Uncovering Nonlinearities with Local Projections: Online Appendix

Paul Bousquet*

July 3, 2025

Description: Many parts of the original paper which were technical in nature now appear here, and there are several references to the main body of the paper. My plan is to develop some parts into a short, standalone companion paper. Please email me with any comments.

Table of Contents

B.1 Formal Result for Disjoint Indicator Functions

First, for clarity the main environment will be restated.

Consider an arbitrary data generating process (DGP) $\psi_h : \mathbb{R} \times \mathbb{R}^L \to \mathbb{R}$ for an outcome variable Y at time $t+h$

$$Y_{t+h} = \psi_h(\varepsilon_t, \boldsymbol{S_{t+h}}) \tag{1}$$

Here, ε_t is the structural shock of interest at time t and $\boldsymbol{S_{t+h}}$ is "everything else" in the system, which could for instance include the information set at time t as well as leads and lags of ε_t (and other shocks). Following Rambachan and Shephard (2025) and Kolesár and Plagborg-Møller (2025), the working definition of a shock, with respect to a DGP of the form in (1), is that it satisfies $\varepsilon_t \perp \boldsymbol{S_{t+h}}$ $\forall h \geq 0$. In that case, note that the conditional mean $m_h(a) \equiv \mathbb{E}[\psi_h(a, \boldsymbol{S_{t+h}})|\varepsilon_t = a]$ is equal to the average structural function $\Psi_h(a) \equiv \mathbb{E}[\psi_h(a, \boldsymbol{S_{t+h}})]$.

*Department of Economics, University of Virginia, pbousquet@virginia.edu

Now we turn to the estimands of interest. For a group of N functions $\{f_i(\cdot)\}_{i=1}^N$ and control set $\boldsymbol{W_t}$, suppose we regress Y_{t+h} on $\left\{1, \{f_i(\varepsilon_t)\}_{i=1}^N, \boldsymbol{W_t}\right\}$. The specification is

$$\begin{aligned} Y_{t+h} &= \alpha + \beta_1 f_1(\varepsilon_t) + \cdots + \beta_N f_N(\varepsilon_t) + \gamma' \boldsymbol{W_t} + u_{t+h} \\ &= \alpha + \boldsymbol{\beta}' \boldsymbol{X_t} + \gamma' \boldsymbol{W_t} + u_{t+h} \end{aligned} \tag{2}$$

where $\boldsymbol{X_t}$ is a concatenation of $\{f_i(\varepsilon_t)\}_{i=1}^N$. If ε_t is a shock and continuously distributed on an interval $I \subset \mathbb{R}$, Kolesár and Plagborg-Møller (2025)'s Proposition 1 can be extended to show that

$$\beta_i = \int_I \omega_i(a) \cdot m_h'(a) \mathrm{d}a \tag{3}$$

$$\text{with} \quad \omega_i(a) = \frac{\mathrm{Cov}(\mathbf{1}_{\{a \leq \varepsilon_t\}}, X_i^\perp)}{\mathrm{Var}(X_i^\perp)} \tag{4}$$

where $X_i^\perp$ is the residual from projecting the ith element of $\boldsymbol{X_t}$ on the remaining $N-1$ elements.[1]. **Definition**. Call a collection of disjoint intervals $\{I_i\}_{i=1}^N$ a sign partition (of $\mathbb{R}$) if there exists O_0 (which we can call the center set) such that $0 \in O_0$, $O_0 \cup \left(\cup_{i=1}^N I_i\right) = \mathbb{R}$, and $O_0 \cap \left(\cup_{i=1}^N I_i\right)$ is measure zero.

Definition. Call a collection of indicator functions $\{f_i(x_t)\}_{i=1}^N$ a normalized collection on a sign partition $\{I_i\}_{i=1}^N$ if their concatenation $\boldsymbol{X}_t^f$ has full rank, $x \in I_i \Longleftrightarrow f_i(x) \neq 0$, and a normalization:

- $x < 0$ and $f_i(x) \neq 0 \Longrightarrow f_i(x) = -1$
- $x > 0$ and $f_i(x) \neq 0 \Longrightarrow f_i(x) = 1$.

Also recall the earlier notation: $f_i^\perp(\varepsilon_t)$ are the residuals in a projection of $f_i(\varepsilon_t)$ on $\{f_k(\varepsilon_t)\}_{k \neq i}^N$ and a constant.

Proposition 1. *Suppose ε_t is a continuously distributed shock on $I \subset \mathbb{R}$ and Y_{t+h} follows a data generating process of the form (1) satisfying the conditions of Kolesár and Plagborg-Møller (2025) Proposition 1. Let $m_h(a)$ be the mean of Y_{t+h} conditional on $\varepsilon_t = a$. For a normalized collection of indicator functions $\{f_i(\varepsilon_t)\}_{i=1}^N$ on sign partition $\{I_i\}_{i=1}^N$ with center set O_0, define $\{g_i(\varepsilon_t)\}_{i=1}^N$ by $g_i(x) = \alpha_i f_i(x)$, where $\alpha_i = \frac{\mathrm{Cov}(\varepsilon_t, f_i^\perp(\varepsilon_t))}{\mathrm{Var}(f_i^\perp(\varepsilon_t))}$, and let $\boldsymbol{X}_t$ be their concatenation. If we project Y_{t+h} on $\boldsymbol{X}_t$ (and a constant and control set as in (2)), then $\beta_i = \beta_j \ \forall i, j$ if $m_h(\cdot)$ is linear in ε_t.*
Let $S_{ij} = O_0 \cup I_i \cup I_j$. $\beta_i = \beta_j$ for $i \neq j$ if $m_h(\cdot)$ is linear in ε_t on $\left(\inf\{S_{ij}\}, \sup\{S_{ij}\}\right) \cap I$.

In plain terms: if the DGP is linear on the space where the weights on β_i and β_j are non-zero, then $\beta_i = \beta_j$. The statement of the result is a bit technical because of a couple subtle points. Notice that the total weight for big and small shocks of the same sign in Figure 15 is not comparable. So we might be concerned the results are distorted by a scaling issue. Of course, the functions can easily be rescaled, but this scaling is sample-dependent so in principle a more direct correction is needed. Indicator functions turn out to have a very easy correction that

[1] And a constant. Also need $\{f_i(\varepsilon_t)\}_{i=1}^N$ s.t rank condition holds

boils down to a two-stage estimator. The other piece is what regions the indicator functions can be active. Disjoint intervals are not necessary but it makes stating the result easier. Ironically, letting intervals overlap in general allows for a more targeted statement of where nonlinearities exist because the region where weight is placed actually shrinks. More discussion is in the rest of the paper and Appendix B.2, as well as a fuller proof.

To sketch out the rest of the result, it's perhaps most instructive to show why Example 1 *didn't* work, which has similar structure but two functions: $f_1(y) = y \cdot \mathbb{1}_{y<0}$ and $f_2(y) = y \cdot \mathbb{1}_{y>0}$. For the estimand on f_1, the weights follow

$$\omega_1(a) \propto \text{Cov}(\mathbf{1}_{a \leq y_t}, X_i^{\perp}), \quad \text{with } X_1^{\perp} = f_1(y) - \mathbb{E}[f_1] - \frac{\text{Cov}(f_1, f_2)}{\text{Var}(f_2)}(f_2(y) - \mathbb{E}[f_2]).$$

Even when $a > 0$, and the indicator is not active, these weights will vary significantly (and eventually turn negative) because they have a term $-\text{Cov}(\mathbf{1}_{a \leq y_t}, y_t \cdot \mathbb{1}_{y_t>0})$. But the solution is not as simple as dropping the interaction; notice in Example 6, the indicator functions used a lower bound of .01 because a collinearity problem emerges as the floor approaches 0. So if Example 1 had instead used $f_1(y) = -\mathbb{1}_{y<-b}$ and $f_2(y) = y \cdot \mathbb{1}_{y>b}$, for some small b bounded away from 0, the weights (and $X_1^{\perp}$) would not have the same problematic term because if we project f_1 on $\{1, f_2\}$, the projection constant and coefficient have the same magnitude (i.e., $X_1^{\perp} = -\mathbb{1}_{y_t<-b} - \beta(\mathbb{1}_{1_{y_t>b}} - 1)$). This is mechanical and occurs even in finite sample estimation. So the sample analog of $\text{Cov}(\mathbf{1}_{a \leq y_t}, X_1^{\perp})$ will be a sum of terms that are non-zero only if the "irrelevant" indicator $\mathbb{1}_{y>b}$ is inactive. Even on the interval $[-b, b]$, we have a guarantee of non-negative weights because $\text{Cov}(f_1, f_2) = -\mathbb{E}[f_1]\mathbb{E}[f_2] > 0$. So incredibly, these disjoint indicator functions guarantee non-negativity and relevance and the seemingly innocuous choice to interact them with the shock makes these nice properties go away. The tacit importance of the center set O_0 should also not be overlooked. This was discussed in Example 5, where the center set is essentially $(1, \infty)$, and as a result each estimand placed weight in that region. Relegating inconvenient weight to a slice around the origin allows for more targeted hypothesis testing and general interpretation.

B.2 Proof of Result

First, restating some of the key definitions.

Definition. Call a collection of disjoint intervals $\{I_i\}_{i=1}^N$ a sign partition (of $\mathbb{R}$) if there exists O_0 (which we can call the center set) such that $0 \in O_0$, $O_0 \cup \left(\cup_{i=1}^N I_i\right) = \mathbb{R}$, and $O_0 \cap \left(\cup_{i=1}^N I_i\right)$ is measure-0.

Definition. Call a collection of indicator functions $\{f_i(x_t)\}_{i=1}^N$ a normalized collection on a sign partition $\{I_i\}_{i=1}^N$ if their concatenation $\boldsymbol{X}_t^f$ has full rank, $x \in I_i \Longleftrightarrow f_i(x) \neq 0$, and a normalization:

- $x < 0$ and $f_i(x) \neq 0 \implies f_i(x) = -1$
- $x > 0$ and $f_i(x) \neq 0 \implies f_i(x) = 1$.

Also recall the earlier notation: $f_i^{\perp}(x_t)$ is the residuals in a projection of $f_i(x_t)$ on $\{f_k(x_t)\}_{k\neq i}^N$ and a constant.

The strategy of the proof will be to first show that for a normalized collection of indicator functions $\left\{f_i(x_t)\right\}_{i=1}^N$ on a sign partition $\left\{I_i\right\}_{i=1}^N$, if we project f_i on the rest of the functions (and a constant), all the projection estimands will have the same magnitude. This will allow us to show a piecewise form for $f_i^{\perp}(x_t)$ that proves the weights in the estimands on the functional regressors in a projection of Y_{t+h} on $\left\{f_i(x_t)\right\}_{i=1}^N$ (and a control set and a constant) will be non-negative. This warrants the interpretation of each as representing a positively weighted average of marginal effects. To actually compare coefficients, we need to normalize them so that the integrated weight is the same across coefficients, which thankfully is very simple. One thing important to highlight before proceeding is the "normalization" aspect of the indicator functions. If we did not have this, the correlation with x_t would naturally be negative for the indicator functions active on the negative real line.

Step 1: Uniform Magnitude in Residualization Projections

Consider a normalized collection of indicator functions $\left\{f_i(x_t)\right\}_{i=1}^N$ on a sign partition $\left\{I_i\right\}_{i=1}^N$. For a projection of f_1 ($i=1$ WLOG) on the rest of the functions (and a constant) we have

$$f_1 = b_0 + \sum_{k=2}^{N} b_{k-1} f_k$$

The constant solves $b_0 = \mathbb{E}[f_1] - \sum_{k=2}^N b_{k-1}\mathbb{E}[f_k]$. The other estimands solve $b_0\mathbb{E}[f_k] = -b_k\mathbb{E}[f_k^2]$. Using the definition of f_k (normalized indicator function), $b_{k-1} = -b_0 \text{sign}(I_k)$ for $k \geq 2$.[2] Substituting into the equation for the constant and defining $\mu_i = \mathbb{P}(x \in I_i)$ and $\mu_0 = \mathbb{P}(x \in O_0)$, we get $\text{sign}(I_1)\mu_1 = b_0(\mu_1 + \mu_0)$. Therefore

$$|b_j| = \frac{\mu_1}{\mu_1 + \mu_0} \quad (j \geq 0)$$

Note that (i) sample analogs will have this same property and (ii) center set O_0 must have positive measure in order for these projections not to be perfectly collinear (hence the full rank condition is critical).

Step 2: Form of Projection Residuals and Implications

Now switching to the general case, define $b_i^{\perp} = \frac{\mu_i}{\mu_i + \mu_0}$. We have shown that $f_i^{\perp}(x)$ is equal to $\text{sign}(I_i)b_i^{\perp}$ when $x \in I_i$, $-\text{sign}(I_i)b_i^{\perp}$ when $x \in O_0$, and 0 otherwise. So now we return to the form of the weights in (3). Again, we assume ε_t is a continuously distributed shock on $I \subset \mathbb{R}$. The weights will be non-negative if $\text{Cov}(\mathbb{1}_{\varepsilon_t \geq a}, f_i^{\perp}(\varepsilon_t)) \geq 0$. Because $f_i^{\perp}(\varepsilon_t)$ is a mean-0 function

$$\text{Cov}(\mathbb{1}_{\varepsilon_t \geq a}, f_i^{\perp}(\varepsilon_t)) = \int_a^{\infty} f_i^{\perp}(x)\text{d}F(x)$$

where $F(\cdot)$ is the distribution function of ε_t. This now illuminates the necessity of normalizing the indicator functions to not simply be binary but instead to be -1 if they are active on negative regions. The formula above

[2] $\text{sign}(I_k) \equiv \text{sign}(i_k)$ for any $i_k \in I_k$ (given our definition of normalized collection and sign partition).

shows that the weights represent the remaining mass ε_t has left on the real line (from a onward) weighted by the function's values. Because the function is mean-0, from $-\infty$ to the left endpoint of I_i, the weights are 0.

- For the case of the indicator functions relating to an interval where $\text{sign}(I_i) = -1$, as a increases from its left endpoint, the weights increase as the function has less mass remaining with negative values. Therefore the weights peak at the right endpoint, where all of the negatively-weighted mass has been shed. If this endpoint is not at the border of O_0, they will remain at this peak until a hits the left border of O_0, then they will decrease until they hit 0 at the right endpoint of O_0.
- For the other case ($\text{sign}(I_i) = 1$), the weights follow the opposite pattern. The negatively-valued parts are on O_0, so moving along the real line towards ∞ increases $\text{Cov}(\mathbb{1}_{\varepsilon_t \geq a}, f_i^{\perp}(\varepsilon_t))$ until it hits its peak at the right endpoint of O_0, and remains there until the beginning of I_i

So not only have we shown that the weights will be non-negative, we've also traversed out the values they will take along the entire support.[3] Combined with the extensions of Kolesár and Plagborg-Møller (2025) shown in Section 2, we have shown these coefficients represent positively weighted sums of average marginal effects.

Some discussion on the mechanics demonstrated above before proceeding with the proof. This underscores both the importance and the tension of the O_0 region: if we make O_0 singleton (simply 0), the function collection will not have full rank because the functions will be perfectly collinear (plug in $\mu_0 = 0$ to the earlier expressions). At the same time, the larger the O_0 region, we are increasing the areas where the estimand on f_i is putting positive weight on areas not in I_i. This motivates the generated regressors approach. As will be discussed later in this Appendix, another fix is to allow for the indicators to overlap, but this of course introduces a different kind of collinearity problem. One nice thing about the O_0 region in practice is many of these shock series have lots of zeros, which may introduce separate problems (Barnichon and Mesters, 2025) but as far as this application is concerned it's helpful because we can make O_0 small without having $\mu_0 \approx 0$.

Step 3: Re-Scaling the Functions We have shown that each β_i in a projection of Y_{t+h} on a relatively generic set of indicator function $\{f_i(x_t)\}_{i=1}^N$ will be be weighted sum of average marginal effects. However, we still have to confront a scaling problem to make comparisons between coefficients. Namely, recall from Proposition 1 that $m_h(a)$ is the expectation of Y_{t+h} conditional on $\varepsilon_t = a$. Suppose we run the aforementioned projection and compare β_1 and β_2. If $\int_I \omega_1(a)\text{d}a < 1$ and $\int_I \omega_2(a)\text{d}a = 1$, then even if $m_h(\cdot)$ is linear in ε_t, $\beta_1 \neq \beta_2$. Rather than trying to define indicator functions along equal regions of probability mass, we can instead just scale the estimands so that their integrated area is the same. It makes sense to normalize the weights so that they all integrate to 1 so we can interpret them as proper weighted averages (of marginal effects). This normalization is simple, and while it makes these new functions generated regressors, the requisite delta method correction will be negligible in practice.

[3] Note if there are gaps in the support $I \subset \mathbb{R}$, the behavior is the same just in a discontinuous fashion.

To be explicit: given the same $\{f_i(x_t)\}_{i=1}^N$, our goal is to create a new collection $\{g_i(x_t)\}_{i=1}^N$. For each g_i, we are looking for α_i such that in a projection of Y_{t+h} on $\{g_i(x_t)\}_{i=1}^N$ (and a constant and control set), the resulting estimand weights (given by (4)) on the new set of functional regressors will have the property $\int_I \omega_i^g(a)\mathrm{d}a = 1$. First, note that we are effectively creating indicator functions out of indicator functions, though in a broad sense where the outputs are a binary other than 0 and 1. So the resulting weights in these new functions will have the property $\alpha_i\omega_i^g(a) = \omega_i(a)$, where $\omega_i(a)$ are the weights in the projection using the collection $\{f_i(x_t)\}_{i=1}^N$. Integrating over both sides, the correction is simply to divide by the total weighted area from the original projection, which is given by $\frac{\text{Cov}(\varepsilon_t, f_i^\perp(\varepsilon_t))}{\text{Var}(f_i^\perp(\varepsilon_t)}$ (proof in next Appendix section), i.e., α_i are the projections coefficients from x_t on $\{f_i(x_t)\}_{i=1}^N$ and a constant. For the standard error correction, the projection estimands for g_i defined implicitly in terms of all the α_i. Because the corrections are essentially just scaling 1 estimand that is orthogonal to the others, differentiating the usual OLS form of $(X'X)^{-1}X'Y$ yields a variance correction of $\left(\frac{\partial\tilde{\beta}_i}{\partial\alpha_i}\right)^2\text{Var}(\alpha_i) = \frac{\tilde{\beta}_i^2\text{Var}(\alpha_i)}{\alpha_i^2}$, where $\tilde{\beta}_i$ is the new projection estimand for g_i, meaning for the sample analog, we simply divide the estimate for $\tilde{\beta}$ by corresponding first stage t-statistic. So another reason to not increase the number of functions from the baseline of $N = 4$ is because these corrections will become less negligible. The covariance correction is $\frac{\tilde{\beta}_i\tilde{\beta}_j}{\alpha_i\alpha_j}\text{Cov}(\alpha_i, \alpha_j)$, where $|\text{Cov}(\alpha_i, \alpha_j)|$ is actually just $\text{Var}(\alpha_0)$.

Now we are done: the estimands represent weighted averages of marginal effects. Recall the discussion from Step 2 on the areas at which functions will have weight. For functional regressor f_i, weight will be placed on $[\min\{I_i, O_0\}, \max\{I_i, O_0\}] \cap I$. So comparing two estimands for f_i, f_j mean the total area of weight covered is the same S_{ij} given in the proposition. Therefore, if $m_h(\cdot)$ is linear in ε_t on S_{ij}, then $\beta_i = \beta_j$. Next, we will discuss trying to get a "better" result because S_{ij} may be large, especially when the number of functions grows.

The Practical Drawbacks of a Stronger Result

Using the same steps, we can prove a stronger result. Suppose we drop the requirement of the sign partition that the intervals be disjoint and instead define several overlapping intervals. Since O_0 cannot be measure 0 to satisfy the definition of a normalized collection, define o_-, o_+ such that $O_0 = [o_-, o_+]$. So instead of a sign partition, we can call an overlap sign partition a collection of intervals such that each I_i satisfies $[L_i, o_-)$ for some L_i or $[o_+, R_i)$ for some R_i (in slight abuse of notation, L_i may be $-\infty$). This just creates two groups: negative and positive shock intervals. In the body of the paper, we define the indexing of the intervals so that the first member of each group corresponds to the smallest shock magnitudes and order the negative group first. Continue to assume that is the case, so that $I_1, \ldots, I_{n^-}$ are the group of negative intervals and $I_{n^-+1}, \ldots, I_N$ are the group of positive intervals. Then we for an overlapping sign partition, we have the same results in Proposition 1 but a different S_{ij} region. Other for i that relates to a beginning of a group (i.e., for $i \neq 1, n^- + 1$), the weights $\omega_i(a)$ for the β_i will non-zero for $a \in \left(I_{i-1} \cup I_i\right)$. For if $i = 1, n^- + 1$, there is non-zero weight for $a \in \left(O_0 \cup I_i\right)$. Two immediate takeaways.

First, this is incredibly ironic. The regions of overlap across functions are considerably *tighter* if we allow for the intervals themselves to overlap. For $N = 4$, the regions are essentially the same, but you can also show that the weights placed in the estimand for β_i are comparatively much smaller outside of I_i. Second, this seems to be a much better approach, taken at face value, especially for $N > 4$. Our goal would be to interpret each β_i as an estimate of average marginal effects on I_i. Because of the considerable weight placed outside of I_i in the default Proposition 1 case, this really isn't possible. If we allow for overlap, the interpretation is much more reasonable.

There is however no free lunch to this result in practice, even if the identification result is strictly more desirable. Allowing the intervals to overlap means the regressors have much more correlation between them. This will of course show up in standard errors. Further, the expansive S_{ij} may actually be a benefit once we are in realm of having a proxy for the structural shock, rather than the structural shock itself. With a proxy, we have no way to know exactly where weight is being placed. Proposition 1 shows that even in a proxy world, the region where weight is being placed will be anchored by O_0 across estimands. So in practice, we will define an O_0 in terms of functions of the shock, but the center set we are actually using with respect to the proxy is unknown. In the case of using disjoint intervals, we at least know that however the center set shifts, the weights will all shift together, which gives some regularity. These two drawbacks ultimately mean the best path forward is simply to use disjoint intervals. But if the primary intent is to get point estimates of average marginal effects on specific regions of the shocks support, it may be worth the inefficiency to use overlapping intervals.

B.3 Orthogonal Generated Regressors

Again consider the premise of a shock ε_t with functions of the shock $\{f_i(\varepsilon_t)\}_{i=1}^N$ included in a regression on Y_t. If the functions are uncorrelated and mean 0, the weight form (4) simplifies to

$$\omega_i(a) = \frac{\text{Cov}(\mathbf{1}_{\{a \leq \varepsilon_t\}}, f_i(\varepsilon_t))}{\text{Var}(f_i(\varepsilon_t))}$$

Suppose ε_t follows distribution F with support I and the collection $\{f_i\}_{i=1}^N$ corresponds to a partition $\{I_i\}_{i=1}^N$ of I. If $f_i \neq 0$ only on I_i, the weights will have no overlap – $\omega_j(a) > 0$ for only one j. A strict no overlap requirement is not one of the weight targets, but if we restrict ourselves to collections of uncorrelated mean 0 functions, it's easy to construct a collection satisfying our objectives from the ground up. First, note that for any mean 0 function

$$\text{Cov}(\mathbf{1}_{\{a \leq \varepsilon_t\}}, f_i(\varepsilon_t)) = \int_a^\infty f_i(x) \text{d}F(x).$$

The next step is to find N functions, staying within this class, producing weights that are non-negative, relevant, and hump-shaped. The expression above shows a clear route to satisfaction. WLOG, consider the interval $I_i = [0, 1]$.

For a fixed $c \in (0,1)$, ε_t has probability mass $F(c)-F(0)$ on $[0,c]$ and mass $F(1)-F(c)$ on $[c,1]$. Consider[4]

$$f_i(a) = \begin{cases} 0 & a \notin [0,1] \\ -[F(c)-F(0)]^{-1} & a \in [0,c) \\ [F(1)-F(c)]^{-1} & a \in (c,1] \end{cases}$$

This function abides by our constraint and targets:

- It's mean 0 (expected value of 0 on $[0,1]$ and it's exactly 0 everywhere else) and will inherently be uncorrelated with other functions defined the same way for all of $\{I_i\}_{i=1}^N$.
- The weights are non-negative, relevant, and hump-shaped. $\int_a^\infty f_i(x)\mathrm{d}F(x)$ is increasing initially at $a=0$ as the area with only negative values shrinks, then begins to decrease once the area with only positive values shrinks. Eventually, it hits the boundary and becomes 0.
- It can also easily be modified to be smooth or scaled so that $\int_I \omega_i(a)\mathrm{d}a = 1$.

We can also directly interpret the estimands as positively-weighted averages of marginal effects on I_i. There are some clear downsides, however. Recall that R_i denotes the region where it's permissible for weight to be placed. The weight targets in general allow for some weight overlap because we don't want to marry qualitative descriptions for the partitioning of a shock's support (e.g., "$a = .99$ is small, $a = 1.01$ is big"). In this case, $R_i = I_i$, so such paradoxes are unavoidable. The point at which weights peak must also be set explicitly. In practice, the solution is to see how sensitive results are to changes in the partitioning and peaks. A deeper problem is the distribution function is unknown. The procedure still works with the empirical CDF, but we would much rather the functions not vary with repeated sampling. With these generated regressors, there would need to be a standard error correction, outlined later in this section and in Appendix A.3, on top of the generated regressor implied by Proposition 1. The direct correction is actually marginal but the standard errors themselves are intrinsically large.

B.4 Deep Learning

To motivate the use of deep learning, we will briefly get a sense of the can of worms we are opening if we allow there to be correlation between the functions used in the regression. The $N = 2$ specification is

$$Y_{t+h} = \alpha + \beta_1 f(\varepsilon_t) + \beta_2 g(\varepsilon_t) + u_t$$

Appendix A.2 shows the integral of the weights in β_1 is proportional to

$$\mathrm{Cov}(f(\varepsilon_t), \varepsilon_t) - \frac{\mathrm{Cov}(f(\varepsilon), g(\varepsilon))}{\mathrm{Var}(g(\varepsilon))} \mathrm{Cov}(g(\varepsilon_t), \varepsilon_t)$$

The first two goals to hit target weighting are non-negative and relevant weights. Since the quantity above represents the "total weight", it's important this quantity be positive to help ensure β_1 represents a positively

[4] This has some precedent in applications of the Haar wavelet (Mallat, 1999).

weighted average of marginal effects.[5] Equally, we need the analogous expression for β_2 to be positive. The simplest path to joint satisfaction is the functions are correlated with ε yet uncorrelated with each other. As the number of functions grows, the potentially paradoxical paths become more unwieldy. For the second goal, we know from Section 2 the weights in β_1 will be large where ε_t has more density and $f_1(\varepsilon_t)$ is large (provided $\mathbb{1}_{a \leq \varepsilon_t} = 1$).

All these "steps to success" contextualize the moderate success of disjoint indicator functions for the $N = 4$ case seen in the baseline specification. The focus of this paper will be on the targeting the same 4 combinations of {big, small} and {positive, negative} along the dimensions of a shock's size and sign. Like the orthogonal regressor approach, the deep learning procedure can naturally be extended to larger collections, but the constraint sets are already difficult to manage and increasing N will become impractical much sooner. Some anecdotal evidence to this effect – in the $N = 4$ case with slight abuse of notation we have

$$Y_{t+h} = \alpha + \beta_1 f_{\text{small, neg}} + \beta_2 f_{\text{big, neg}} + \beta_3 f_{\text{small, pos}} + \beta_4 f_{\text{big, pos}} + u_t$$

For this case, one instance of training with standard normal shocks produces "small" functions resembling indicators and "big" functions that look like a ReLu. Their plots (Figure 1) roughly look like (chronologically)

$$f_1(x) = \mathbb{1}_{x>-0.5} - 1 \ \underline{\textbf{and}} \ f_2(x) = \min\{-.8x + 2, 0\}$$

$$f_3(x) = \mathbb{1}_{x>-0.1} - .1 \ \underline{\textbf{and}} \ f_4(x) = \max\{0, .8x - 2\}$$

However, actually using these functions fails spectacularly; notice the approximations for f_1 and f_3 are highly collinear. It turns out the neural network introduces lots of slight idiosyncrasies to slither through the monstrous constraint set. So the complexity cost for expanding beyond $N = 4$ may not be worth the added specificity.

Deep learning carries a stigma of being opaque, but in this case neural network training is perfectly analogous to generic minimization routines in your programming language of choice. The modal minimization application is to find a vector $\boldsymbol{x} \in \mathbb{R}^k$ that minimizes $F(\boldsymbol{x})$. The only difference here is the search is over a space of functions, rather than a subset of the real numbers, and the space of functions that can be approximated by neural networks is vast. Again, turning to deep learning is even more natural because we are more precisely looking for a collection of functions with complicated dependencies. To search effectively in such a setting, a minimizer must jump through lots of "hoops" in order to even take a step, meaning the extensive parameterization endemic to deep learning is likely a necessary condition for this to even be a feasible venture.

In principle, a deep learning algorithm for the objectives (weighting targets) described at the beginning of this section is simple. Each iteration of training (epoch) will generate a candidate collection of functions $\left\{f_i(\cdot)\right\}_{i=1}^4$.

[5] Though recall this is not a sufficient condition on its own, as many of the examples in Section 2 show.

Given a sample for a shock $\{\varepsilon_t\}_{t=0}^T$, this yields a set of weights defined by sample analogs of (4). The candidate collection will be evaluated by a loss function which penalizes instances where weighting targets are not being hit. For example, a penalty will be incurred if there is negative weight, if there is weight where there definitively shouldn't be, and if the weight functions are not initially increasing. There are a myriad of implementation flavors for actually encoding this algorithm, which are discussed in more detail in Online Appendix C.2. One stumbling block arising from the complicated nature of the problem is approaches that are functionally equivalent (e.g., different ways of estimating LP) can have very different complexity and convergence properties. The basic strategy I've found most effective is to train with relatively few epochs, see what aspects of target weighting are being violated most intensely, adjust the penalty weights for those components, and start again. The goal here is not really about getting the loss value within a tolerance threshold, but rather to plot the weights after training and be happy with the allocations (Kolesár and Plagborg-Møller, 2025).

B.5 Standard Errors for Generated Orthogonal Regressors

When using the generated orthogonal regressor approach, one must specify intervals $\{I_i\}_{i=1}^N$ and a collection of points $\{c_i\}_{i=1}^N$ within each interval where the weights will peak. Here, I only focus on the case where we set c_i equal to the median of the interval I_i (this is what was used for the applications in the paper). The full derivations can be found in Online Appendix C.3, as well as derivations for the case that was initially presented in Appendix B.3 that defined the function in terms of the Empirical CDF.

To be explicit, define $I_i = [L_i, R_i)$, where L_i may be $-\infty$ in slight abuse of notation. Here, we are thinking about a case where we have a time series for a shock (or a proxy) $\left\{\varepsilon_t\right\}_{t=0}^T$. When we set c_i equal to the median, our functions in the basic case where they are piecewise-linear are

$$f_i(x) = \begin{cases} 0 & \text{if } x \notin I_i \\ \frac{-k_i}{n_i - k_i} & \text{if } x \in [L_i, c_i) \\ 1 & \text{if } x \in [c_i, R_i) \end{cases}$$

where n_i, k_i are the number of observations where $\varepsilon_t \in I_i$ and $\varepsilon_t \in [L_i, c_i)$, respectively. This definition ensures that the function will be hump shaped and place weight only within I_i.[6] The necessary delta error adjustment turns out to be simple. The potential complications relating to the probability density at c_i are neutralized the term appears in both the variance of c_i as well as $\frac{\partial \beta_i}{\partial c_i}$. The cancellation allows the correction term to simplify to $\widehat{\beta}_i^2 / n_i$.

[6] The same adjustment described in Online Appendix B.1 can be performed to let the weights integrate to 1

B.6 Illustrations of Functions and Their Weights

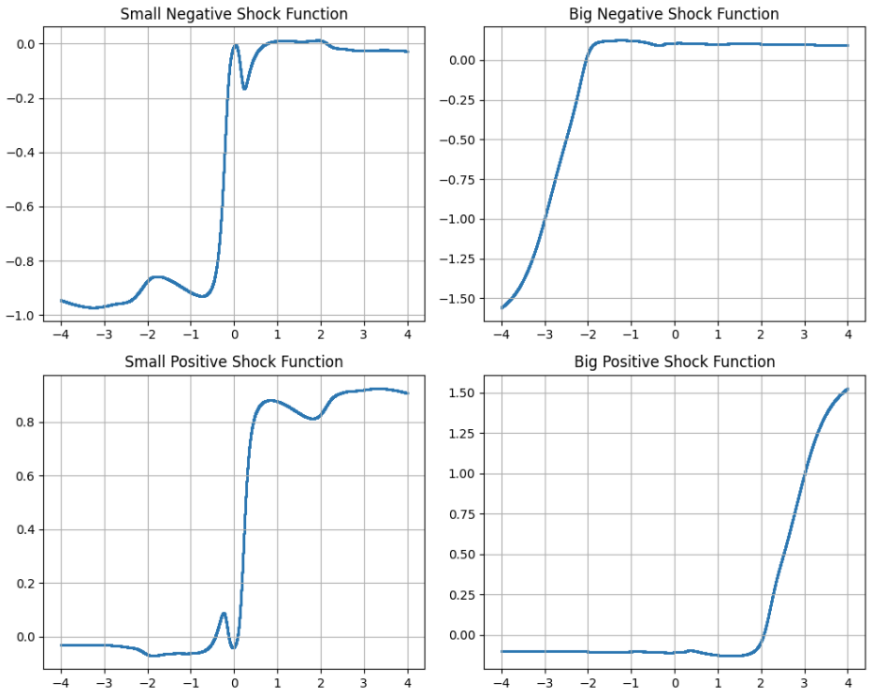

Figure 1: Neural Network Output with Standard Normal Shocks

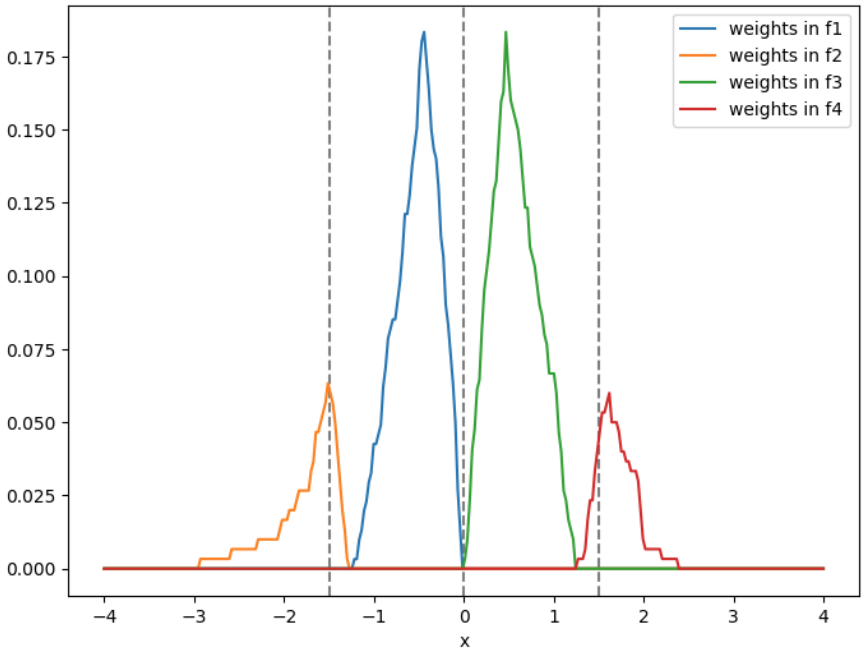

Figure 2: Generated Regressor Weights, Standard Normal Shocks

C.1 Full Expansion of FWL

Recall the general form from Section 2 of the paper

$$\omega_i = \frac{\text{cov}(1_{a \leq \varepsilon_t}, X_i^{\perp})}{\text{var}(X_i^{\perp})}$$

where $X_i^{\perp}$ is the residual from regressing X_i on the other elements in $\boldsymbol{X}_t$. We can unpack this definition to get things soley in terms of covariances and variance of terms of $\boldsymbol{X}_t$, which amounts to an expansion of the FWL theorem. To my knowledge, this expansion has not been done previously and for good reason – the full form amounts to several messy recursions that offer absolutely no insight to write out. However to motivate the use of deep learning to address one of the central issues in this paper, it may be useful to see why it's difficult to conjure up functional forms that will produce appropriate weighting.

For what follows, consider $\boldsymbol{X}$ to be a generic matrix of N covariates in a regression (which can include a vector of 1s) and X_i to be its i-th element. Keeping with notation from earlier, $X_i^{\perp}$ is the residual from X_i on the remaining elements of $\boldsymbol{X}$. WLOG, we will initially look at an example where $i = 1$. Further consider $X_n^{\perp_1}$ to be regressing the n-th element of $\boldsymbol{X}$ on its the remaining parts excluding X_1. Then

$$X_1^{\perp} = X_1 - \sum_{n=2}^{N} \frac{\text{cov}(X_1, X_n^{\perp_1})}{\text{var}(X_n^{\perp_1})} X_n$$

We can keep unpacking these terms but it should be clear that indexing is quickly going to become a nightmare because the "exclusions" will not be in a consistent ordering across the components (and sub-components, and sub-sub-components,...) of this summation. Things would have already got a bit messy notation wise had we done a formula for a generic $X_i^{\perp}$. So we will have to break this up into several parts.

First, we will deal with the covariance terms and keep the variance terms fixed. Again using $i = 1$ for indexing coherence and only focusing on the first term term in the sum above ($n = 2$), if we unpack a bit more we will have

$$\frac{\text{cov}(X_1, X_n^{\perp_1})}{\text{var}(X_n^{\perp_1})} = \frac{\text{cov}(X_1, X_2) - \frac{\text{cov}(X_2, X_3^{\perp_{1:2}})}{\text{var}(X_3^{\perp_{1:2}})} \text{cov}(X_1, X_3) + \dots}{\text{var}(X_2^{\perp_1})}$$

where $X_3^{\perp_{1:2}}$ are the residuals of regressing X_3 on the remaining elements of $\boldsymbol{X}$ excluding X_1, X_2. To even coherently define the remaining terms in the numerator, new notation has to be introduced to deal with the order in which variables are excluded from the "sub-regressions". To address this, we will define things in chunks. Again keeping with the $i = 1$ and $n = 2$ case because it's the cleanest, note that

$$\frac{\text{cov}(X_1, X_2^{\perp_1})}{\text{var}(X_2^{\perp_1})} = \frac{C_{1,2} - C_{1,3}\left(\frac{C_{2,3}}{V_{2,3}} - C_{2,4}\left(\frac{C_{3,4}}{V_{3,4}V_{2,3}} - C_{3,5}(\dots) - \dots\right) - \dots\right) + \dots}{V_{1,2}}$$

where $C_{p,q}$ represents the covariance between elements p and q of $\boldsymbol{X}$ and $V_{p,q}$ is $\text{var}(X_q^{\perp_{1:p}})$. Ignoring the unavoidably

ugly denoting of what terms correspond to, we can see a light of coherency at the end of the recursion tunnel. While there may now be a decipherable pattern to latch onto, this hasn't made the cases that are not $i = 1$ and $n = 2$ any less difficult to notate. So we will define a function relative to indexing. First observe

$$C_{1,3}\left(\frac{C_{2,3}}{V_{2,3}} - C_{2,4}\left(\frac{C_{3,4}}{V_{3,4}V_{2,3}} - C_{3,5}(\dots)\right)\right) = \sum_{k=2}^{N-1}(-1)^k \frac{C_{k,k+1}\prod_{j=1}^{k-1} C_{j,j+2}}{\prod_{j=2}^{k} V_{j,j+1}}$$

Everything else follows this structure, conditional on indexing. We can make an (ugly) generalization as follows. Let A be generic rearrangement of $\boldsymbol{X}$; i.e., $\boldsymbol{X}$ has indexing $\{1,\dots,N\}$ and A's indexing can be any permutation of this order. Define $I_k^{\{A\}}$ as the index of $\boldsymbol{X}$ corresponding to the k-th element of A (formally: a mapping $I(k;\{A\})$). Now define

$$S(\{A\}) = \sum_{k=2}^{N-1}(-1)^k \frac{C_{I_k^{\{A\}},I_{k+1}^{\{A\}}}\prod_{j=1}^{k-1} C_{I_j^{\{A\}},I_{j+2}^{\{A\}}}}{\prod_{j=2}^{k} V_{j,j+1}^{\{A\}}} \quad \text{with } V_{j,j+1}^{\{A\}} = \text{var}\left(X_{I_{j+1}^{\{A\}}}^{\perp_{I(1:j;\{A\})}}\right)$$

This will allow for a crawl towards completeness, burying as much of the index stumbling blocks as possible. Let P_N denote all permutations of $\{1,2,\dots,N\}$. Define $P_N^{i,n} \subseteq P_N$ as permutations with i,n as the first elements:

$$P_N^{i,n} = \left\{\sigma \in P_N : \sigma(1) = i \;\&\; \sigma(2) = n\right\}$$

Then we can write the earlier expression $\text{cov}(X_1, X_2^{\perp_1})$ in the general case as

$$\text{cov}(X_i, X_n^{\perp_i}) = C_{i,n} - \boldsymbol{\Sigma}_{i,n} \quad \text{with } \boldsymbol{\Sigma}_{i,n} = \sum_{\sigma\in P_N^{i,n}} S(\{\sigma\})$$

At the very beginning we started with

$$\omega_i = \frac{\text{cov}(1_{a\le\varepsilon_t}, X_i^{\perp})}{\text{var}(X_i^{\perp})}$$

And now we can write $\text{cov}(1_{a\le\varepsilon_t}, X_i^{\perp})$ compactly as

$$\text{cov}(1_{a\le\varepsilon_t}, X_i^{\perp}) = \text{cov}(1_{a\le\varepsilon_t}, X_i) - \sum_{n\ge 1:n\ne i}^{N} \text{cov}(1_{a\le\varepsilon_t}, X_n)\cdot\left(\frac{C_{i,n} - \boldsymbol{\Sigma}_{i,n}}{\text{var}(X_n^{\perp_i})}\right)$$

We are not out of the woods yet because we skipped unpacking the variance terms. But once that has been done we will have finished "simplifying", in that arbitrarily complex regressions can be defined in terms of estimands that feature only explicit variance and covariance terms.[7]

The strategy to deal with the variance terms will be very similar and hopefully easier to digest now that we have some machinery to work with. Again to deal with the simplest case ($i = 1$) first,

$$\text{var}(X_1^{\perp}) = \text{var}(X_1) + \sum_{n=2}^{N}\sum_{m=2}^{N} \frac{\text{cov}(X_1, X_n^{\perp_1})}{V_{1,n}} \frac{\text{cov}(X_1, X_m^{\perp_1})}{V_{1,m}} C_{n,m} - \sum_{n=2}^{N} \frac{\text{cov}(X_1, X_n^{\perp_1})}{V_{1,n}} 2C_{1,n}$$

using the notation as before for V and C. Once more, we have a situation where everything will follow this pattern,

[7] Of course, $(X'X)^{-1}X'Y$ is a better simplification under any sensible definition. "disambiguating" may be more appropriate

less indexing. The first layer is simple to write

$$\text{var}(X_i^{\perp}) = \text{var}(X_i) + \sum_{m,n\geq 1:m,n\neq i}^{N}\sum^{N} \frac{\mathbf{\Sigma}_{i,n}}{V_{1,n}}\frac{\mathbf{\Sigma}_{i,m}}{V_{1,m}}C_{n,m} - \sum_{n\geq 1:n\neq i}^{N}\frac{\mathbf{\Sigma}_{i,n}}{V_{1,n}}2C_{i,n}$$

The only thing remaining is to expand this definition so that it holds as terms are continually added to $\perp$ in $X_i^{\perp}$ (i.e., in the FWL regressions, some terms have already been partialed out and won't be included). To do this, we need to make the indexing of $\mathbf{\Sigma}_{i,n}$ a bit more flexible. Define

$$SV(i;\{O\}) = \text{var}(X_i^{\perp_{\{O\}}}) = \text{var}(X_i) + \sum_{m,n\geq 1:m,n\notin O}^{N}\sum^{N} \frac{\mathbf{\Sigma}^c_{\{O\},n}}{SV(n;\{O,i\})}\frac{\mathbf{\Sigma}^c_{\{O\},m}}{SV(m;\{O,i\})}C_{n,m} - \sum_{n\geq 1:n\notin O}^{N}\frac{\mathbf{\Sigma}^c_{\{O\},n}}{SV(n;\{O,i\}}2C_{i,n}$$

where O is a set of unique integers $o \in [1,N]\backslash\{i\}$ and

$$\mathbf{\Sigma}^c_{\{O\},n} = \sum_{\sigma\in P_-^{\{O\}}} S(\{\sigma\})$$

$$\text{with } P_-^{\{O\}} = \left\{\sigma \in \mathbb{Z}^{N-|O|} : \sigma \cup \{O\} \in P_N \ \&\ \forall n, \nexists m \text{ s.t } \sigma(m) = O(n)\right\}.$$

Noting that for any $\sigma \in P_N$, $SV\Big(i, \big\{\sigma\backslash\{i\}\big\}\Big) = \text{var}(X_i)$, our nightmare is finally over.

C.2 Deep Learning Implementation

Work is still ongoing to fine tune the algorithm and therefore to sharpen these recommendations. Also, the Online Appendix from this point forward is a work in progress, but for the remaining sections its a matter of compilation, rather than work that remains to be completed.

One consideration is complexity. I've found that convergence occurs rapidly, but convergence may not be to a collection of satisfactory functions (i.e., the neural net essentially gets stuck at a local minima). So while this may be more feasible without a GPU, the iterative nature of refining the loss computation may make this less feasible without access to hardware designed for efficient deep learning training. I hope to eventually share functions on this paper's GitHub repository that work well for standard normal shocks. When applied to other shock series, the performance will not be perfect but may be acceptable.

C.3 Standard Errors for Generated Regressors

Recall the generated regressor functions defined in Section 3 $\left\{f_i\right\}_{i=1}^{4}$. For clarity, we are interested in functions of a shock ε_t that is continuously distributed on $a \in I \subset \mathbb{R}$. Each of this function has a designated "peak" c_i and a set I_i with endpoints left_i and right_i). These functions are defined in terms of the empirical CDF $F_N(\cdot)$.

Specifically, for each $a \in I$

$$f_i(a) = \begin{cases} 0 & a \notin [\text{left}_i, \text{right}_i) \\ -[F_N(c_i) - F_N(\text{left}_i)]^{-1} & a \in [\text{left}_i, c) \\ [F_N(\text{right}_i) - F_N(c_i)]^{-1} & a \in (c, \text{right}_i) \end{cases}$$

with slight abuse of notation if $\text{left}_i = -\infty$. In a regression of y on $\{f_i\}_{i=1}^4$, the estimands will be defined in terms of the CDF $F(\cdot)$. Define p_{iL} as $F(c_i) - F(\text{left}_i)$ and $p_{iR} = F(\text{right}_i) - F_N(c_i)$. The estimand β_i on f_i is

$$\beta_i = \frac{\text{cov}(y, f_i)}{\text{Var}(f_i)} = \frac{\bar{y}_{iR} - \bar{y}_{iL}}{\frac{1}{p_{iL}} + \frac{1}{p_{iR}}},$$

where $\bar{y}_{iL}$ and $\bar{y}_{iR}$ are the means of y on the subsets of I_i given by p_{iL} and p_{iR}. To see this, recall f_i is mean 0. So

$$\text{cov}(y, f_i) = \mathbb{E}[y f_i] = -\frac{1}{p_{iL}} \mathbb{E}[y \cdot \mathbb{1}_{[\text{left}_i, c_i)}] + \frac{1}{p_{iR}} \mathbb{E}[y \cdot \mathbb{1}_{[c_i, \text{right}_i)}] = -\frac{1}{p_{iL}} \bar{y}_{iL} \cdot p_{iL} + \frac{1}{p_{iR}} \bar{y}_{iR} \cdot p_{iR} = \bar{y}_{iR} - \bar{y}_{iL}$$

Because this estimand is formed with respect to a generated regressor, we need to adjust the standard errors. Adjustment is done using the delta method. Differentiating

$$\frac{\partial \beta_i}{\partial p_{iL}} = \beta_i \cdot \frac{p_{iR}}{p_{iL}(p_{iL} + p_{iR})} \quad \underline{\textbf{and}} \quad \frac{\partial \beta_i}{\partial p_{iR}} = -\beta_i \cdot \frac{p_{iL}}{p_{iR}(p_{iL} + p_{iR})}$$

The adjustment takes the form of [8]

$$\left(\frac{\partial \beta_i}{\partial p_{iL}}\right)^2 \text{Var}(p_{iL}) + \left(\frac{\partial \beta_i}{\partial p_{iR}}\right)^2 \text{Var}(p_{iR}).$$

Using sample analogs, standard errors are the square soot of the sum of the usual Huber-White variance and

$$\frac{\hat{\beta}_i^2}{N} \left(\frac{\hat{p}_{iR}(1 - \hat{p}_{iL})}{\hat{p}_{iL}(\hat{p}_{iL} + \hat{p}_{iR})^2} + \frac{\hat{p}_{iL}(1 - \hat{p}_{iR})}{\hat{p}_{iR}(\hat{p}_{iL} + \hat{p}_{iR})^2} \right)$$

where $\widehat{\text{Var}}(p_{iL}) = \frac{\hat{p}_{iL}(1 - \hat{p}_{iL})}{N}$ (and similar for p_{iR}).

C.4 Details for Empirical Application in Section 4.2

Following Ramey (2016), outcome variables are the Consumer Price Index (CPI), industrial production, 1 year treasury yields, excess bond premium (Favara et al., 2016), unemployment, and add real consumption expenditures (all monthly frequency).[9] Control variables also include lagged interest rates, monetary policy uncertainty (Husted et al., 2020), an indicator for the ZLB, and a healthy number lags (12) of both outcome and controls following our discussion of standard errors. Data is sourced from FRED unless noted otherwise and the maximum sample periods are retained. Based on the availability of the shock series used in the paper, this is ultimately not an issue, but in general pre-1983 target funds rate data should be discarded to reflect incongruities in Fed policy norms

[8] Note that f_i is not differentiable at c_i

[9] Some of these variables are highly non-stationary (McCracken and Ng, 2016). Montiel Olea and Plagborg-Møller (2021) show LP is remarkably robust to the presence of unit roots and non-stationary variables. I find some anecdotal support for this: estimating in differences and summing for the cumulative effect in levels produces very similar IRFs to estimating in levels directly

(Thornton, 2006; Aruoba and Drechsel, 2025). Related, to aggregate to changes as intended by policymakers, a few earlier instances of "intermediate" changes (e.g., adjust 12 basis points immediately and 25 more in a few weeks) are cumulated. This adds another reason to be concerned about temporal aggregation bias (Jacobson et al., 2024) and likely to discard results at early horizons. I focus on CPI and the joint picture of output painted by industrial production, consumption, and unemployment in order to take the findings directly to models. Outcome variables are cumulative log differences, yielding an approximate percent change interpretation: $\widehat{\alpha}_h$ is represents the percent change in levels h periods after a shock at t. At $h = 12$, this takes a nice form of year over year growth.

The LP framework described in the paper can be used to illustrate possible nonlinearities, which I refer to as size and size effects. The most straightforward way to think of these effects is as functions of parameters. For the simple case of plotting in levels, the objects of interest are

$$\text{Size Effect}_h : \hat{\alpha}_h^B - \hat{\alpha}_h^S \qquad \text{Sign Effect}_h : \hat{\alpha}_h^P + \hat{\alpha}_h^N$$

A size effect exists if we can conclude the difference in the big and small (regime) coefficients are distinct from 0 and a sign effect exists if positive and negative coefficients have different magnitudes. This is complicated slightly by wanting to allow for both types of non-linearities simultaneously: we want to see if a size effect exists for both cuts and hikes and a sign effect exists for both big and small changes, in other words 4 graphs per outcome variable. Grouping is thus by type of nonlinearity, rather than outcome.

For sign effects, we can interpret positive statistically significant results as evidence for a "pushing on a string" narrative, the idea that it's more difficult (especially in recessions) for expansionary monetary policy to stimulate the economy than it is for contractionary policy to suppress it. Even if the individual point estimates go against what standard theory might suggest, we abstract from the notion of puzzles by simply focusing on one estimate relative to the other. For instance, if the coefficient of a big cut on CPI is -3 and the coefficient for a big hike is 2, these estimates are consistent with the string story because the the big hike's contractionary power, albeit a lack of one, is still greater than the expansionary effect of a big cut.

Before gauging how models holds up to the data-based findings presented in the main body of the paper, it's important to have a sense of what, if anything, can make these results weaken when pushed. Changing the lag order, adding and removing controls, estimating in differences vs. levels, different measures for inflation, bias-correcting point estimates (Herbst and Johannsen, 2024) and using LP instead of LP-IV in general do not yield IRFs with different interpretations, even under various combinations of these factors. One area where there is some sensitivity is sample selection and size, which is especially not surprising given the zero lower bound period. One option would be using a non-linear filtering procedure (e.g., Farmer, 2021) to construct a shadow interest rate, or a measure of how interest rates "would have moved" if the ZLB didn't bind.

A more involved critique of model-free estimation is an inability to account for state-dependence. For example,

many past efforts try to allow for responses in boom and bust cycles to be asymmetric. With respect to interest rate shocks of a given size, another worry could be that beliefs about the future path of policy may not be updated in the same for different histories of action. The econometric concern is that these local projection coefficients amount to weighted averages and these weights could be biased if the joint distribution of the shock and state space has disparate behavior from a product of their marginals. In a regression context, this essentially amounts to the difference between including a variable as a control and additionally interacting it with the shock. The work of Rambachan and Shephard (2025), Kolesár and Plagborg-Møller (2025), and results in the main body of the paper show that these estimates average out under arbitrary nonlinearities, so within this setting there is less room for concern. But because of the limited sample size, it's worth taking note of other, more directed approaches. Jordà (2023), following the revelation of Gonçalves et al. (2024) that the previous default methodology can severely distort impulse responses, provides a framework incorporating interaction terms to estimate state-dependent effects. Gonçalves et al. (2024) themselves suggest non-parametric estimation, which has an over-parameterization problem with or without instrumenting (i.e., in either case, control variables must be shed).

C.5 Details for Quantitative Application in Section 4.2

Description	Equation	#
Consumption Euler Equation	$1 = \beta \mathbb{E}_t \left[\left(\frac{C_{t+1}}{C_t} \right)^{-\tau} \frac{R_t}{\Pi_{t+1} \tilde{A}_{t+1}} \right]$	(1)
Definition for Real Wages	$\Delta_t^w = \frac{W_t}{W_{t-1}} \cdot \tilde{A}_t$	(2)
Resource Constraint	$\frac{G_t - 1}{G_t} \cdot Y_t + C_t = Y_t(1 - \Phi_t^p) - W_t Y_t \cdot \Phi_t^w$	(3)
Wage Equation, Household's problem	$\frac{\chi_h}{\lambda_w W_t} C_t^{\tau} Y_t^{\frac{1}{\nu}} + (1 - \Phi_t^w)\left(1 - \frac{1}{\lambda_w}\right) =$ $\Delta_t^{w_{nom}} \cdot \Phi_t^{'w} - \beta \mathbb{E}_t \left[\left(\frac{C_{t+1}}{C_t} \right)^{-\tau} \frac{\Pi_{t+1} (\Delta_{t+1}^w)^2}{\tilde{A}_{t+1}} Y_{t+1} \cdot \Phi_{t+1}^{'w} \right]$	(4)
Price Equation, Intermediate Firms problem	$(1 - \Phi_t^p) + \beta \mathbb{E}_t \left[\left(\frac{C_{t+1}}{C_t} \right)^{-\tau} \Phi_{t+1}^{'p} \Pi_{t+1} Y_{t+1} \right] = \frac{\mu_t}{\Lambda_t} + \Phi_t^{'p} \Pi_t$	(5)
Hours Equation	$W_t = (1 - \Phi_t^p) - \mu_t$	(6)
Adjustment Costs, Nominal Wages	$\Phi_t^w = \frac{\phi_w}{\psi_w^2} \left(e^{-\psi_w (\Delta_t^{w_{nom}} - \gamma \pi^*)} + \psi_w (\Delta_t^{w_{nom}} - \gamma \pi^*) - 1 \right)$	(7)
Adjustment Costs, Prices	$\Phi_t^p = \frac{\phi_p}{\psi_p^2} \left(e^{-\psi_p (\Pi_t - \pi^*)} + \psi_p (\Pi_t - \pi^*) - 1 \right)$	(8)
Derivative, Adjustment Costs Nominal Wages	$\Phi_t^{'w} = \frac{\phi_w}{\psi_w} \left(1 - e^{-\psi_p (\Delta_t^{w_{nom}} - \gamma \pi^*)} \right)$	(9)
Derivative, Adjustment Costs to Prices	$\Phi_t^{'p} = \frac{\phi_p}{\psi_p} \left(1 - e^{-\psi_p (\Pi_t - \pi^*)} \right)$	(10)
Taylor Rule	$R_t = \exp(r_t); \quad r_t = \rho_r r_{t-1} + (1 - \rho_r) r_t^* + \sigma_r \varepsilon_r$	(11)
TFP Growth	$\tilde{A}_t = \exp(a_t); \quad a_t = (1 - \rho_a) \log \gamma + \rho_a a_{t-1} + \sigma_a \varepsilon_a$	(12)
Government Spending Shocks	$G_t = \exp(g_t); \quad g_t = (1 - \rho_g) \log g^* + \rho_g g_{t-1} + \sigma_a \varepsilon_g$	(13)
Price Markup Shock	$\Lambda_t = \exp(\lambda_t); \quad \lambda_t = (1 - \rho_p) \log \lambda_{p_{ss}} + \rho_p \lambda_{t-1} + \sigma_p \varepsilon_p$	(14)
Output change	$\Delta_t^y = Y_t / Y_{t-1}$	(15)
Nominal wage change	$\Delta_t^{w_{nom}} = \Delta_t^w \Pi_t$	(16)
Interest rate target	$r_t^* = \log\left(\frac{\gamma}{\beta} \cdot \pi^* \right) + \psi_1 \log\left(\Pi_t / \pi^* \right) + \psi_2 \left(a_t + \log\left(\Delta_t^y / \gamma \right) \right)$	(17)

- For the set of draws that came out of our Metropolis- Hastings routine, I simulated data of 400 observations for each group of parameters to align with the US data sample size. Analogous control variables are included (lagged interest rates, zero lower bound, unemployment, output and interest rate variance) and plots are in terms of standard deviations to abstract away from any differences between model-simulated and US data. This is described and justified further in the next subsection.

- The priors are largely from Aruoba et al. (2017). Because of the difference in sample period, I scaled down the priors for annualized output growth (μ_y) and inflation (μ_π), as well as β^{-1}. In fact, it's actually not possible for this model to generate a steady state that matches the data. Steady state interest rates are $\mu_y + \mu_\pi + 400(\beta^{-1} - 1)$. If μ_y and μ_π are picked to match inflation data, you must pick $\beta > 1$ to match interest rate data.

- Some other fixes to the original replication code are detailed here.

- For consistency in the comparative static illustrations, it was necessary to make sure this mode line was the same across plots, but that meant the same series of shocks would need to be used for all sets of simulated data, which could paint a misrepresentative picture for a short sample size. So I plotted the median estimate for 100 samples for each parameter group (for simulation i, seed was set to $i = \{1, \ldots, 100\}$).
  - Ideally, this would be done for the Bayesian IRFs, but that would take a month to run. Implementation needed to be extremely parsimonious – because each loop of the LP file performs 25*number of outcome variables calls to regress, I randomly selected 10,000 draws of the post burn-in M-H output.

	h		
	0	1	2
Big Cut	-18.8%	-5.5%	-1.7%
Big Hike	36.6%	5.5%	0.4%

Table 1: Average % Deviation from i^*, h periods after large change in i_t

Next, I show that the model is capable of generating any type of nonlinearity on impact, but the effects dissipate quickly. To make efficient use of space, the exact figures are relegated to the very end of the Appendix, but there are hyperlinks to each. Again, what we learned from this exercise is that any sort of nonlinearity desired can be generated on impact using the right combination of asymmetry parameters, but it does not last for even one period longer in most cases.

Size Effects

	Description	**Anything Interesting? (all at $h = 0$)**	**Link**
1	$\psi_p \uparrow$	(slightly) amplifies negative size effect for hikes on π at $h = 0$	Figure 3
2	$\psi_p \downarrow$	(slightly) amplifies all $h = 0$ size effects except for hike on π	Figure 4
3	$\psi_w \uparrow$	(slightly) increases the positive size effect for cuts on Y at $h = 0$	Figure 5
4	$\psi_w \downarrow$	No.	Figure 6
5	$\psi_p \uparrow$, $\psi_w \uparrow$	amplifies size effect (-) of cuts on Y, depresses size effect (+) of hikes on Y	Figure 7
6	$\psi_p \downarrow$,$\psi_w \downarrow$	(slightly) amplifies negative size effect for hikes on π at $h = 0$	Figure 8

Sign Effects

	Description	Anything Interesting? (all at $h = 0$)	Link
1	$\psi_p \uparrow$	depressed all $h = 0$ sign effects except for small changes on π	Figure 9
2	$\psi_p \downarrow$	low values reversed the direction of the sign effect for big changes on π.	Figure 10
3	$\psi_w \uparrow$	Depresses small change on Y size effect and amplifies everything else	Figure 11
4	$\psi_w \downarrow$	(slightly) amplified sign effect of big changes on π and small changes on Y	Figure 12
5	$\psi_p \uparrow$, $\psi_w \uparrow$	depressed sign effect of small changes on Y and amplified everything else	Figure 13
6	$\psi_p \downarrow$,$\psi_w \downarrow$	reversed the direction of sign effect for big changes on π	Figure 14

C.6 Estimation in Terms of Standard Deviations

In a linear regression, coefficients are the estimated effect of a marginal (size), positive (sign) change. If we normalize our previous definitions by the standard deviation of the coefficient corresponding to this linear "default", we have an alternative formulation of size and sign effects in terms of standard deviations instead of percent change in levels at a given horizon. For example, if $\frac{\hat{\alpha}_{BC} - \hat{\alpha}_{SC}}{\sigma_{SC}} = 3$, the interpretation is that the big cut coefficient amounts to a 3 standard deviations away realization of the small cut coefficient. Additional intuition can be gleaned by noticing that if we instead normalized by the standard deviation of the entire (original) definition, we would simply have a t-statistic. This approach has the advantage of the y-axis having a uniform representation across all outcomes of interest and arguably removes some of the subjectivity implicit in deciding what % constitutes a meaningful effect for a given variable-horizon combination. Put differently, this representation sends a similar signal to the results of a hypothesis test (is there enough evidence from data to infer these parameters are drawn from distinct distributions), but unlike a t-statistic the units lend themselves more to economic meaning (moment of the distribution for our baseline coefficient, rather than a general normal distribution). Another motivation is model comparison. In principle, percent change is a "unitless" point of comparison. But in a small sample setting, having parameters in a DSGE model that dictate growth rates can induce distortions in scaling and correlations relative to time series data that add meaningless noise to analogous estimations of our objects of interest. By using standard deviations, everything is normalized by whatever scale persists in the DGP, meaning the finite sample idiosyncrasies are softened. Ultimately, percent change in levels carries more real world weight, but standard deviations add indispensable context.

References

Aruoba, S. Boragan, and Thomas Drechsel. 2025. "Identifying Monetary Policy Shocks: A Natural Language Approach." Working paper.

Aruoba, S. Borağan, Luigi Bocola, and Frank Schorfheide. 2017. "Assessing DSGE model nonlinearities." *Journal of Economic Dynamics and Control*, 83 34–54.

Barnichon, Régis, and Geert Mesters. 2025. "Innovations meet Narratives -improving the power-credibility trade-off in macro." January, Working paper.

Farmer, Leland E. 2021. "The discretization filter: A simple way to estimate nonlinear state space models." *Quantitative Economics*, 12(1): 1–45.

Favara, Giovanni, Simon Gilchrist, Kurt F. Lewis, and Egon Zakrajšek. 2016. "Updating the Recession Risk and the Excess Bond Premium."

Gonçalves, Sílvia, Ana María Herrera, Lutz Kilian, and Elena Pesavento. 2024. "State-dependent local projections."

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

Husted, Lucas, John Rogers, and Bo Sun. 2020. "Monetary Policy Uncertainty." *Journal of Monetary Economics*, 115 20–36.

Jacobson, Margaret M., Christian Matthes, and Todd B. Walker. 2024. "Temporal Aggregation Bias and Monetary Policy Transmission." March, Working Paper.

Jordà, Òscar. 2023. "Local Projections for Applied Economics."

Kolesár, Michal, and Mikkel Plagborg-Møller. 2025. "Dynamic Causal Effects in a Nonlinear World: the Good, the Bad, and the Ugly." Working Paper.

Mallat, St'ephane. 1999. *A Wavelet Tour of Signal Processing*.: Academic Press, i–xxiv, 1–637.

McCracken, Michael W., and Serena Ng. 2016. "FRED-MD: A Monthly Database for Macroeconomic Research." *Journal of Business & Economic Statistics*, 34(4): .

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

Rambachan, Ashesh, and Neil Shephard. 2025. "When do common time series estimands have nonparametric causal meaning?", Working paper.

Ramey, Valerie. 2016. "Macroeconomic Shocks and Their Propagation." In *Handbook of Macroeconomics*. Chap. 2.

Thornton, Daniel L. 2006. "When Did the FOMC Begin Targeting the Federal Funds Rate? What the Verbatim Transcripts Tell Us." *Journal of Money, Credit, and Banking*, 38(8): 2039–2071.

D.1 Figure and Equation Reference

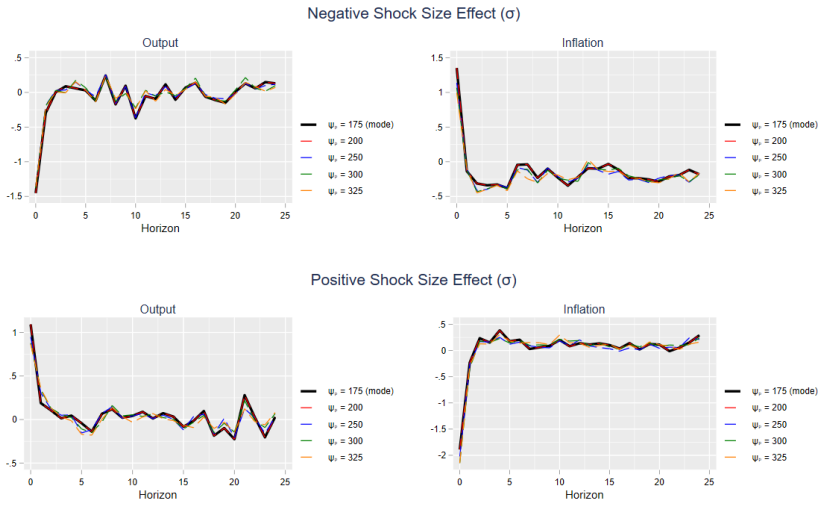

Figure 3: (click to go back to tables)

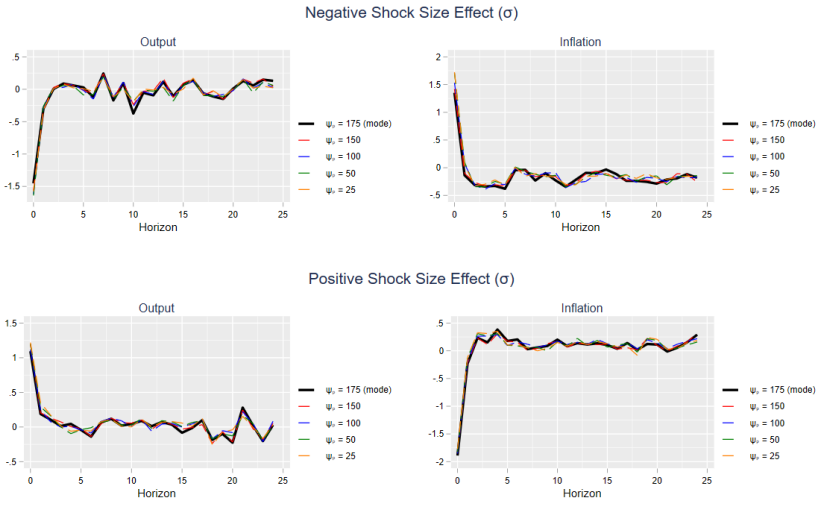

Figure 4: (click to go back to tables)

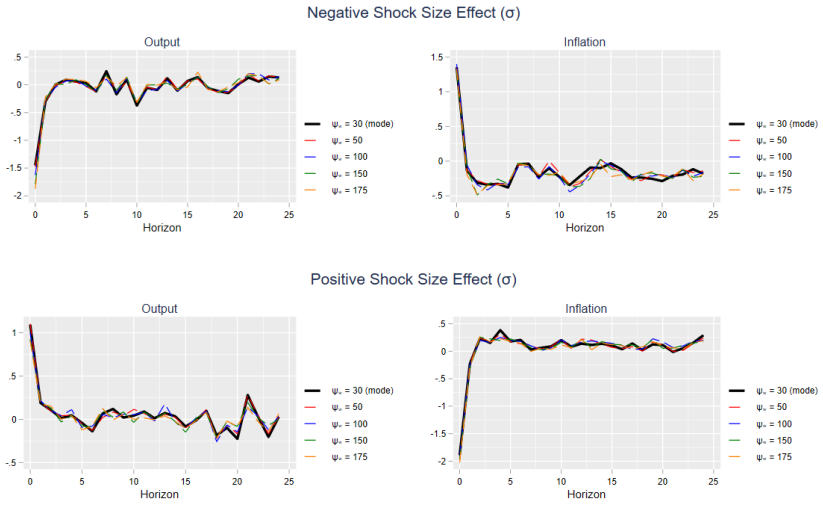

Figure 5: (click to go back to tables)

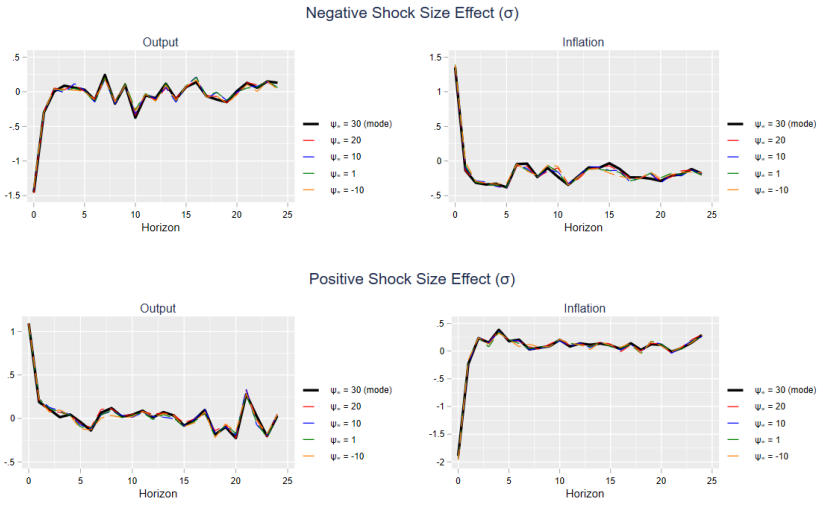

Figure 6: (click to go back to tables)

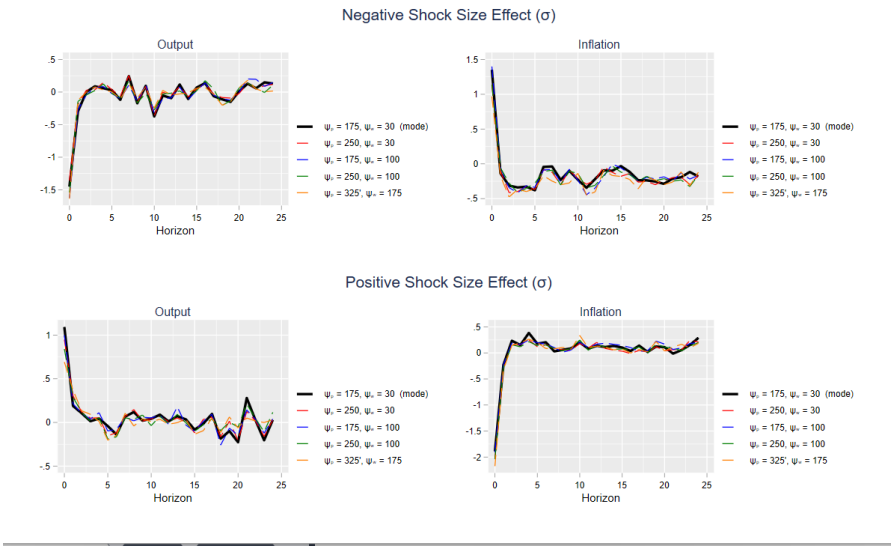

Figure 7: (click to go back to tables)

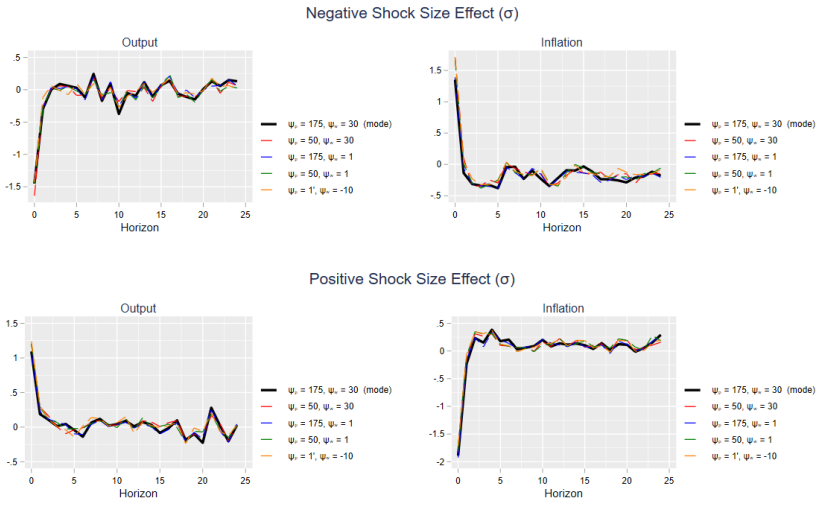

Figure 8: (click to go back to tables)

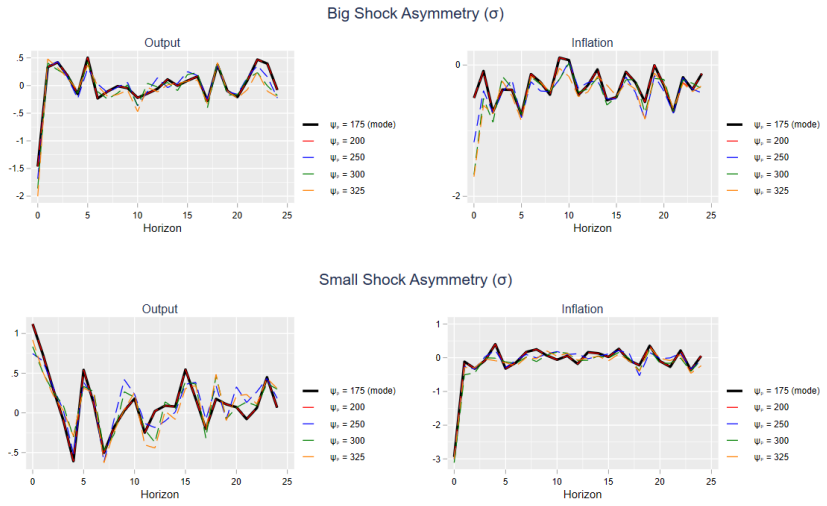

Figure 9: (click to go back to tables)

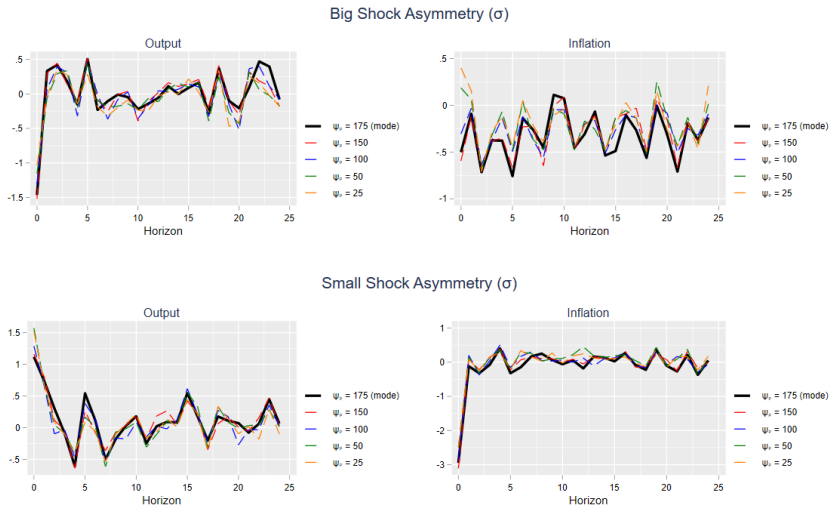

Figure 10: (click to go back to tables)

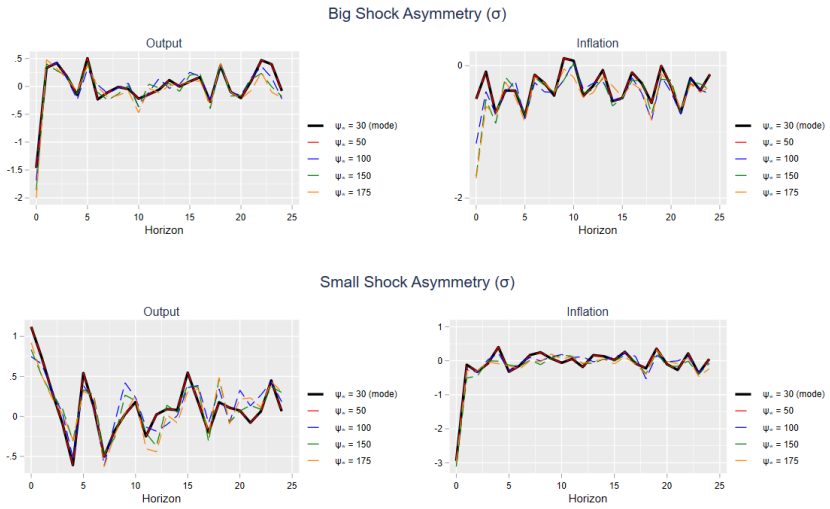

Figure 11: (click to go back to tables)

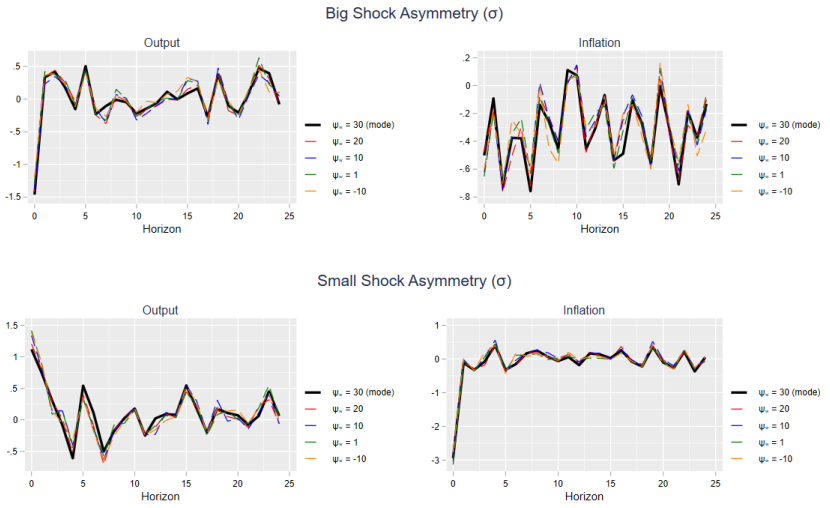

Figure 12: (click to go back to tables)

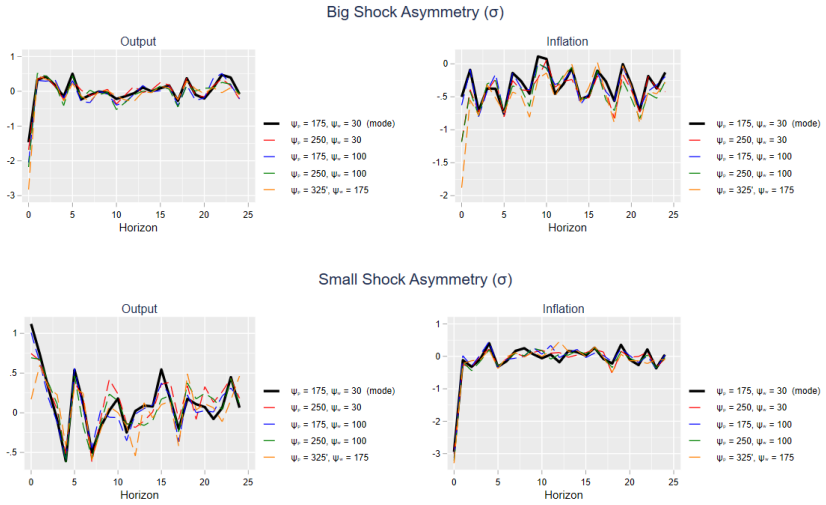

Figure 13: (click to go back to tables)

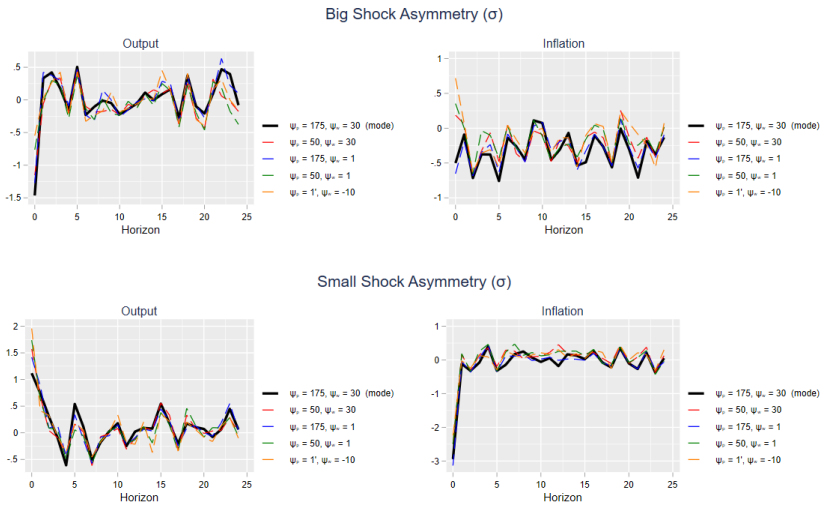

Figure 14: (click to go back to tables)

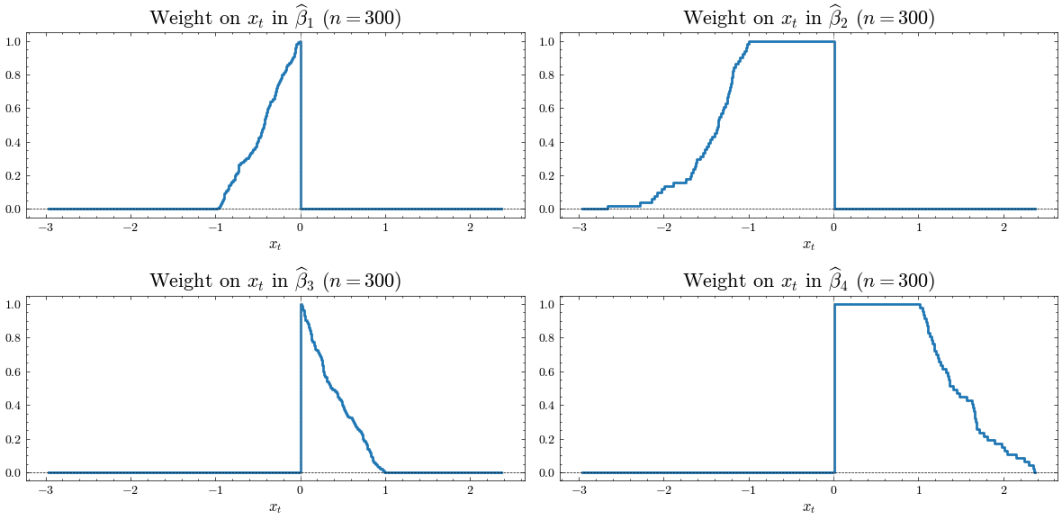

Figure 15: Standard Normal Shock with Example 6

Example 1: $$\pi_t = \alpha + \beta_1 x_t \cdot \mathbf{1}_{x_t \leq 0} + \beta_2 x_t \cdot \mathbf{1}_{x_t > 0} + u_t$$

Example 5: $$\pi_t = \alpha - \beta_2 \cdot \mathbb{1}_{-x_t \in (0,1]} - \beta_3 \cdot \mathbb{1}_{-x_t < 1} - \beta_4 \cdot \mathbb{1}_{x_t \in [0,1]} + u_t$$

Example 4: $$\pi_t = \alpha - \beta_{\text{small, neg}} \cdot \mathbb{1}_{-x_t \in [\frac{1}{100}, \frac{5}{4}]} - \beta_{\text{big, neg}} \cdot \mathbb{1}_{x_t < \frac{5}{4}} + \beta_{\text{small, pos}} \cdot \mathbb{1}_{x_t \in [\frac{1}{100}, \frac{5}{4}]} + \beta_{\text{big, pos}} \cdot \mathbb{1}_{x_t > \frac{5}{4}} + u_t$$