered marginal p-values. Let $\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(m)}$ denote the p-values ordered from smallest to largest. Let $(1), \ldots, (m)$ denote the corresponding ordering of indices in $\mathcal{A}$. Define the BH cutoff rank by

$$\hat{r}_{\mathrm{BH}} \equiv \max\left\{0 \leq r \leq m : \hat{p}_{(r)} \leq \frac{r}{m}\alpha\right\}, \tag{C.20}$$

with the convention that $\hat{p}_{(0)} \equiv 0$. The BH rejection set is

$$\widehat{\mathcal{R}}_{\mathrm{BH}} \equiv \{(1), \ldots, (\hat{r}_{\mathrm{BH}})\}. \tag{C.21}$$

BH controls the FDR at level α under independence of the null p-values. BH also controls the FDR under certain positive dependence conditions, including PRDS on the subset of true null hypotheses (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001).

In the present IRF setting, the issue is not that BH "fails" mechanically. The issue is that BH requires a dependence restriction that is not implied by the econometric dependence structure of IRF t-statistics.

To make the dependence condition explicit, define the vector of evidence statistics as follows.

$$S \equiv (S_a)_{a\in\mathcal{A}} \in \mathbb{R}^m. \tag{C.22}$$

A set $D \subseteq \mathbb{R}^m$ is called increasing if $x \in D$ and $y \geq x$ componentwise imply $y \in D$. The vector S is said to be *positive regression dependent on a subset* (PRDS) on $\mathcal{A}_0$ if, for every increasing set D and every $a \in \mathcal{A}_0$, the map

$$s \mapsto \mathbb{P}(S \in D \,|\, S_a = s) \tag{C.23}$$

is nondecreasing in s. Because $\hat{p}_a$ is a strictly decreasing function of S_a, one can equivalently formulate PRDS directly in terms of p-values by reversing the partial order.

The delta-method representation $\Sigma = GV_\theta G'$ implies that the dependence of $(T_a)_{a\in\mathcal{A}}$ can be strong and can involve both positive and negative correlations. Because PRDS is a global monotonicity restriction on conditional distributions, it is not generically implied by nonzero covariances. Therefore, even if a specific IRF application happened to satisfy PRDS, it is not a transparent or verifiable implication of standard VAR or LP implementations. In other words, BH's sufficient conditions for FDR control do not map cleanly to IRF practice.

A dependence-robust alternative is the Benjamini–Yekutieli (BY) modification, which replaces α by α/c_m with $c_m \equiv \sum_{\ell=1}^{m} 1/\ell$. The BY procedure controls the FDR under arbitrary dependence (Benjamini and Yekutieli, 2001). The tradeoff is that BY can be very conservative when m is large.

C.4.1 A concrete PRDS violation in VAR and LP IRF panels

PRDS is a joint distributional restriction on the full evidence vector $S = (S_a)_{a\in\mathcal{A}}$ under the true nulls. In particular, PRDS fails as soon as it fails for any finite subvector corresponding to indices in $\mathcal{A}_0$. This subsection gives an explicit three-hypothesis construction in which PRDS is violated even under an (asymptotically) Gaussian approximation.

A three-statistic Gaussian counterexample. Let U and V be independent $\mathcal{N}(0,1)$ random variables. Define three (null) studentized statistics by

$$T_1 \equiv U, \qquad T_2 \equiv \frac{U+V}{\sqrt{2}}, \qquad T_3 \equiv \frac{-U+V}{\sqrt{2}}, \tag{C.24}$$

and define evidence statistics $S_\ell \equiv |T_\ell|$ for $\ell \in \{1,2,3\}$. Consider the increasing set

$$D_c \equiv \{(s_1,s_2,s_3) \in \mathbb{R}^3_+ : s_2 > c,\ s_3 > c\}, \qquad c > 0. \tag{C.25}$$

PRDS on the (true) null corresponding to S_1 would require the map

$$s \mapsto \mathbb{P}(S \in D_c \,|\, S_1 = s) \tag{C.26}$$

to be nondecreasing in s.

To see that this monotonicity can fail, fix $c > 0$ and write $a \equiv c\sqrt{2}$. Because conditioning on $S_1 = s$ means $U = \pm s$ with equal probability, and because the event in D_c is invariant

to the sign of U, we have

$$\mathbb{P}(S \in D_c \,|\, S_1 = s) = \mathbb{P}\left(\left|\frac{s+V}{\sqrt{2}}\right| > c,\ \left|\frac{-s+V}{\sqrt{2}}\right| > c\right) \tag{C.27}$$

$$= \mathbb{P}(|V+s| > a,\ |V-s| > a) \tag{C.28}$$

$$= \mathbb{P}(V \notin [s-a, s+a] \ \cup\ [-s-a, -s+a])\,. \tag{C.29}$$

At $s = 0$, the two intervals coincide and

$$\mathbb{P}(S \in D_c \,|\, S_1 = 0) = \mathbb{P}(|V| > a) = 2\big(1 - \Phi(a)\big). \tag{C.30}$$

At $s = a$, the two intervals exactly meet at 0 and their union is $[-2a, 2a]$, so

$$\mathbb{P}(S \in D_c \,|\, S_1 = a) = \mathbb{P}(|V| > 2a) = 2\big(1 - \Phi(2a)\big) \ <\ 2\big(1 - \Phi(a)\big) = \mathbb{P}(S \in D_c \,|\, S_1 = 0)\,. \tag{C.31}$$

On the other hand, as $s \to \infty$ the union $[s-a, s+a] \cup [-s-a, -s+a]$ moves out to $\pm\infty$ and leaves the central region uncovered, implying

$$\lim_{s\to\infty}\ \mathbb{P}(S \in D_c \,|\, S_1 = s) = 1. \tag{C.32}$$

Therefore, the function $s \mapsto \mathbb{P}(S \in D_c \mid S_1 = s)$ is not nondecreasing (it falls from $s = 0$ to $s = a$ but eventually rises toward 1), which violates PRDS. This shows that PRDS can fail in a simple Gaussian setting driven purely by linear combinations of common estimated objects.

How this arises in a VAR IRF. Consider a stable VAR(1) in two variables with coefficient matrix

$$A \ \equiv\ \rho \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \qquad 0 < \rho < \frac{1}{\sqrt{2}}, \tag{C.33}$$

so that the reduced-form MA coefficient at horizon 1 is $\Phi_1 = A$. Let the structural impact vector for shock $j = 1$ be $b \equiv Be_1 = (b_1, b_2)'$. Then three IRF entries in the reported panel satisfy

$$\psi^{(1)}_{1,0} = b_1, \tag{C.34}$$

$$\psi^{(1)}_{1,1} = e_1'\Phi_1 b = \rho(b_1 + b_2), \tag{C.35}$$

$$\psi_{2,1}^{(1)} = e_2' \Phi_1 b = \rho(-b_1 + b_2). \tag{C.36}$$

Suppose these three hypotheses are true nulls (e.g. $b_1 = b_2 = 0$), and suppose $(\hat{b}_1, \hat{b}_2)$ is asymptotically normal with a nonsingular covariance matrix (as is standard when b is an estimated function of reduced-form moments and identifying objects). After studentization, the associated t-statistics are asymptotically a standardized linear transformation of $(\hat{b}_1 - b_1, \hat{b}_2 - b_2)$. In particular, in the special case where the studentized errors of $\hat{b}_1$ and $\hat{b}_2$ are asymptotically independent standard normals, the three t-statistics above converge to the (T_1, T_2, T_3) construction. Hence PRDS fails for this three-element IRF subvector, and therefore PRDS is not a generic implication of VAR IRF practice.

How this arises in an LP IRF. Consider three LP regressions at a fixed horizon h with a common shock regressor x_t and no loss of generality normalize $\mathbb{E}[x_t^2] = 1$:

$$y_{\ell,t+h} = \beta_{\ell,h}\, x_t + \varepsilon_{\ell,t}, \qquad \ell \in \{1, 2, 3\}. \tag{C.37}$$

Suppose the three null hypotheses $\beta_{\ell,h} = 0$ are true, and suppose the (time-t) disturbance vector has a simple two-factor representation

$$\varepsilon_{1,t} \equiv \eta_t, \qquad \varepsilon_{2,t} \equiv \eta_t + \xi_t, \qquad \varepsilon_{3,t} \equiv -\eta_t + \xi_t, \tag{C.38}$$

where (η_t, ξ_t, x_t) are i.i.d. over t, mean-zero, mutually independent, and (for concreteness) Gaussian with unit variances. Then the OLS score contributions are $x_t \varepsilon_{\ell,t}$, and the joint CLT implies that the studentized LP t-statistics converge to the same (T_1, T_2, T_3) construction above (with U and V arising from the partial sums of $x_t \eta_t$ and $x_t \xi_t$). Consequently, the associated evidence vector $(|T_1|, |T_2|, |T_3|)$ violates PRDS. This factor structure is economically interpretable: two responses (or two state-specific equations, or two shock series built from a shared first stage) can load with opposite signs on a common disturbance component, while also sharing additional common variation. Thus, even in a textbook LP setup with Gaussian primitives and valid (marginal) studentization, PRDS need not hold for the IRF evidence vector.

C.5 Why the RSW stepdown procedure is well-suited for IRFs

RSW construct stepdown multiple testing procedures that achieve asymptotic FDR control under dependence by calibrating critical values from a *joint* resampling distribution of the full vector of studentized statistics (Romano et al., 2008). This is a direct match to IRF panels: the empirically relevant dependence across horizons, response variables,

shocks, and states is induced by common estimated objects (e.g. shared reduced-form parameters, shared first-stage shocks, shared state classification) and by reuse of the same time-series sample. Unlike BH, which reduces the problem to marginal p-values and therefore requires additional dependence restrictions for validity, RSW explicitly uses the joint distribution of the entire evidence vector in its calibration step.

C.5.1 Stepdown structure

Let $S_{(1)} \geq \cdots \geq S_{(m)}$ denote the evidence statistics ordered from largest to smallest, with associated ordered hypotheses $H_{0,(1)}, \ldots, H_{0,(m)}$. Given critical values $\hat{c}_1 \geq \cdots \geq \hat{c}_m$, a stepdown rule rejects the most significant hypotheses until the first comparison fails. Define

$$\hat{j} \equiv \max\left\{0 \leq j \leq m : S_{(1)} \geq \hat{c}_1, \ \ldots, \ S_{(j)} \geq \hat{c}_j\right\}, \tag{C.39}$$

and the rejection set

$$\widehat{\mathcal{R}} \equiv \{(1), \ldots, (\hat{j})\}. \tag{C.40}$$

The substantive problem is therefore the construction of $(\hat{c}_j)_{j=1}^m$ in a way that controls the FDR when $(S_a)_{a\in\mathcal{A}}$ is dependent.

RSW choose $(\hat{c}_j)$ via a recursion that is defined in terms of the *joint* distribution of order statistics of a resampled evidence vector. Operationally, the recursion uses resampling probabilities of events involving the ordered resampled statistics evaluated on index sets tied to the observed ordering of S (see, e.g., their bootstrap critical values in Eq. (17) and surrounding discussion) (Romano et al., 2008). Because these probabilities depend on the full joint law of the resampled vector, the procedure incorporates cross-coordinate dependence by construction.

C.5.2 Resampling, centered studentization, and dependence preservation

Let $X_{1:T}$ denote the observed sample, and let $\widehat{P}_T$ denote an estimated distribution used to generate a resample $X^*_{1:T}$ (for time series data, $\widehat{P}_T$ should be induced by a bootstrap that preserves serial dependence). Given $X^*_{1:T}$, recompute the full IRF panel using the same estimation pipeline as in the original data:

$$\hat{\Psi}^* \equiv (\hat{\psi}^*_a)_{a\in\mathcal{A}}, \qquad \hat{\sigma}^*_a \text{ for each } a \in \mathcal{A}. \tag{C.41}$$

To approximate the joint distribution of the studentized estimation error, use centered resampled studentized statistics,

$$T_a^* \equiv \frac{\hat{\psi}_a^* - \hat{\psi}_a}{\hat{\sigma}_a^*}, \qquad S_a^* \equiv |T_a^*|, \qquad S^* \equiv (S_a^*)_{a \in \mathcal{A}}. \tag{C.42}$$

A single resample produces the entire vector S^*. Because S^* is computed jointly from one resampled dataset and through the same IRF construction map used in the original sample, it inherits (by design) the same cross-horizon, cross-variable, cross-shock, and cross-state dependence present in the empirical IRF pipeline. This is exactly the dependence information needed to calibrate critical values for a stepdown rule in a setting where marginal evidence statistics are not independent and where PRDS is not generically implied.

C.5.3 Where exchangeability enters, and what it means for IRFs

The RSW bootstrap stepdown relies on an exchangeability condition for the limiting joint distribution of the studentized statistics corresponding to *true* nulls (Romano et al., 2008). In this context, exchangeability means that (after studentization) the joint limiting law of $(T_a)_{a \in \mathcal{A}_0}$ is invariant to permutations of the indices in $\mathcal{A}_0$. Intuitively, this symmetry allows the procedure to use the *least significant* coordinates (those with small observed S_a) as an asymptotic proxy for the unknown null subset in the recursion that determines $(\hat{c}_j)$. Absent such symmetry, the recursion can implicitly overweight or underweight certain null coordinates because "small S_a" would no longer be an approximately uninformative label for "likely null" uniformly across a.

For IRF panels, exchangeability is most plausible when the analyst deliberately enforces comparability of null distributions across coordinates by (i) using a single, uniform studentization recipe across all (i, h, j, k) (same variance estimator class and tuning rules, same lag-selection conventions, same normalization of shocks), and (ii) avoiding mixtures of fundamentally different testing objects within the same FDR family (e.g. combining statistics produced by different estimators or different identification regimes in one pooled multiple-testing problem).

C.5.4 What this delivers in practice

In an IRF discovery-selection problem, the object of interest is a rejection set that is robust to the empirically unavoidable dependence induced by the IRF construction. RSW is well-suited to this task because it (a) uses a joint resampling distribution of the entire evidence

vector and therefore carries dependence into the calibration step, and (b) produces a stepdown rejection set with critical values chosen to target FDR control under dependence (Romano et al., 2008). Consequently, when the bootstrap is valid for the joint studentized IRF vector and exchangeability is a reasonable approximation for true-null coordinates, RSW provides a principled alternative to BH in settings where BH's dependence conditions are not transparently satisfied.

D Asymptotic validity of RSW discovery control for IRF inference

SECTION PRELIMINARY & INCOMPLETE.

This appendix establishes theoretical guarantees of the bootstrap stepdown procedure in Section 3 for the VAR- and LP-based IRF statistics. We examine a fixed dimension of the declared IRF family ($m < \infty$) as the sample size increases ($T \to \infty$).

D.1 Preliminaries

Denote $\theta = (\theta_1, \ldots, \theta_m)'$ the stacked IRF vector defined in Section 3, and $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)'$ its estimator computed from a sample of length T. We test the m two-sided null hypotheses $H_{0,j} : \theta_j = 0$, $j \in \{1, \ldots, m\}$. Let $\mathcal{H}_0 \subseteq \{1, \ldots, m\}$ denote the (unknown) index set of true nulls.

Define the absolute studentized test statistics used for two-sided testing each hypothesis $j \in \{1, \ldots, m\}$, where $\hat{s}_j$ is the estimated standard error of the respective estimator.

$$T_j \equiv \left| \frac{\hat{\theta}_j}{\hat{s}_j} \right| \tag{D.1}$$

Let $\hat{\mathcal{R}} \subseteq \{1, \ldots, m\}$ denote the set of rejections produced by the RSW stepdown algorithm at target level q (Section 3). Write $R \equiv |\hat{\mathcal{R}}|$ for the total number of rejections and $V \equiv |\hat{\mathcal{R}} \cap \mathcal{H}_0|$ for the number of false rejections. By definition, FDR $\equiv \mathbb{E}[V/\max\{R, 1\}]$.

Let $\hat{P}_T$ denote the bootstrap law used to generate the joint bootstrap IRF draws (residual bootstrap for VARs; application-specific bootstrap schemes for LPs). Define $\hat{\theta}^*$ the IRF estimator computed on a bootstrap sample drawn from $\hat{P}_T$, and let $\hat{s}_j^*$ be the corresponding bootstrap standard error.

Directly following Romano et al. (2008), the bootstrap stepdown theory is stated in terms of centered studentized bootstrap statistics

$$T_j^* \equiv \left| \frac{\hat{\theta}_j^* - \theta_j(\hat{P}_T)}{\hat{s}_j^*} \right|. \tag{D.2}$$

It is possible to center at $\hat{\theta}_j$ rather than $\theta_j(\hat{P}_T)$, and obtain identical asymptotic results if $\hat{\theta}_j - \theta_j(\hat{P}_T) = o_p(T^{-1/2})$. If the bootstrap distribution is constructed to be centered at $\hat{\theta}$ (as in the standard residual bootstrap where the resampled residuals are mean-zero by construction), this condition holds exactly. In cases involving bias-corrected estimators,

recentering is still required to ensure the bootstrap null aligns with the sample estimate. For further detail, see Romano et al. (2008), Remark 5.

For any $K \subseteq \{1, \ldots, m\}$ and any law P governing the data, define $J_{T,K}(P)$ as the joint distribution under P of the (signed) studentized estimation errors

$$\left(\frac{\hat{\theta}_j - \theta_j(P)}{\hat{s}_j} \right)_{j \in K}. \tag{D.3}$$

We enumerate the following assumptions for theoretical results (Romano et al., 2008).

Assumption D.1. As $T \to \infty$, $J_{T,\{1,\ldots,m\}}(P)$ weakly converges to a limit law $J_{\{1,\ldots,m\}}(P)$ such that $J_{T,\mathcal{H}_0}(P) \Rightarrow J_{\mathcal{H}_0}(P)$. Further, $J_{\mathcal{H}_0}(P)$ has continuous, one-dimensional marginal distributions and connected support.

Assumption D.2. The limiting joint distribution $J_{\mathcal{H}_0}(P)$ is exchangeable (invariant under permutations of indices in $\mathcal{H}_0$).

Let d denote any metric characterizing weak convergence on $\mathbb{R}^m$.

Assumption D.3. $\hat{P}_T$ is an estimate of P such that $d\left(J_{T,\{1,\ldots,m\}}(P), J_{T,\{1,\ldots,m\}}(\hat{P}_T)\right) \xrightarrow{p} 0$.

Assumption D.4. For every $j \in \{1, \ldots, m\}$, $\sqrt{T}\hat{s}_j \xrightarrow{p} \sigma_j$, where $\sigma_j \in (0, \infty)$.

Theorem D.1 is a direct result of Romano et al. (2008). Thus, we provide the statement without proof. See Romano et al. (2008) for further detail.

Theorem D.1. *Suppose Assumptions D.1 - D.4 are satisfied. Then, the RSW bootstrap stepdown procedure applied to* $(T_1, \ldots, T_m)$ *at target level* q *satisfies* $\limsup_{T\to\infty} \mathrm{FDR}_P \leq q$.

Lemma D.1 establishes the limiting power of the test tending to one as $T \to \infty$.

Lemma D.1. *Suppose Assumptions D.1 and D.4 are satisfied. Then, for any fixed* $\theta_j \neq 0$, $|T_j| \xrightarrow{p} \infty$.

Proof. As a consequence of Assumption D.1, $\hat{\theta}_j$ is a $\sqrt{T}$-consistent estimator and we can write $\hat{\theta}_j = \theta_j + O_p(T^{-1/2})$. Further, $\hat{s}_j = T^{-1/2}(\sigma_j + o_p(1))$ by Assumption D.4. Then,

$$|T_j| = \left| \frac{\hat{\theta}_j}{\hat{s}_j} \right| = \left| \frac{\theta_j + O(T^{-1/2})}{T^{-1/2}(\sigma_j + o_p(1))} \right| = \sqrt{T} \left| \frac{\theta_j + O(T^{-1/2})}{\sigma_j + o_p(1)} \right|.$$

By Slutsky's Theorem,

$$|T_j| = \sqrt{T}\left(\left|\frac{\theta_j}{\sigma_j}\right| + o_p(1)\right) = \sqrt{T}\left|\frac{\theta_j}{\sigma_j}\right| + o_p(\sqrt{T}).$$

As $T \to \infty$, the first term tends to infinity. □

Remaining sections of Appendix D discuss verification of Assumptions D.1 , D.3, and D.4 for VAR and LP IRF estimators. Assumption D.2 is not directly implied by generic VAR or LP primitives, but is an additional symmetry restriction on the dependence structure of the true-null studentized statistics. In the settings considered below, the limiting joint law of the true-null studentized vector is Gaussian. For a mean-zero Gaussian vector, exchangeability is equivalent to invariance of its covariance matrix under permutations of the true-null coordinates: for every permutation matrix Π acting on $\mathcal{H}_0$, $\Pi\Sigma_0\Pi' = \Sigma_0$. A simple sufficient condition is compound symmetry, $\Sigma_0 = (1-\rho)I + \rho\,\mathbf{1}\mathbf{1}'$ for some $\rho \in (-1/(|\mathcal{H}_0|-1), 1)$.

D.2 Verification for VAR impulse responses

Assumption D.5 (Stable VAR and regular identification)**.** Assume the observed K-vector $\{y_t\}_{t=1}^T$ is generated by a stable VAR(p)

$$y_t = c + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t,$$

with $\{u_t\}$ i.i.d. (or, more generally, a martingale difference sequence) satisfying $\mathbb{E}[u_t] = 0$, $\mathbb{E}\|u_t\|^{2+\delta} < \infty$ for some $\delta > 0$, and $\Sigma_u \equiv \mathbb{E}[u_t u_t']$ positive definite. Assume stability: $\det(I_K - A_1 z - \cdots - A_p z^p) \neq 0$ for all $|z| \leq 1$.

Let the structural impact matrix C be either known and fixed, or estimated by $\widehat{C} = g(\widehat{\Sigma}_u)$ where $g(\cdot)$ is continuously Fréchet differentiable at Σ_u and $C = g(\Sigma_u)$ is invertible. Let $\theta = h(\eta)$ denote the stacked finite-horizon IRF panel as a function of the reduced-form parameter vector η (VAR coefficients and any additional identification moments), and let $\widehat{\theta} = h(\widehat{\eta})$ be the plug-in estimator. Assume the horizon grid and the outcome/shock indices defining m are fixed, and that $h(\cdot)$ is continuously differentiable in a neighborhood of the truth (automatic for finite-horizon VAR IRFs under stability). Finally, assume the standard errors used in (D.1) satisfy (**??**). (**??**)

Proposition D.1 (Joint CLT for studentized VAR IRF estimation errors)**.** *Under*

Assumption D.5, the m-vector

$$\boldsymbol{Z} \equiv \left(\frac{\widehat{\theta}_1 - \theta_1}{\widehat{s}_1}, \ldots, \frac{\widehat{\theta}_m - \theta_m}{\widehat{s}_m} \right)'$$

converges in distribution to an m-variate normal law $\mathcal{N}(0, \Sigma)$ with ones on the diagonal. In particular, the limiting law has continuous one-dimensional marginals and connected support.

Proof. Under Assumption D.5, the OLS estimator $\widehat{\eta}$ of the finite-dimensional reduced-form VAR parameter vector satisfies a multivariate CLT: $\sqrt{T}(\widehat{\eta} - \eta) \Rightarrow \mathcal{N}(0, V_\eta)$ for some finite covariance matrix V_η (see, e.g., Lütkepohl (2005)). Because $\theta = h(\eta)$ is continuously differentiable and m is fixed, the multivariate delta method yields

$$\sqrt{T}(\widehat{\theta} - \theta) = Dh(\eta)\sqrt{T}(\widehat{\eta} - \eta) + o_p(1) \Rightarrow \mathcal{N}(0, V_\theta), \qquad V_\theta \equiv Dh(\eta) V_\eta Dh(\eta)'.$$

Consistent studentization (**??**) and Slutsky's theorem then imply

$$\boldsymbol{Z} = \left(\frac{\sqrt{T}(\widehat{\theta}_1 - \theta_1)}{\sigma_1}, \ldots, \frac{\sqrt{T}(\widehat{\theta}_m - \theta_m)}{\sigma_m} \right)' + o_p(1) \Rightarrow \mathcal{N}(0, \Sigma),$$

where Σ is the correlation matrix induced by V_θ after scaling by $(\sigma_1, \ldots, \sigma_m)$. A nondegenerate multivariate normal distribution has continuous marginals and connected support. □

Proposition D.2 (Joint bootstrap validity for the VAR IRF panel)**.** *Suppose Assumption D.5 holds and the bootstrap law $\widehat{P}_T$ is generated by the residual bootstrap described in Section 3: resample reduced-form residuals $\{\widehat{u}_t\}$ i.i.d. (or apply i.i.d. wild multipliers) to generate innovations $\{u_t^*\}$, simulate y_t^* recursively from a stable fitted VAR, re-estimate the VAR (and re-estimate identification objects if treated as estimated), and recompute the full IRF panel to obtain $\widehat{\theta}^*$ and (if needed) $\widehat{s}^*$. Then the bootstrap approximation condition (**??**) holds for the VAR IRF vector, i.e.,*

$$\rho\Big(J_{T,\{1,\ldots,m\}}(P),\ J_{T,\{1,\ldots,m\}}(\widehat{P}_T) \Big) \to_p 0.$$

Proof. Under Assumption D.5, the residual bootstrap for the reduced-form VAR parameters is valid: the conditional distribution of $\sqrt{T}(\widehat{\eta}^* - \widehat{\eta})$ converges (in probability) to the same Gaussian limit as the distribution of $\sqrt{T}(\widehat{\eta} - \eta)$. Because the IRF map $h(\cdot)$ is continuously differentiable and m is fixed, the bootstrap delta method yields the corresponding result for $\sqrt{T}(\widehat{\theta}^* - \theta(\widehat{P}_T))$, and consistent studentization transfers this to the studentized vector (D.3) via Slutsky and the continuous mapping theorem.

Residual bootstrap validity for stable finite-order VARs with i.i.d. innovations is standard; see, e.g., Kilian (1998) and Lütkepohl (2005). □

Corollary D.1 (FDR control for the VAR IRF implementation)**.** *If Assumption D.5, Proposition D.1, Proposition D.2, and Assumption D.2 hold, then the RSW stepdown procedure applied to the VAR IRF panel satisfies*

$$\limsup_{T\to\infty} \mathrm{FDR}_P \;\le\; q.$$

Proof. Proposition D.1 verifies Theorem D.1(i)–(ii), Assumption D.5 imposes consistent studentization (iii), Proposition D.2 verifies (iv), and Assumption D.2 is (v). Apply Theorem D.1. □

D.3 Verification for LP impulse responses

Assumption D.6 (Stationary LP design and consistent studentization)**.** Let $\{(y_t, \varepsilon_t)\}$ be strictly stationary and ergodic with weak dependence (e.g. strong mixing with summable mixing coefficients) and $\mathbb{E}\|(y_t, \varepsilon_t)\|^{2+\delta} < \infty$ for some $\delta > 0$. Fix a finite horizon set $\{0, \dots, H\}$ and a finite set of outcomes/shocks defining the m coefficients in the declared family. For each coefficient, let $\widehat{\theta}_j$ be the corresponding LP estimator based on the horizon-by-horizon regressions used in the paper (including lag augmentation where implemented).

Assume that the stacked LP estimator admits an asymptotically linear representation

$$\sqrt{T}(\widehat{\theta} - \theta) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi_t + o_p(1)$$

for some m-vector influence process $\{\psi_t\}$ that is strictly stationary with finite second moments and satisfies a multivariate CLT. Assume further that the standard errors used in (D.1) satisfy (**??**).

Proposition D.3 (Joint CLT for studentized LP IRF estimation errors)**.** *Under Assumption D.6, the studentized LP estimation error vector*

$$\boldsymbol{Z} \;\equiv\; \left(\frac{\widehat{\theta}_1 - \theta_1}{\widehat{s}_1}, \dots, \frac{\widehat{\theta}_m - \theta_m}{\widehat{s}_m} \right)'$$

converges in distribution to an m-variate normal law with continuous marginals and connected support.

Proof. By the assumed asymptotic linear representation and multivariate CLT for $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\psi_t$, we have $\sqrt{T}(\widehat{\theta}-\theta) \Rightarrow \mathcal{N}(0, V_\theta)$ for some finite covariance matrix V_θ. Consistent studentization (**??**) and Slutsky's theorem then imply $\boldsymbol{Z} \Rightarrow \mathcal{N}(0, \Sigma)$ for the corresponding correlation matrix Σ. □

Assumption D.7 (Joint bootstrap validity for the LP IRF panel)**.** Assume the bootstrap law $\widehat{P}_T$ used for LP inference yields joint conditional consistency for the signed studentized LP error vector in the sense of Romano et al. (2008), namely

$$\rho\Big(J_{T,\{1,\dots,m\}}(P),\ J_{T,\{1,\dots,m\}}(\widehat{P}_T)\Big) \to_p 0.$$

Corollary D.2 (FDR control for the LP IRF implementation)**.** *If Assumption D.6, Proposition D.3, Assumption D.7, and Assumption D.2 hold, then the RSW stepdown procedure applied to the LP IRF panel satisfies*

$$\limsup_{T\to\infty} \mathrm{FDR}_P \ \le\ q.$$

Proof. Proposition D.3 verifies Theorem D.1(i)–(ii), Assumption D.6 imposes (iii), Assumption D.7 is (iv), and Assumption D.2 is (v). Apply Theorem D.1. □

D.4 Validity of FCR-adjusted post-selection confidence intervals

This subsection justifies the post-selection confidence intervals reported in the second step of the paper's two-step procedure. The relevant error criterion is the false coverage–statement rate (FCR) of Benjamini and Yekutieli (2005), which measures the expected fraction of noncovering reported confidence intervals among all reported confidence intervals.

False coverage–statement rate. Let $\widehat{\mathcal{R}} \subseteq \{1,\dots,m\}$ denote a data-dependent set of indices for which confidence intervals will be reported. In the implementations in this paper, $\widehat{\mathcal{R}}$ is the rejection set produced in the first step (e.g., by the RSW stepdown rule), but the FCR arguments below allow $\widehat{\mathcal{R}}$ to be any measurable selection rule. Let $\widehat{R} := |\widehat{\mathcal{R}}|$. For each $j \in \widehat{\mathcal{R}}$, let $\widehat{C}_j^{\mathrm{sel}}$ denote the reported confidence interval for θ_j. Define the number of noncovering reported intervals, the false coverage–statement proportion, and the FCR as

$$U := \sum_{j\in\widehat{\mathcal{R}}} \mathbf{1}\{\theta_j \notin \widehat{C}_j^{\mathrm{sel}}\}, \tag{D.4}$$

$$\mathrm{FCP} := \frac{U}{\max(\widehat{R}, 1)}, \tag{D.5}$$

$$\mathrm{FCR}_P := \mathbb{E}_P[\mathrm{FCP}]\,. \tag{D.6}$$

A family of marginal confidence intervals. The FCR adjustment presumes access to a confidence interval construction that is (at least asymptotically) valid marginally for each coefficient at any desired level. Throughout this paper, such intervals are obtained from the bootstrap distribution of the estimator (basic bootstrap) or of a studentized estimator (percentile-t).

Assumption D.8 (Monotone marginal confidence intervals)**.** For each $j \in \{1, \ldots, m\}$ and each $\alpha \in [0, 1]$, the procedure delivers an interval $\widehat{C}_j(\alpha)$ satisfying:

(i) (Asymptotic marginal validity) For each fixed $\alpha \in [0, 1]$,

$$\sup_{1 \le j \le m} \left| \mathbb{P}_P\Big(\theta_j \in \widehat{C}_j(\alpha)\Big) - (1 - \alpha) \right| \longrightarrow 0 \qquad \text{as } T \to \infty.$$

(ii) (Monotonicity in the confidence level) If $\alpha \ge \alpha'$, then $\widehat{C}_j(\alpha) \subseteq \widehat{C}_j(\alpha')$ almost surely.

Benjamini–Yekutieli FCR adjustment. Following Definition 3 in Benjamini and Yekutieli (2005), let $W = (W_1, \ldots, W_m)$ denote the random vector on which the selection rule is based (e.g., the vector of p-values or studentized test statistics), and write $\widehat{\mathcal{R}} = S(W)$ for a measurable selection mapping S. For each j, partition W into (W_j, W_{-j}), where W_{-j} collects all components except W_j. For each selected index $j \in S(W)$, define

$$\widehat{R}_{\min,j} := \min_{w_j:\ j \in S(w_j, W_{-j})} \big|S(w_j, W_{-j})\big|\,. \tag{D.7}$$

The level-q_{FCR} FCR-adjusted selective confidence interval for θ_j is then

$$\widehat{C}_j^{\mathrm{sel}} := \widehat{C}_j\Big(\alpha_j^*\Big)\,, \qquad \alpha_j^* := q_{\mathrm{FCR}} \frac{\widehat{R}_{\min,j}}{m}, \qquad j \in \widehat{\mathcal{R}}, \tag{D.8}$$

with no interval reported when $j \notin \widehat{\mathcal{R}}$. The adjusted level α_j^* increases with the number of reported intervals and equals the Bonferroni level q_{FCR}/m when $\widehat{R}_{\min,j} = 1$.

Remark D.1 (When the common adjustment $\alpha^* = q_{\mathrm{FCR}}\widehat{R}/m$ is valid). If the selection rule is *simple* in the sense of Benjamini and Yekutieli (2005, Remark 1)—that is, for each j and each

fixed W_{-j}, the cardinality $|S(w_j, W_{-j})|$ is constant over all w_j such that $j \in S(w_j, W_{-j})$—then $\widehat{R}_{\min,j} = \widehat{R}$ for every selected j and (D.8) simplifies to the common adjustment

$$\widehat{C}_j^{\text{sel}} = \widehat{C}_j(\alpha^*), \qquad \alpha^* = q_{\text{FCR}} \frac{\widehat{R}}{m}, \qquad j \in \widehat{\mathcal{R}}.$$

FCR control: assumptions and guarantees. The next two theorems summarize the FCR guarantees established by Benjamini and Yekutieli (2005). Theorem D.2 yields exact level control under independence, while Theorem D.3 provides a dependence-robust bound and the corresponding BY correction.

Theorem D.2 (FCR control under independence (Benjamini–Yekutieli, 2005))**.** *Suppose that the components of W are independent and that Assumption D.8 holds with exact (finite-sample) marginal coverage, $\mathbb{P}_P(\theta_j \in \widehat{C}_j(\alpha)) \geq 1 - \alpha$ for all j and α. Then the FCR-adjusted selective confidence intervals in (D.8) satisfy*

$$\text{FCR}_P \;\leq\; q_{\text{FCR}}.$$

If Assumption D.8(i) holds only asymptotically, then the same conclusion holds asymptotically:

$$\limsup_{T \to \infty} \text{FCR}_P \;\leq\; q_{\text{FCR}}.$$

Proof. The finite-sample statement is Theorem 1 in Benjamini and Yekutieli (2005). If the marginal coverage property holds only up to $o(1)$ uniformly in j (Assumption D.8(i)) and m is fixed, the same argument implies $\text{FCR}_P \leq q_{\text{FCR}} + o(1)$, which yields the stated lim sup bound. □

Theorem D.3 (FCR bound under arbitrary dependence (Benjamini–Yekutieli, 2005))**.** *Assume Assumption D.8 and construct FCR-adjusted selective confidence intervals as in (D.8). Then, without any restriction on the dependence structure of W,*

$$\limsup_{T \to \infty} \text{FCR}_P \;\leq\; q_{\text{FCR}} \cdot c_m, \qquad c_m := \sum_{\ell=1}^{m} \frac{1}{\ell}.$$

Consequently, setting $q_{\text{FCR}} := q/c_m$ ensures $\limsup_{T\to\infty} \text{FCR}_P \leq q$ under arbitrary dependence.

Proof. Theorem 4 in Benjamini and Yekutieli (2005) establishes the non-asymptotic bound $\text{FCR}_P \leq q_{\text{FCR}} c_m$ under monotonicity and exact marginal coverage. Under Assumption D.8(i), the same proof yields $\text{FCR}_P \leq q_{\text{FCR}} c_m + o(1)$ for fixed m, and the stated lim sup bound follows. □

Verification of Assumption D.8 for the bootstrap IRF intervals. We now verify that the bootstrap pointwise confidence intervals used for IRFs satisfy the conditions needed for the BY adjustment, under the same regularity conditions used above for the stepdown testing step.

Proposition D.4 (Bootstrap pointwise IRF intervals are marginally valid: VAR)**.** *Under Assumption D.5, Proposition D.1, and Proposition D.2, the bootstrap percentile-t pointwise confidence intervals obtained by inverting the bootstrap distribution of the signed studentized estimation error satisfy Assumption D.8.*

Proof. Fix $j \in \{1, \dots, m\}$ and define the signed studentized estimation error $Z_{T,j} := (\hat{\theta}_j - \theta_j)/\hat{s}_j$, where $\hat{s}_j$ is the standard error used in the test statistic. Let $\hat{z}_{j,\tau}$ denote the τ quantile of the bootstrap distribution of the corresponding signed studentized bootstrap error $\tilde{Z}^*_{T,j} := (\hat{\theta}^*_j - \hat{\theta}_j)/\hat{s}_j$ conditional on the data. The percentile-t interval at nominal miscoverage α can be written as

$$\widehat{C}_j(\alpha) = \left[\hat{\theta}_j - \hat{s}_j\, \hat{z}_{j,1-\alpha/2},\ \hat{\theta}_j - \hat{s}_j\, \hat{z}_{j,\alpha/2}\right].$$

Proposition D.2 implies that the conditional bootstrap distribution function of $\tilde{Z}^*_{T,j}$ converges uniformly to the distribution function of $Z_{T,j}$ in probability, which yields $\hat{z}_{j,\tau} - z_{j,\tau} \to 0$ in probability for each continuity point τ of the limit distribution, where $z_{j,\tau}$ is the τ quantile of $Z_{T,j}$. Proposition D.1 and Assumption D.5(iii) imply that $Z_{T,j}$ converges in distribution to a continuous limit law, so the quantile map is continuous. Therefore,

$$\mathbb{P}_P\Big(\theta_j \in \widehat{C}_j(\alpha)\Big) = \mathbb{P}_P\Big(\hat{z}_{j,\alpha/2} \le Z_{T,j} \le \hat{z}_{j,1-\alpha/2}\Big) \longrightarrow 1 - \alpha.$$

Because m is fixed, the convergence is uniform in j. Monotonicity in α follows because $\tau \mapsto \hat{z}_{j,\tau}$ is nondecreasing and the interval endpoints are built from these quantiles. □

Proposition D.5 (Bootstrap pointwise IRF intervals are marginally valid: LP)**.** *Under Assumption D.6, Proposition D.3, and Assumption D.7, the bootstrap percentile-t pointwise confidence intervals satisfy Assumption D.8.*

Proof. The argument is identical to the proof of Proposition D.4, replacing Proposition D.2 by Assumption D.7 and Proposition D.1 by Proposition D.3. □

Corollary D.3 (Simultaneous FDR and FCR control for the two-step IRF procedure)**.** *Fix $q \in (0,1)$ and let $\widehat{\mathcal{R}}$ denote the rejection set produced in the first step (e.g., by the RSW stepdown rule in Section 3) at nominal FDR level q. For each $j \in \widehat{\mathcal{R}}$, report a post-selection confidence interval $\widehat{C}^{\mathrm{sel}}_j$ constructed using the BY adjustment (D.8) with some $q_{\mathrm{FCR}} \in (0,1)$. Suppose the*

conditions of Corollary D.1 hold for the VAR implementation or the conditions of Corollary D.2 hold for the LP implementation. Then

$$\limsup_{T\to\infty} \mathrm{FDR}_P \le q, \qquad \limsup_{T\to\infty} \mathrm{FCR}_P \le q_{\mathrm{FCR}} \cdot c_m, \qquad c_m = \sum_{\ell=1}^{m} \frac{1}{\ell}.$$

In particular, choosing $q_{\mathrm{FCR}} := q/c_m$ *implies that both* $\limsup_{T\to\infty} \mathrm{FDR}_P \le q$ *and* $\limsup_{T\to\infty} \mathrm{FCR}_P \le q$ *hold under arbitrary dependence.*

Proof. The FDR statement is Corollary D.1 (VAR) or Corollary D.2 (LP). Proposition D.4 (VAR) or Proposition D.5 (LP) verifies Assumption D.8 for the bootstrap pointwise confidence intervals. Apply Theorem D.3 to obtain the FCR bound under arbitrary dependence. □