

# Uncovering Nonlinearities with Regression Anatomy

Paul Bousquet\*

February 22, 2025

[\[Click here for latest version\]](#)

## Abstract

[Kolesár and Plagborg-Møller \(2024\)](#) show (i) even under an arbitrarily complex data generating process, linear regression frameworks can estimate a weighted average of a shock's marginal effects (ii) previous, more involved efforts to directly estimate non-linear effects produce flawed inference. The price of a vanilla regression's lack of sensitivity is a black box: one point estimate cannot reveal where and to what extent nonlinearities exist. I show how to exploit the mechanics of least-squares regression and develop specifications to jointly test if marginal effects have sign and size dependence. Using monetary policy shocks as an application, I find persistent nonlinearities in US data that cannot be replicated by a New Keynesian model with asymmetric rigidities in price and wage setting.

---

\*Department of Economics, University of Virginia, [pbousquet@virginia.edu](mailto:pbousquet@virginia.edu). I thank Rachel Childers and Eric Young for helpful comments.

# 1 Introduction

Imbens and Angrist (1994), Yitzhaki (1996), and Angrist et al. (2000) show simple regression estimands are a weighted average of true underlying marginal effects. In some respects, this remarkable claim was indeed too good to be true. The "weights" in this averaging do not have a straightforward interpretation and the result breaks in many standard settings (Masten, 2024). The implicit weights can also be negative, opening the door for the worst case scenario in causal inference: the regression estimand could have the opposite sign as marginal effects, so our estimates will be wrong no matter how much data we have (Small et al., 2017; Goldsmith-Pinkham et al., 2024).

While the particulars of the credibility revolution were being ironed out in microeconometrics, pure least-squares regression became more popular in macroeconomics thanks to Jordà (2005) local projection (LP), often used to estimate a shock's macroeconomic effects using a constructed shock proxy (e.g., Romer and Romer, 2004). Rambachan and Shephard (2021) and Kolesár and Plagborg-Møller (2024) revitalize the weighted average result with analogous propositions for LP and vector autoregression (VAR) that hold under commonly assumed conditions for these shock series. Kolesár and Plagborg-Møller (2024) also make a pragmatic point that while the regression weights aren't readily interpretable, they are easily estimatable. By digging into the mechanics of these weights, they show a vanilla linear regression is in general a much better tool to recover non-linear effects than specifications that explicitly try to capture nonlinearities (e.g., by including  $x^2$  as a regressor). This is because, unlike a standard local projection, non-linear regression is sensitive to misspecification (White, 1980).

Though inference is clean in a standard linear regression on a shock, Kolesár and Plagborg-Møller (2024) note their work cannot indicate if nonlinearity exists, an important consideration for modeling in Macro. One gauge of a model's match to data are impulse response functions (IRFs), which depend on a shock's size  $\alpha$  and time since it occurred  $h$ . Linear models have separable IRFs:  $f(\alpha, h) = \alpha g(h)$ , ruling out nonlinearities like *size effects* (disproportionate impact of big and small shocks) and *sign effects* (asymmetry of positive and negative shocks). Caravello and Martínez Bruera (2024) demonstrate how to (separately) test for size and sign effects in data. They focus on ensuring testing for size nonlinearities is not contaminated with traces of sign nonlinearities (and vice versa), but their identification result and method is sensitive to the distribution of the shock. The approach also cannot jointly account for size and sign effects (i.e., how size-dependence differs for positive and negative shocks) and is limited by relying on pure significance tests, which may be inherently unrevealing given the large variance of local projection estimates (Li et al., 2024).

This paper builds on past work looking carefully "under the hood" of regressions and presents a method to jointly identify size and sign nonlinearities in data. The procedure exploits that implicit regression weights depend only on the shock (not the outcome variable) and seeks out specifications placing weight in the desired parts of a shock's support. Broadly, if we consider 4 types of shocks along the dimensions of big vs. small and positive vs. negative, the goal is to have a regression with 4 corresponding coefficients. "Corresponding" in this context

means including just the right combinations of regressors so that, for example, the regression weights  $\omega(\varepsilon)$  on the big, positive shock coefficient are only non-zero for  $\varepsilon$  sufficiently large. Appropriate weighting justifies labeling  $\beta_{i,j}$  with combinations of  $i = \{\text{big, small}\}$  and  $j = \{\text{positive, negative}\}$ . Testing for nonlinearities is then a simple task: for size effects, the null hypothesis is  $\beta_{\text{big},j} = \beta_{\text{small},j}$  and for sign effects it's  $\beta_{i,\text{pos}} = -\beta_{i,\text{neg}}$ . An advantage is coefficient differences may be significant even if the underlying coefficients aren't.

Formally, the purpose of the paper is showing what functional regressors have the best weighting properties to detect nonlinearities. A naive starting point of disjoint indicator functions turn out to be a safe way to carry out the procedure: under an arbitrary shock distribution, there will be no false positives in population estimates and the weights converge quickly in finite samples. To retain this feature while adding robustness to false negatives, we have to confront the tangled mapping between functions and their weights. Weights can be expressed compactly using Frisch–Waugh–Lovell (c.f., "regression anatomy" in Angrist and Pischke (2009)), but really they are complex combinations of second moments, so it's not obvious how to get the weighting we want. It's not even obvious exactly *what* we want because setting a definitive threshold for a "big" shock is an impossible task (i.e., the paradox of the heap).<sup>1</sup> Rather than adhering to a strict threshold, we can plot the weights generated by the inclusion of candidate regressors and decide ex-post if it's sufficient (Kolesár and Plagborg-Møller, 2024). I show two methods of constructing functions with good weighting properties, each with their own appeal: orthogonal generated regressors (closed-form and easy to extend beyond 4 shock types) and deep learning (don't vary with sampling). Given the sample sizes in most settings, simple indicator functions may often be the best choice.

I also synthesize recent work on local projections to form an implementation guide. This includes selection criteria for the vast array of structural shock proxies and some notes of caution about potential misuses of indicator functions in regressions. I apply these recommendations to assess the effects of monetary policy shocks on U.S. fundamentals and find nonlinearities for all variables generally peaking in the medium to long-run, with firmest indications for size effects for both positive and negative shocks and sign effects in big shocks. Barnichon and Matthes (2018) find similar sign effects using unemployment and inflation and conjecture they can be rationalized by a New Keynesian model with asymmetric adjustment costs in price and wage setting (Kim and Ruge-Murcia, 2009). I use a Metropolis-Hastings routine to estimate the Aruoba et al. (2017) extension of the model and find these nonlinearities do appear on impact but quickly vanish. This lends support to Friedman (1960)'s "long and variable lags", but in an era where central banks don't exert control over monetary aggregates (Cochrane, 2024), it's not clear what mechanism would yield such a transmission path.

---

<sup>1</sup>"One grain of sand is not a heap of sand, two grains of sand is not a heap of sand, ..., one million grains of sand is a heap of sand"

## 2 Current Paradigm

### 2.1 Environment

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

$$Y_{t+h} = g_h(\varepsilon_t, \mathbf{S}_{t+h}) \quad (1)$$

Here,  $\varepsilon_t$  is the structural shock of interest at time  $t$  and  $\mathbf{S}_{t+h}$  is "everything else" in the system, which could for instance include the information set at time  $t$  as well as leads and lags of  $\varepsilon_t$  (and other shocks). Following [Rambachan and Shephard \(2021\)](#) and [Kolesár and Plagborg-Møller \(2024\)](#), the working definition of a shock, with respect to a data generating process of the form in (1), is that it satisfies  $\varepsilon_t \perp \mathbf{S}_{t+h} \forall h \geq 0$ . In that case, note that the conditional mean  $\mathbb{E}[g_h(a, \mathbf{S}_{t+h}) | \varepsilon_t = a]$  can be written as  $m_h(a)$  for some function  $m_h(\cdot)$ .

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

$$\begin{aligned} Y_{t+h} &= \alpha + \beta_1 f_1(\varepsilon_t) + \dots + \beta_N f_N(\varepsilon_t) + \gamma' \mathbf{W}_t + \varepsilon_{t+h} \\ &= \alpha + \boldsymbol{\beta}' \mathbf{X}_t + \gamma' \mathbf{W}_t + \varepsilon_{t+h} \end{aligned} \quad (2)$$

where  $\mathbf{X}_t$  is a concatenation of  $\{f_i(\varepsilon_t)\}_{i=1}^N$ . If  $\varepsilon_t$  is a shock and continuously distributed on an interval  $I \subset \mathbb{R}$ , [Kolesár and Plagborg-Møller \(2024\)](#)'s Proposition 1 can be extended to show that

$$\beta_i = \int_I \omega_i(a) \cdot m'_h(a) da \quad (3)$$

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

where  $X_i^\perp$  is the residual from projecting the  $i$ th element of  $\mathbf{X}_t$  on the remaining  $N - 1$  elements.<sup>2</sup> Thus, the estimands can be described as a weighted average of the data generating process' true marginal effects. In Appendix A.4, the Fresh-Waugh-Lovell truncation is expanded to provide more explicit closed form solutions.

Estimands are a weighted average of marginal effects that can be arbitrarily non-linear, but estimation is a black box with output that sheds no light on the existence of nonlinearities, namely *size effects* (disproportionate impact of big and small shocks) and *sign effects* (asymmetric impact of positive and negative shocks). A more few things to take stock of before proceeding. Notice the weights (4) only depend on  $\varepsilon_t$ . This of course will not hold if  $\varepsilon_t$  is not actually a shock (it will also depend on the control set  $\mathbf{W}_t$ ). In addition, if we instead use a proxy  $z_t$  in place of  $\varepsilon_t$  (if  $\varepsilon_t$  is not observable), the weights still depend on  $\varepsilon_t$ . Also, the estimand's form says nothing about the finite sampling properties of an estimator  $\widehat{\beta}_i$ . These issues will be discussed at length in the next section.

---

<sup>2</sup>And a constant. Also need  $\{f_i(\varepsilon_t)\}_{i=1}^N$  s.t rank condition holds, as will be discussed in [Proposition 1](#). Notice that the weight form means the weights can be estimated easily using a loop of OLS regression.

## 2.2 Past Efforts To Estimate and Identify Non-Linear Marginal Effects

A large literature in applied macroeconomics has tried to estimate the effects of policy (e.g., interest rates or government spending) by using [Jordà \(2005\)](#) local projection or vector autoregression in conjunction with a constructed shock series meant to represent plausibly exogenous change (e.g., [Romer and Romer, 2004](#)). The default is to use a completely linear structure. Relative to the framework of (2), this means the only regressors are the identity function of the shock and the control set. Some work has included other functions of the shock, like  $f(\varepsilon) = \varepsilon^2$ , in addition to the identity function in an attempt to capture non-linear effects of shocks. [Caravello and Martínez Bruera \(2024\)](#) provide a survey of many past efforts and finds such specifications are sometimes incorrectly characterized. They consider a special case of (2)

$$Y_{t+h} = \alpha + \beta_1 \varepsilon_t + \beta_2 f(\varepsilon) + \gamma' \mathbf{W}_t + \varepsilon_{t+h} \quad (5)$$

With respect to (5), they show if  $\varepsilon_t$  is a shock that follows a symmetric distribution then

$$(i): f(\cdot) \text{ is even \& DGP features no sign effects} \implies \beta_2 = 0$$

$$(ii): f(\cdot) \text{ is odd \& DGP features no size effects} \implies \beta_2 = 0$$

These results provide important clarity on past work (e.g.,  $\varepsilon^2$  as a regressor isn't informative about size effects) and provide a clear strategy to test for nonlinearities. Because these statements hold regardless of the DGP's other properties, the presence of sign-dependence won't distort the detection of size-dependence and vice versa. While this separation property is valuable, it still leaves some questions left unanswered. For example, if we include  $f(\varepsilon) = \varepsilon^3$  and reject the null hypothesis that  $\beta_2 = 0$ , we might feel comfortable concluding there are size effects but cannot say more. There are many possibilities for the nature of the nonlinearity – in the extreme case, only negative shocks have size effects (and positive shocks don't) or vice versa. These possibilities, which we can't distinguish between at present, carry vastly different implications. This is also merely an identification result; it says nothing about finite sample properties of hypothesis testing coefficients in (5). Later parts of the paper will show simulations illustrating instances where performance may be lacking, even in ideal circumstances where the identification results hold exactly because the shock is symmetrically distributed. As the distribution becomes more asymmetric, as is the case for the monetary policy shock application in [Section 4](#), their approach is less useful. Related, the procedure is relatively inflexible, as the best choices for  $f(\cdot)$  are the same across shock series (ex ante). Because of sample size restrictions and the variety of distributions  $\varepsilon_t$  could follow, this is a notable limitation.

Besides a conflation of size and side effects, some past work with specifications like (2) incorrectly ascribed causal meaning to the estimands. [Kolesár and Plagborg-Møller \(2024\)](#) show that unless the data generating process (1) matches the regression structure exactly, causal inference is not possible. For example, suppose we use (5) with  $f(\varepsilon) = \varepsilon^2$ . Unless the conditional mean of  $Y$  is a quadratic function in  $\varepsilon$ ,  $\beta_1 + 2\beta_2\varepsilon$  is not a consistent estimate for the average marginal effects of  $\varepsilon$ . This is because a corollary to their Proposition 1 is in specification (5),

there must be negative weight placed on  $\beta_2$  (see Appendix A.2 for a proof). In general, specifications that include functions of  $\varepsilon$  as regressors cannot be used to estimate causal effects (White, 1980) but are rather a means to detect nonlinearities. In contrast, the form of (3) shows simply using a linear specification will consistently estimate a positively weighted average of the true average marginal effects. There's perhaps a counterintuitive takeaway from the above: we often think of regressions as measuring the effect of a "unit change", but the statement only applies in full to predicted values. For example, if we project  $y_t$  on  $\varepsilon_t$  and also  $y_t$  on  $\mathbb{1}_{\varepsilon_t \geq 0}$ , both estimands are an average of the *same* object (marginal effects of  $\varepsilon_t$  on  $y_t$ ) – the difference is the weights in this averaging. The next section will unpack the relationship between functional regressors and their weights.

In sum, linear regression is a surprisingly powerful tool for estimating non-linear marginal effects of a shock. The important qualifier is estimates represent an approximation to a weighted average across a shock's entire support. While the weights' form is known, underlying marginal effects are not; in other words  $\sum_i^M \omega_i \cdot m'_i = \beta$  is still one equation with  $M$  unknowns. Recovering the exact marginal effect of a given value of  $\varepsilon$  is not possible, but it is possible to test whether the marginal effect function is non-linear by augmenting linear regressions with the proper functions. There does seem to be room to expand past approaches along the extensive margin (i.e., what kinds of nonlinearities) which may even open the door to statements about the intensive margin (i.e., how non-linear). The rest of the paper will focus on how to use linear regression to be more descriptive about the types of nonlinearities that exist in a DGP.

### 2.3 An Illustration of The Problem

The objective of Section 2 is to describe the status quo as concisely as possible, which thus far mostly involved extending the analysis of more technical papers like Rambachan and Shephard (2021) and Kolesár and Plagborg-Møller (2024). But what it means to have a weighting scheme  $\frac{\text{Cov}(\mathbb{1}_{\{a \leq \varepsilon_t\}}, X_t^+)}{\text{Var}(X_t^+)}$  is not obvious, so an example is useful. Under the following DGP<sup>3</sup>

$$y_t = \varepsilon_t^d, \pi_t = c(y_t) + \beta \mathbb{E}_t[\pi_{t+1}] + \varepsilon_t^s \quad \text{where } \varepsilon_t^d \sim \mathcal{U}[-a, a], \varepsilon_t^s \sim \mathcal{N}(0, \sigma^2), c(y) = \begin{cases} \kappa y^b & \text{if } y > 0 \\ 0 & \text{o.w} \end{cases} \quad (6)$$

note that  $\mathbb{E}_t[\pi_{t+1}]$  will be a constant. So a regression of  $\pi_t$  on  $y_t$ , or functions of  $y_t$ , should be revealing. Because of the simple structure, we might expect a specification of

**Example 1:** 
$$\pi_t = \alpha + \beta_1 y_t \cdot \mathbf{1}_{y_t \leq 0} + \beta_2 y_t \cdot \mathbf{1}_{y_t > 0} + \varepsilon_t$$

to perform well in estimating marginal effects. In context of the previous discussion, the logic is the following: (3) showed regression estimands are weighted averages, so shouldn't weight only be placed where the indicator

---

<sup>3</sup>Motivated by a basic New Keynesian model. Caravello and Martínez Bruera (2024) use a special case to illustrate their separation result, and I found tinkering with it was very helpful to understand the broader mechanics of the weights.

functions are equal to 1 (active)? But this is not the case. Using the form in (4) we can plot the weights. Figure 1 shows that while the *aggregate* weight where the indicators are not active is indeed 0, this is only because there is positive and negative weight that cancels out. For the estimand on  $y \cdot \mathbb{1}_{y>0}$ , there is no issue because marginal effects are 0 where the indicator is not active. However,  $\widehat{\beta}_1$  will not converge to 0 unless marginal effects are constant for  $y > 0$  (i.e., only if  $b = 0, 1$ ). This result holds more generally under standard choices for the distribution of  $y_t$ . Related, another possibly surprising revelation from Figure 1 is weights are not relatively equal across the relevant parts of the shock's support, even though it follows a uniform distribution. In fact, the weight plots look similar when  $y_t$  follows a normal distribution.

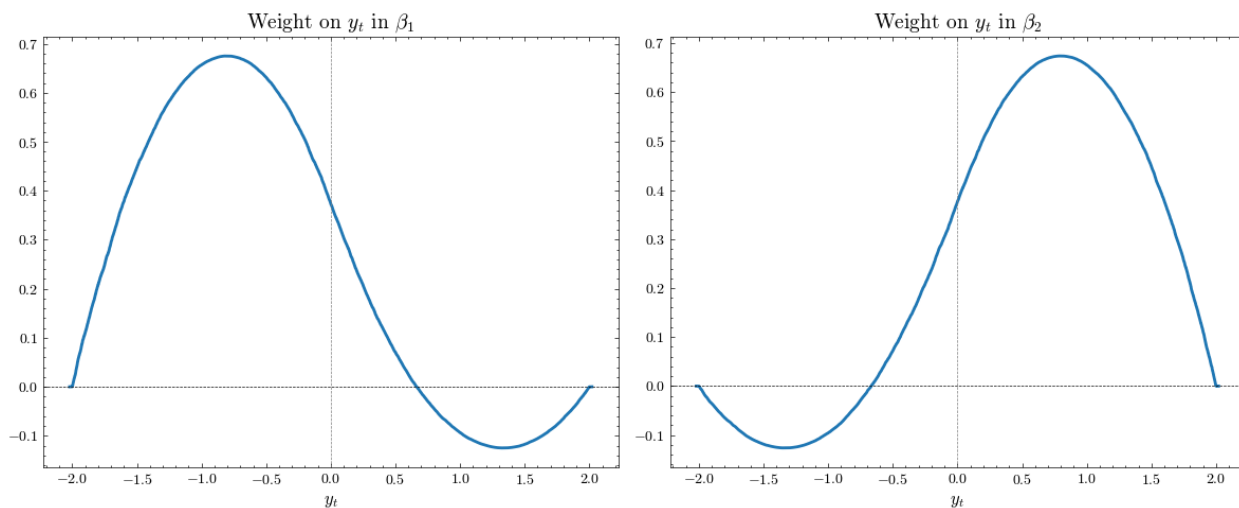


Figure 1: Uniform Shock with Example 1

Next, the Caravello and Martínez Bruera (2024) benchmark for nonlinearity detection for this case is based on

**Example 2:**

$$\pi_t = \alpha + \beta_1 y_t + \beta_2 f(y_t) + \epsilon_t$$

The success of the framework varies widely by how it's parameterized. For our example DGP in the  $b = 1$  case, there are only sign effects. With uniform shocks, even though the structure of the DGP is simple and the shock distribution is symmetric, the detection performance is poor with an realistic sample size: in only 16% of 10,000 simulations with  $n = 300$ , a null hypothesis of no size effects is rejected in a level-.05 test. The performance is better with standard normal shocks, rejecting in 73% of simulations. Similarly, for  $b = 2$  and standard normal shocks, there are now size effects for positive shocks, but the null of no size effects is not rejected in 35% of simulations. This highlights the limited power an identification result has in finite samples. To make the failure more transparent, Figure 2 plots the weights in the size effect specification for one of the simulations next to its limit.<sup>4</sup> Even in the most aspirational scenario when shocks follow a well-behaved, symmetric distribution, the weights may be far from converging.<sup>5</sup>

<sup>4</sup>Here,  $f(y) = \mathbb{1}_{y \geq \bar{y}} \cdot (y - \bar{y}) + \mathbb{1}_{y \leq -\bar{y}} \cdot (y + \bar{y})$ , where  $\bar{y}$  is  $\sigma$  away from the mean (0). Results are similar for  $f(y) = y^3$ .

<sup>5</sup>The  $n = 300$  graph in Figure 2 varies across samples. The median simulated error relative to the sum of the area in each quadrant is 20%. In the language of Caravello and Martínez Bruera (2024), we would say this is problematic because the odd weights are non-trivial.

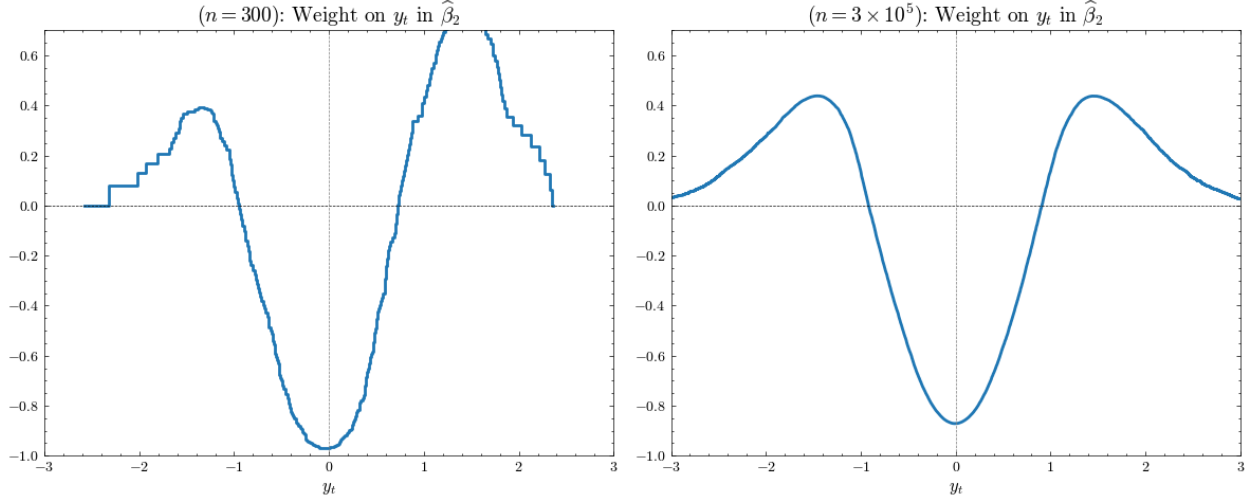


Figure 2: Standard Normal Shock with Selected Simulation of [Example 2](#)

Another issue is interpretability. After conducting hypothesis tests on the [Example 2](#) specification, we are still pretty much in the dark about the underlying DGP. Even if the nulls are properly rejected, we can at best operate under the belief that positive shocks have generally larger effects than negative shocks and (in the  $b = 2$  case) big shocks generally have disproportionately larger effects than small shocks, but this is imprecise. At a minimum, we should seek to get more specific than "generally".

Finally, we can preview the ultimate strategy outlined in [Section 3](#) by looking at similar approaches that don't work. There are two common approaches for incorporating indicator functions into regressions. One strategy is to just include indicator functions on their own. For the application where we are using indicator functions of a shock  $\varepsilon_t$ , we have established that if  $\varepsilon_t$  is included in the regression, there must be some negative weighting in the estimands on the indicator functions. But some may conjecture that in this case, the negative weights have a purpose. To fix ideas, we again use DGP (6) with standard normal shocks and consider

**Example 3:**

$$\pi_t = \alpha + \beta_1 \varepsilon_t + \beta_2 \cdot \mathbb{1}_{-y_t \in (0,1]} + \beta_3 \cdot \mathbb{1}_{y_t < 1} + \beta_4 \cdot \mathbb{1}_{y_t \in [0,1]} + \varepsilon_t$$

$\mathbb{1}_{y_t > 1}$  is excluded for collinearity (otherwise we would have linearly dependent regressors). We might intuitively expect negative weights to be a necessary component to estimate effects relative to  $\mathbb{1}_{y_t > 1}$ . Because estimands are weighted averages of the same object ( $m_h(\cdot)$  in [Section 2.1](#)), a coefficient difference  $\beta_i - \beta_j$  is equivalent to one object  $\tilde{\beta} = \int_I \tilde{\omega}(a) m_h(a) da$  where  $\tilde{\omega}(a) = \omega_i(a) - \omega_j(a)$ . By the same token, we might expect that the weights in the specification in [Example 3](#) take on a similar form. In the ideal case, the weights in  $\beta_i$  (for  $i = 2, 3, 4$ ) are positive when the corresponding indicator function is equal to 1 and are negative when  $\varepsilon_t > 1$ . This would support the earlier described interpretation as these coefficients being "effects in region  $i$  relative to the excluded region". [Figure 3](#) shows this is not the case. While the weights in  $\beta_1$  (the coefficient on  $\varepsilon_t$ ) take on a possibly useful form where weight is more evenly dispersed throughout the shock's support, the indicator function weights are erratic and frequently negative. Had we instead used "negative indicators" (i.e., codomain of  $\{-1, 0\}$ ,



rather than  $\{1, 0\}$ ), the indicator weights would be reflected across the  $x$ -axis (and thus substantially reduce the presence of negative weights), but those estimands are not of any more use. [Section 3](#) and [Appendix A.1](#) shed some more light on the properties of the linear regression weights that result in this behavior.

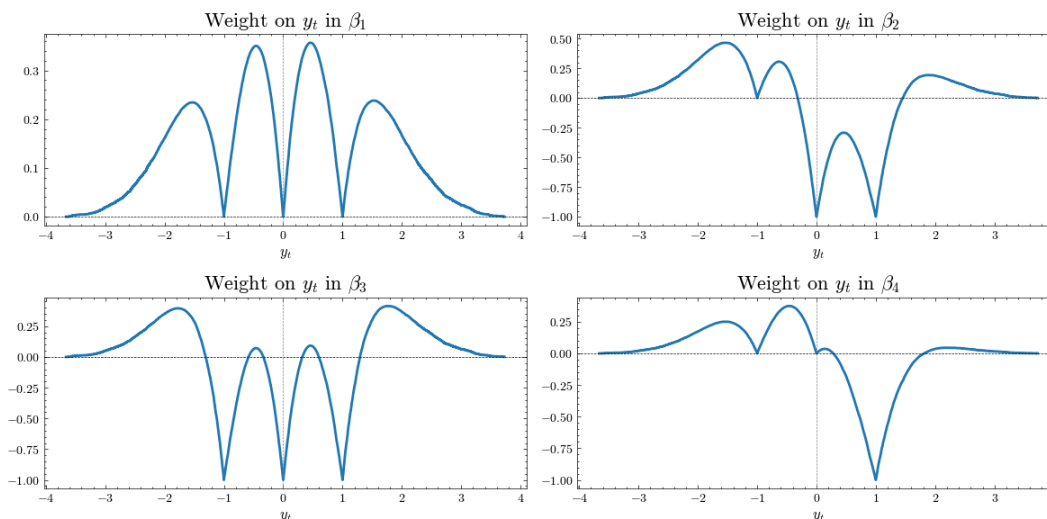


Figure 3: Standard Normal Shock with [Example 3](#)

Again, [Figure 3](#) highlights the disconnect between our instinct when looking at regression coefficients as representing the effect of a unit change and what the estimands actually imply. All estimands on functions of  $f(\varepsilon_t)$  are weighted averages of the same object. The only difference is how the weighting occurs, and it's often not what we would expect. Similarly, if we modify [Example 3](#) by removing  $\varepsilon_t$ , the results might also be surprising. Instead using

**Example 4:**

$$\pi_t = \alpha - \beta_2 \cdot \mathbb{1}_{-y_t \in (0,1]} - \beta_3 \cdot \mathbb{1}_{y_t < 1} - \beta_4 \cdot \mathbb{1}_{y_t \in [0,1]} + \varepsilon_t$$

[Figure 4](#) shows we no longer have negative weights, which makes sense ex-post given the removal of  $\varepsilon_t$ . At the same time, this means we still cannot describe what we're estimating as average marginal effects relative to the excluded indicator region. The estimands now do have a clear interpretation, but in some cases it initially seems at odds with the [Example 4](#) specification. For each estimand  $\beta_i$ , say the corresponding indicator is active on interval  $I_i$ . [Figure 4](#) shows the estimands are weighted averaged of marginal effect on the interval from the left endpoint of  $I_i$  to  $\infty$ .<sup>6</sup> So these objects may have some use; under a linear DGP, these estimands will all be the same. But the weights are, with respect to the indicators they relate to, poorly targeted. [Proposition 1](#) in [Section 3](#) will formalize why this happens. The ultimate revelation is it matters significantly what indicator region is "excluded". Because structural shock proxies often have many 0 values, it will make sense to instead exclude a small region around 0. This is because there will be enough excluded probability mass to not have linearly dependent regressors while at the same time enforcing that indicators on one sign of the real line don't put significant weight on the other.

<sup>6</sup>Because weights is trivial where  $\varepsilon_t$  has low density, weights are essentially 0 starting from  $\approx 3$  standard deviations.

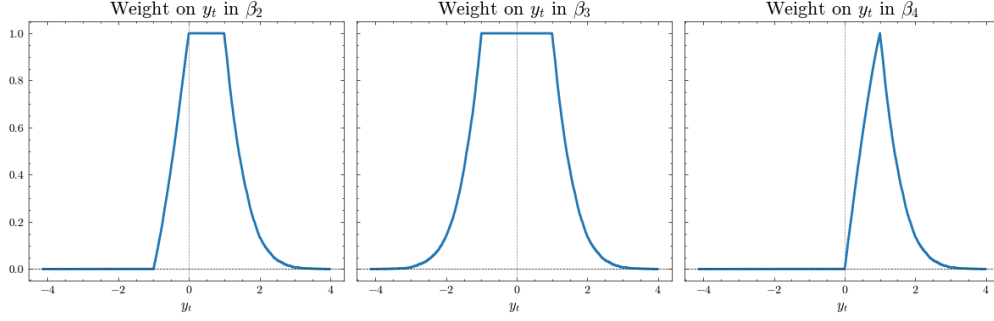


Figure 4: Standard Normal Shock with [Example 4](#)

Another use of indicator functions would be to interact them with a shock as in [Example 1](#)

**Example 5:** 
$$\pi_t = \alpha + \varepsilon_t \left( \beta_1 + \beta_2 \cdot \mathbb{1}_{-y_t \in (0,1]} + \beta_3 \cdot \mathbb{1}_{y_t < 1} + \beta_4 \cdot \mathbb{1}_{y_t \in [0,1]} \right) + \varepsilon_t$$

[Figure 5](#) shows we don't recover a "effect relative to the excluded region" interpretation.<sup>7</sup>  $\beta_4$  is a slight exception, but there is still weight placed for shock values less than 0. The weights in  $\beta_3$  follow the clearest pattern, but do not allow the effects of size and sign nonlinearities to be disentangled ([Caravello and Martínez Bruera, 2024](#)).

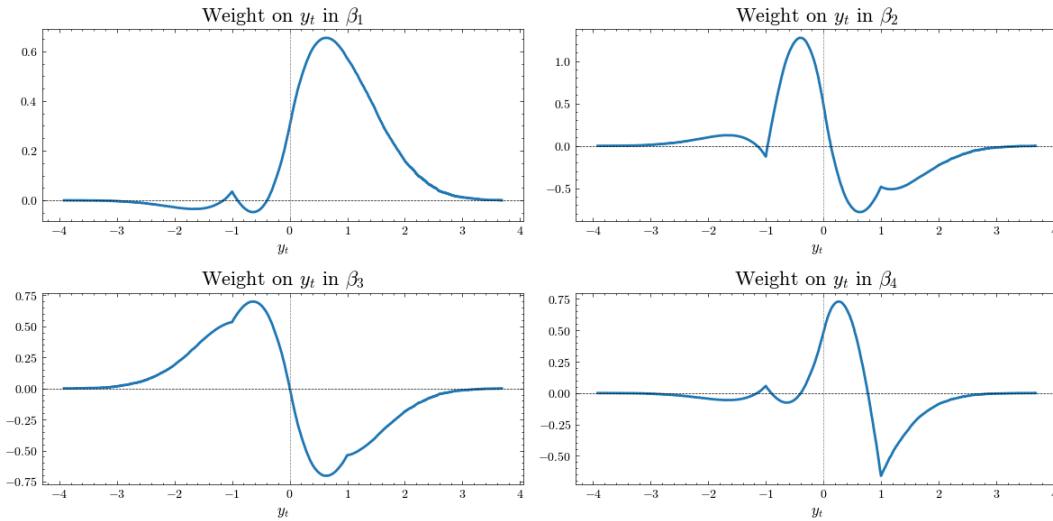


Figure 5: Standard Normal Shock with [Example 5](#)

Overall, one needs to take care when using including indicator functions in any form for a regression. In addition, "exploiting" negative weights is difficult. In the case of [Caravello and Martínez Bruera \(2024\)](#), the weight cancellation underlying their identification result may not occur in finite samples, even with a symmetric shock. For indicator functions, the weights do not follow a straightforward pattern, so it seems more reasonable to focus on estimating positively-weighted averages. [Section 3](#) builds on these lessons and formalizes a path forward.

<sup>7</sup>Excluding  $\varepsilon_t$  as a regressor only exacerbates the issues with [Example 5](#).

### 3 Uncovering Nonlinearities

Section 3.1 explicitly lays out ideal criteria for functional regressors to satisfy for nonlinearity testing. Section 3.2 demonstrates that disjoint indicator functions do a surprisingly good job but have some limitations. Section 3.3 and 3.4 show alternative methods: orthogonal generated regressors and deep learning. Section 3.5 compares all methods while also highlighting how some practical concerns (e.g., measurement error) affect estimation.

#### 3.1 Objective

Formally, we are interested in the effects of a shock  $\varepsilon_t$  on an outcome  $Y_{t+h}$ . Recall from (4) if a collection of shock functions  $\{f_i(\varepsilon_t)\}_{i=1}^N$  is included in a regression, the weights in the estimand on  $f_i$  are  $\omega_i(a) = \frac{\text{Cov}(\mathbf{1}_{\{a \leq \varepsilon_t\}}, X_i^\perp)}{\text{Var}(X_i^\perp)}$ , where the superscript  $\perp$  denotes a projection residual à la Frisch-Waugh-Lovell. Suppose  $\varepsilon_t$  is a shock continuously distributed on  $I \subset \mathbb{R}$ . The explicit objective is to find functions  $\{f_i\}_{i=1}^N$  corresponding to a partition  $\{I_i\}_{i=1}^N$  of  $I$  with their weights  $\{\omega_i\}_{i=1}^N$  satisfying the following targets

- (no negative weight)  $\omega_i(a) \geq 0 \forall a \in I$
- (relevant weight)  $\omega_i(a) > 0 \implies a \in R_i$ , where  $I_i \subset R_i$
- ("hump-shaped")  $\exists$  a peak  $c \in I_i$  s.t  $\omega_i'(a) \geq 0$  on  $(-\infty, c]$  and  $\leq 0$  otherwise

From the second tenet, we say  $f_i$  "corresponds" to  $I_i$  if weight is only placed in a predefined region  $R_i$  nesting  $I_i$ . As discussed in the introduction, it's hard to mix qualitative categorizations of interest (e.g., "big, positive shocks") with quantitative cutoffs. The most practical way to define  $R_i$  is to set a boundary where there is definitely no correspondence. For example, if a function is designated to capture the effects of big, positive shocks, a reasonable baseline would be that no weight is placed on shocks less than 1 standard deviation away from its mean.<sup>8</sup>

The rest of this section will give 3 approaches for satisfying the targets and then compare them. An intuitive guess of disjoint indicator functions turns out to work well. To try to do even better, two other approaches, orthogonal generated regressors and deep learning, are discussed. But as seen in the simulation evidence and Section 4, sample size limitations restrict the gains for moving to the more technical approaches.

#### 3.2 Disjoint Indicator Functions

Disjoint indicator functions feel like they would hit the weighting targets, but Section 2 revealed two necessary conditions: not interacting the indicators with the shock and including all regions of interest. To see this, we will add a final example with standard normal shocks. One really nice property of using regressions to detect nonlinearities is the implicit weighting is invariant to the outcome variable. So really, we don't need a model to

---

<sup>8</sup>So  $R_i = \{a \in I | a > 1\}$  if the shock is standardized. Once  $R_i$  is set, choices for the partitioning of  $I$  follow naturally.

evaluate the weights, just a time series for the shock process. So when we consider

**Example 6:** 
$$\pi_t = \alpha - \beta_{\text{small, neg}} \cdot \mathbb{1}_{-y_t \in [\frac{1}{100}, \frac{5}{4}]} - \beta_{\text{big, neg}} \cdot \mathbb{1}_{y_t < \frac{5}{4}} + \beta_{\text{small, pos}} \cdot \mathbb{1}_{y_t \in [\frac{1}{100}, \frac{5}{4}]} + \beta_{\text{big, pos}} \cdot \mathbb{1}_{y_t > \frac{5}{4}} + \epsilon_t$$

note that the weight plots in [Figure 6](#) are the same no matter the left hand side outcome variable. The motivation for this form is to set reasonable cutoffs for big and small shock magnitudes (e.g., for standard normal,  $y_t = 1$  is a standard deviation and so on). The broader structure seeks to distinguish the effects of both size ( $i = \{\text{big, small}\}$ ) and sign ( $j = \{\text{positive, negative}\}$ ). To test for specific size effects, the null hypothesis is  $\beta_{\text{big},j} = \beta_{\text{small},j}$ , for sign effects it's  $\beta_{i,\text{pos}} = -\beta_{i,\text{neg}}$ , and for general effects a joint test can be used. In over 99.9% of  $n = 300$  simulations of the DGP (6) using the same parameterizations as [Example 2](#), when size or sign effects are present, the appropriate nulls of no effect are rejected. [Figure 6](#) shows weight plots for this specification.<sup>9</sup> The reason for the drastic improvement in performance is evident: the weights here are much further along in converging and also are more directly placing weight where desired. Still, this is not perfect – "big shock" estimates put significant weight on smaller values. But overall it's clearly beneficial to have everything work through a single regression, where each region of interest has its own corresponding estimand. The point estimates themselves are also more revealing, as taking the difference in coefficients provides an indication of how quickly a linear approximation would diverge.

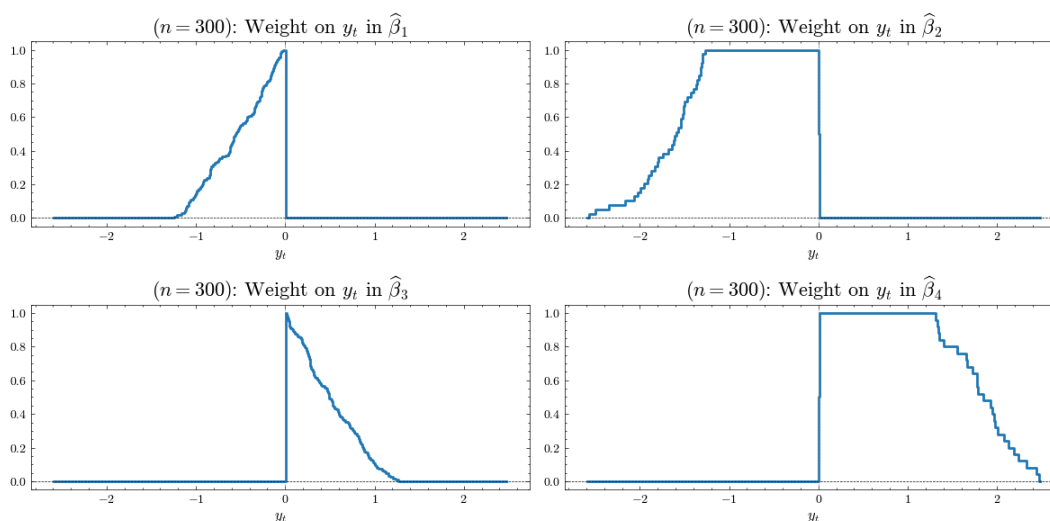


Figure 6: Standard Normal Shock with [Example 6](#)

So now to put [Figure 6](#) in a bigger context: [Section 2](#) discusses the identification result of [Caravello and Martínez Bruera \(2024\)](#), which is essentially that if the shock is symmetric, a population hypothesis test with their proposed specifications will have no false positives. Disjoint indicator functions can achieve the same property under any shock distribution, as well as more descriptiveness about the type of nonlinearity and confidence about false positive robustness in finite samples. We can formalize all the above into a proposition.

<sup>9</sup>Like [Figure 2](#), this will vary across simulations, but the variance here is concentrated exclusively at the endpoints for the big shock weights.

**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  $0 \in O_0$ ,  $O_0 \cup (\cup_{i=1}^N I_i) = \mathbb{R}$ , and  $O_0 \cap (\cup_{i=1}^N I_i)$  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  $\mathbf{X}_t^f$  has full rank,  $x \in I_i \iff f_i(x) \neq 0$ , 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:  $f_i^\perp(\varepsilon_t)$  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  $\varepsilon_t$  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 (2024)* 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  $\alpha_i = \frac{\text{Cov}(\varepsilon_t, f_i^\perp(\varepsilon_t))}{\text{Var}(f_i^\perp(\varepsilon_t))}$ , and let  $\mathbf{X}_t$  be their concatenation. If we project  $Y_{t+h}$  on  $\mathbf{X}_t$  (and a constant and control set as in (2)), then  $\beta_i = \beta_j \forall i, j$  if  $m_h(\cdot)$  is linear in  $\varepsilon_t$ . 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  $\varepsilon_t$  on  $(\inf\{S_{ij}\}, \sup\{S_{ij}\}) \cap I$ .

In plain terms: if the DGP is linear on the space where the weights on  $\beta_i$  and  $\beta_j$  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 6](#) 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 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 decreases. More discussion is in the rest of the paper and [Appendix A.1](#), 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 2 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

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

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. In [Example 5](#), the center set was 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.

Besides the implications for hypothesis tests, this structure is appealing because coefficient differences can be informative about the extent of the nonlinearity in practice. Underlying [Proposition 1](#) is that the implied weights when using indicator functions will be non-negative. In the linear case where marginal effects are constant ( $m$ ), a weighted integral will always be  $m$  multiplied by "total weight". So two estimands meaningfully differ *only if* there is heavy nonlinearity, provided total weight is the same. As with [Proposition 1](#), the converse is not true;  $\beta_i, \beta_j$  being similar does not imply a lack of nonlinearity. But this approach still possibly allows for something to be said about the intensive margin. One thing to keep in mind for this interpretation is that as  $n \rightarrow \infty$ , a hypothesis test will always reject a null hypothesis that two estimates are the same even if the difference is negligible. [Section 4](#) gives some ways to gauge if the estimates of coefficient differences are qualitatively meaningful.

While the false positive result is valuable, there is a risk of false negatives because of weight overlap. There's no reason to think the best of both worlds is impossible, but no alternatives immediately come to mind. To do this, the form of the regression weights must be confronted directly. They can be represented compactly with the help of the Frisch–Waugh–Lovell Theorem, but as detailed in [Appendix A.4](#), they are more precisely a complex non-linear combination of the shock's variance and the covariances of  $\{f_i(\cdot)\}_{i=1}^N$ . We can conceptualize our objective as picking functions to minimize deviations from the weight targets subject to what one might call cross-equation restrictions the functions must abide by. Our two paths forward are either to make these dependencies somehow not matter or use a complex procedure that somehow respects them. The rest of this section will detail two approaches, orthogonal generated regressors and deep learning, one for each path.

First, we can target collections of functions that are uncorrelated with each other. This simplifies the problem tremendously and also makes transparent how to make the weights hit the targets. However, the simplicity comes at the expense of having functions that vary by sample because they are defined in terms of a shock's empirical distribution. Ideally, we would choose a set of fixed functions that perform well across simulations. But this is only possible if we lift the 0 correlation restriction, opening the door to inscrutable dependencies across function. This creates a problem suitable for deep learning, which can finesse through the entanglement constraints to yield the weighting we want. Both approaches have appeal and will be given a detailed treatment. One tension that will emerge is a tradeoff between specificity and variability. Take the earlier example with  $f_1(y) = -\mathbb{1}_{y < -b}$  and

$f_2(y) = y \cdot \mathbb{1}_{y>b}$ , which again involves some weight placed on  $[-b, b]$ . But we can't simply take this floor to 0 to get rid of the unwanted weight because of collinearity, and milder relaxations themselves will cause standard errors to grow. There is a parallel difficulty with moving away from the indicator functions. I find that, under realistic sample sizes, the push to reduce false negatives may come at too high of a cost to standard errors (and thus not be able to say anything). Since limitations will be setting-dependent, they are still worth exploring.

### 3.3 Orthogonal Generated Regressors

Again consider the premise of a shock  $\varepsilon_t$  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  $\varepsilon_t$  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

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

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  $c \in (0, 1)$ ,  $\varepsilon_t$  has probability mass  $F(c) - F(0)$  on  $[0, c]$  and mass  $F(1) - F(c)$  on  $[c, 1]$ . Consider<sup>10</sup>

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

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.  $\int_a^\infty f_i(x) dF(x)$  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  $\int_I \omega_i(a) da = 1$ .

<sup>10</sup>The has some precedent in applications of the Haar wavelet (Mallat, 1999).

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.

### 3.4 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

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

Appendix A.2 shows the integral of the weights in  $\beta_1$  is proportional to

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

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  $\beta_1$  represents a positively weighted average of marginal effects.<sup>11</sup> Equally, we need the analogous expression for  $\beta_2$  to be positive. The simplest path to joint satisfaction is the functions are correlated with  $\varepsilon$  yet uncorrelated with each other. As the number of function grows, the potentially paradoxical paths become more unwieldy. For the second goal, we know from (4) the weights in  $\beta_1$  will be large where  $\varepsilon_t$  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 Example 6. 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

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

<sup>11</sup>Though recall this is not a sufficient condition on its own, as many of the examples in Section 2 show.



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 12) roughly look like (chronologically)

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

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

However, actually using these functions fails spectacularly; notice the approximations for  $f_1$  and  $f_3$  are highly collinear. It turns out the 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  $\mathbf{x} \in \mathbb{R}^k$  that minimizes  $F(\mathbf{x})$ . 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  $\{f_i(\cdot)\}_{i=1}^4$ . 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 B.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, 2024).

### 3.5 Simulation Performance

This section will assess performance across a variety of data generating processes. To give a preview of the prevailing takeaway from this section, first the DGP in (6) from Example 2 and Example 6. For the case of  $b = 2$  with standard normal shocks (and sample size  $n = 300$ ), a deep learning approach only offers a modest

improvement over the [Caravello and Martínez Bruera \(2024\)](#) approach, failing to reject a null hypothesis of no size effects in 26% of simulations. The generated regressor approach fares even worse, correctly rejecting in only 60% of simulations – the methods developed to produce less false negatives failed to do so. This is because, while the weights look more appealing (see [Figure 11](#) in Appendix A.5), the variance of the coefficients limits the usefulness of this property. For the generated regressor approach, the functions corresponding to big shocks in particular have large variance, in part because less of the sample is concentrated there. For deep learning, the aforementioned idiosyncrasies the neural network creates to respect the constraint set (see [Figure 12](#) in Appendix A.5) also create more variance. And to re-emphasize – these occur even with DGP (6), which outside of a single kink, is about as vanilla as it gets (outcome driven entirely by two i.i.d shocks with no autocorrelation). This is an indication that disjoint indicator functions should be the default method of choice, though it’s still useful to report results from all.

One last, previously omitted case to discuss before moving on from DGP (6) is  $b = .5$  (square-root). While the indicator functions perform the best in this environment, nulls are not rejected in nearly 40% of cases, with the other methods performing far worse. This presents a stark limitation – an important nonlinearity of economic interest would be diminishing returns to scale of policy intervention. However, these are intrinsically harder to detect, as the second derivative of the square-root function is essentially constant after moving away from 0. This is an unfortunate downside the framework is not as well-equipped to handle.

To get a more wholistic picture of performance, we now consider a richer class of data generating processes. In particular, [Li et al. \(2024\)](#) assess the finite sample tradeoffs between local projections and vector autoregressions by making thousands of random selections of 5 variables from the [Stock and Watson \(2018\)](#) dataset and fit a dynamic factor model to create a DGP (and back out the structural shocks). I use this as a starting point and compare all the above methods against a multitude of flavors for DGP. Because each DGP has its own structural shock, it’s not feasible to train a neural network on each one. I find the point estimates are very similar to the disjoint indicator approach, which is not surprising given the functional forms it consistently converges to are broadly well-approximated by combinations of indicators linear functions. For the generated regressors, construction here is more feasible, albeit requiring a sorting procedure. Something slightly different from what’s described in [Section 3.3](#) can be used – instead of making the weight peak  $c_i$  an arbitrary point, we can make it the sample median, allowing for the function to be normalized. This is discussed in more detail in [Appendix A.3](#), along with the necessary standard error correction (which is negligible in practice).

I modify a threshold-VAR model from [Loria et al. \(2025\)](#) who argue it captures some fundamental macroeconomic dynamics. The structure is centered around 3 components: growth of real activity  $y_t$ , a financial factor  $f_t$ , and a macroeconomic factor  $m_t$ . I keep this same system of 3 equations but add inflation  $\pi_t$  and additional fundamentals  $W_t$ . I use the [Stock and Watson \(2018\)](#) dataset and randomly select a series from the relevant group of variables for  $y_t, f_t, m_t$ , and  $\pi_t$  and randomly select other variables from the remaining categories for  $W_t$ . The procedure detailed in [Li et al. \(2024\)](#) is used to generate a structural monetary policy shock  $X_t$  for this system

using a dynamic factor model representation. The skeleton of the threshold-VAR DGP is

$$\begin{aligned}
y_t &= \beta_0 + \beta_1 f_t + \beta_2 m_t + \beta_3 \pi_t + \beta'_w \mathbf{W}_{t-1} + \epsilon_t^y \\
f_t &= \alpha_0 + \alpha_1 f_{t-1} + \alpha_2 m_t + \alpha_3 g(x_t) \cdot \mathbb{1}_x(f_{t-1}, m_{t-1}) + \epsilon_t^f \\
m_t &= \gamma_0 + \gamma_1 m_{t-1} + \gamma_2 f_{t-1} + \gamma_3 g(x_t) \cdot \mathbb{1}_x(f_{t-1}, m_{t-1}) + \epsilon_t^m
\end{aligned} \tag{7}$$

where  $g(x_t)$  is some non-linear function of the shock and  $\mathbb{1}_x(f_{t-1}, m_{t-1})$  is a state-dependent multiplier. In the baseline calibration, I set  $\mathbb{1}_x(f_{t-1}, m_{t-1}) = 3$  if the financial and macroeconomic factors are both negative and equal to 1 otherwise. I estimate the other parameters in this model using  $g(x_t) = x_t$  and omitting the state-dependence. Then I simulate a time series for  $y_t$ ,  $f_t$ , and  $m_t$  using (7) with a few different choices for  $g(\cdot)$  and the data for  $\pi_t$  and  $\mathbf{W}_t$ . For each choice of  $g(\cdot)$ , I run several local projections at horizon  $h = 0$  for the different approaches for detecting nonlinearities. The LPs have 196 observations and include 4 lags of all variables except  $\mathbf{W}_t$ , which is not included at all to mimic the presence of omitted variables. The results are averages across 100 variations of (7) with 10,000 simulations each. Inference is performed with Huber-White standard errors (Montiel Olea et al., 2024) and results are materially the same using more involved estimations of the variance-covariance matrix (Xu, 2023).

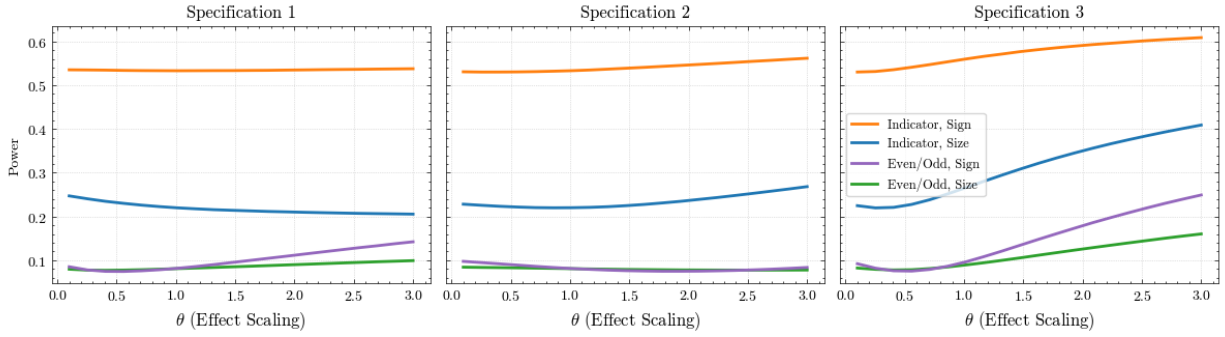
Figure 7 plots the power of hypothesis tests using the indicator function approach and the even/odd weight decomposition of Caravello and Martínez Bruera (2024) across 3 specifications for the non-linear shock function  $g(x_t)$  that feature both size and sign effects. The point of the plots is to show how power changes as we scale a component of  $g(\cdot)$  by  $\theta$ . With a foundation of  $c \cdot \mathbb{1}_{x \geq c} + x \cdot \mathbb{1}_{x < c}$ , the first specification has the first term scaled by  $\theta$ , likewise for the second specification and the second term. This can be thought of as two ways of adjusting a jump then plateau of effects. The third specification is  $\theta x^2 \cdot \mathbb{1}_{x \geq c} + x \cdot \mathbb{1}_{x < c}$ .<sup>12</sup> The shock is standardized and the cutoff  $c$  is set to 1 in the DGP. Across simulations, the structural shocks follow a roughly but not perfectly symmetric distribution, making the even/odd approach a valid choice ex-ante.

The plots show that the indicator function approach strictly dominates the even/odd decomposition, though the gap is decreasing in the size of these specific nonlinearities. The Caravello and Martínez Bruera (2024) procedure mostly dominates the generated regressor approach for the different parameterization, so it's excluded for clarity. The indicator function approach has the advantage of insignificant coefficients not being the end of the story; the estimates may be different enough that a null of linearity can be rejected. In principle, this advantage should extend to the generated regressors, but their unconventional construction clearly leads to even more inefficiency.

---

<sup>12</sup>The first two specifications were selected to give all methods as difficult task, with the third being a more traditional non-linear structure. The results for the first two specifications are similar if we multiply  $x$  through their corresponding  $g(\cdot)$ .

Difference in Power: Indicators vs. Even/Odd Approach



Size for the indicator function approach refers to a rejection of the null hypothesis that the coefficient on the big, positive shock and the small, positive shock are the same. Sign refers to a rejection that the big, positive and big, negative coefficient is the same. Even/Odd are simple significance tests of the coefficient on  $f(x)$ .

Figure 7

In every previous simulation, we have assumed the structural shock  $\varepsilon_t$  is perfectly observed. This will obviously not be the case in practice. As mentioned in Section 2, if the structural shock is not observed the weights are unknown. Assuming that a proxy  $z_t$  is in fact the structural shock represents the ceiling on estimation quality. But this says nothing about how useful what we're actually estimating is. [Kolesár and Plagborg-Møller \(2024\)](#) show that if controls are needed for identification, we cannot have any faith in what we're estimating unless the propensity score is linear. Instead, the proxy should indeed be a proxy in a classical sense – departures from the structural shock amount to noisy measurement. This is merely a conjecture; measurement error itself can be complicated, even if induced from noise alone. [Chen et al. \(2011\)](#) show that the usual "attenuation bias" relationship to estimands does not hold if the measurement error is non-linear. Thankfully, I find in simulations that under a rich set of measurement error types (e.g., nonlinearities, heteroskedasticity, state-dependent noise) the "best case" weights are good approximations for the true weights. Even though the bias can go in either direction, that does not matter for the hypothesis testing procedure. Using a simulation of the first specification of DGP (7) with structural shock  $x_t$ , [Figure 8](#) plots the case of  $z_t = \text{sgn}(x)(x + \epsilon)^2$  where  $\epsilon$  is normally distributed, mean-0 noise with conditional variance  $.01^2(1 + x^2)$ . If we run local projections using  $z_t$  and plot the weights as if  $z_t = x_t$ , the true weights are similar. The indicator function structure constrains deviations; out of several combinations tried, this was about as ugly as it got. One specific area of concern is the true weights are putting much more weight of shocks on the "wrong sign". But the consequences are limited to being less robust to false negatives.<sup>13</sup>

<sup>13</sup>This is another reason to use disjoint indicators. While it forces all coefficients to put some weight on the wrong sign, the values of  $\varepsilon_t$  where "wrong-sign weight" is placed will be the same (within the the coefficient group). If we allow for overlap, we could have one coefficient with much more wrong-sign weight than the rest, and we can't know if it's an issue without the exact form of measurement error.

True Weights vs. "Best Case"

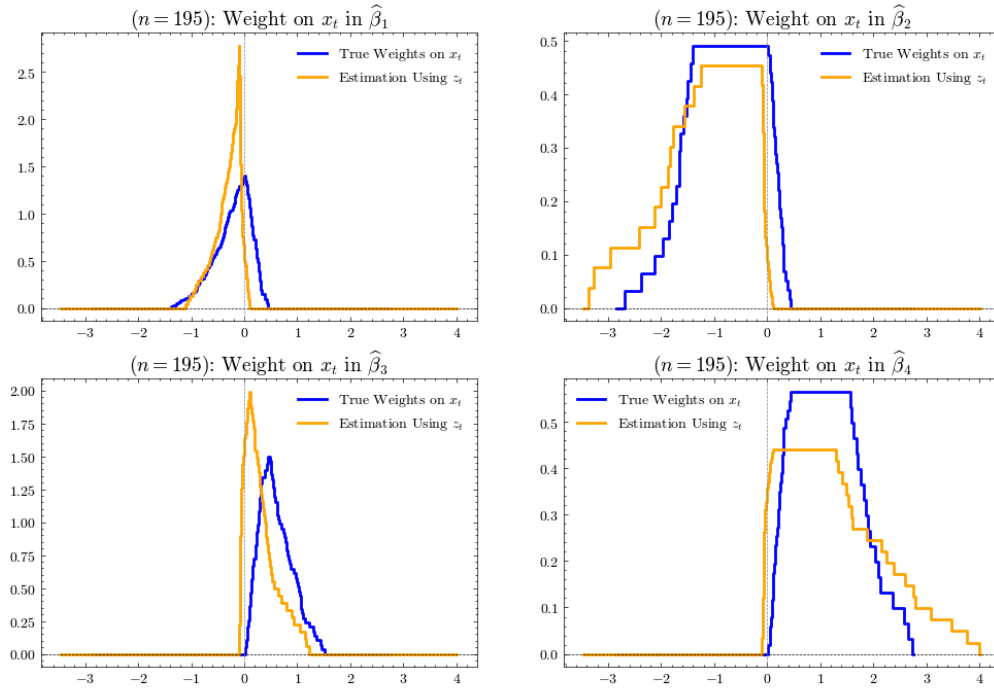


Figure 8: Selected Simulation of DGP (7) with Measurement Error

Finally, one issue that was not highlighted in the previous discussions is the issue of choosing thresholds. It's not ideal to have to mandate when a small shock becomes big etc. The first step to address this is to first standardize the data and then make the partitioning based on data realities. The threshold for magnitude in positive shocks need not be the same as for negative shocks. That should give an inkling as to a reasonable baseline to set, and then robustness exercises could involve moving this threshold around to see if results are sensitive to cutoffs. Interestingly, Figure 6 shows that while it *feels* like indicator functions involve setting paradoxical thresholds, there is a significant amount of weight overlap, so this is actually not as much of an issue. In fact, to decrease the amount of false negatives, one ironic way to address this is to allow indicators themselves to overlap. For example, the weights when using  $f_1 = \mathbb{1}(\varepsilon_t < .01)$  and  $f_2 = \mathbb{1}(\varepsilon_t < -1.5)$  have significantly less overlap than the disjoint case. The tension as mentioned before is this increases standard errors. On the other hand, for generated regressors, there is no weight overlap and this paradox is unavoidable, though it can be diminished by fixing the peak at the median of the interval instead of having to choose than beforehand as well.

## 4 Application

[Section 4.1](#) outlines consideration for estimation based on developments in the local projection literature and the implications for shock series selection from the "regression anatomy" presented in [Section 2](#). [Section 4.2](#) applies the preceding guidance for choosing a monetary shock proxy. Because many great papers and people have been committed to the topic, the discussion is specialized and may not be of interest to general readers. Likewise, interested readers may be disappointed that some of the arguments are not fully fleshed out, which I leave to separate work. [Section 4.3](#) shows the nonlinearity detection results and briefly outlines an attempt to match them with a non-linear equilibrium model.

### 4.1 General Best Practices

The results in [Section 2](#) rely on using the actual structural shock  $\varepsilon_t$  in the regression. In practice, we will have some shock proxy  $z_t$ . [Kolesár and Plagborg-Møller \(2024\)](#) note that if  $\varepsilon_t$  is unobserved, the form of the weights are unknown, so plotting the weights under the assumption that  $\varepsilon_t = z_t$  provides the "best case scenario". The previous section illustrated that if noise alone causes the divergence between  $\varepsilon_t$  and  $z_t$ , then the weights using the proxy are actually a good approximation. There are typically many shock series researchers can choose from, but the regression anatomy representation provides a clear first-order selection criteria:  $\varepsilon_t \approx z_t$ . Because we're working with finite samples and shocks will often be small in size, it's important to note that  $\approx$  here means the bias is purely idiosyncratic. This raises concerns about identification approaches that either rely on some "selection on observables" assumption or estimation in general. Namely, [Kolesár and Plagborg-Møller \(2024\)](#) show that if a proxy is only exogenous conditional on some control set, the weights will depend on the control set (and will have negative weights if there are non-linear relationships).

A complication of the guidelines for assessing the quality of proxies is there's currently no adequate sensitivity analysis procedure. Including controls can improve the efficiency of estimates, even though controls themselves have no effect on a shock's regression estimand, but there is a natural concern that certain control variables could drive the results. In the language of the regression anatomy framework, the concern is the estimand using  $z_t$  is not a useful object because some propensity score squirreliness makes the regression weights badly behaved. [Kolesár and Plagborg-Møller \(2024\)](#) recommend dropping various controls and seeing if results change, a strategy ubiquitous in appendix robustness checks. Unfortunately, these results may be misleading. Under a special case of DGP (7) where shocks enter linearly, point estimates in the regression of  $y_t$  on  $\varepsilon_t$  vary wildly across samples and when different control sets are used. So we cannot distinguish between variation indicating sensitivity to controls or sensitivity to sample size. This raises a broader point: another common robustness check is to redo the main analysis with different shock proxies and see if the results change. But because these proxies are constructed so differently (see, e.g., [Brennan et al. \(2024\)](#) for comparisons of monetary policy shock series), in general there's

no reason the implied weights and therefore results should be similar. Until better sensitivity analysis tests are available, the best strategy is to have a convincing argument  $\varepsilon_t \approx z_t$ .<sup>14</sup>

Once a shock series has been selected, there are other considerations with running local projections (LPs). [Ramey \(2016\)](#)'s handbook chapter provides a good starting point, but there is also important recent work. Inference remains an open question. Newey-West standard errors have been shown to be biased, while Huber-White with lagged control variables is often sufficient ([Herbst and Johansen, 2024](#); [Montiel Olea and Plagborg-Møller, 2021](#)). [Xu \(2023\)](#) shows a more involved estimation of the variance-covariance matrix that is robust in a larger class of DGPs. There have also been methodological advances, namely smooth local projections ([Barnichon and Brownlees, 2019](#)), which uses penalization to salvage the appealing properties of LPs while increasing their efficiency and delivering smoother coefficient plots, and Bayesian local projection ([Ferreira et al., 2024](#)). A great deal of work has been done to clarify the differences between LP and vector autoregression (VAR).<sup>15</sup> ([Plagborg-Møller and Wolf, 2021](#)) prove LP and VAR are asymptotically equivalent in the limit (if lag order is high enough). So even researchers who prefer VAR estimation should run the LP analogue and plot the weights to have a better sense of what is being estimated. While there will be finite sample differences ([Li et al., 2024](#); [Montiel Olea et al., 2024](#)), plotting the weights in the LP will still be informative. Borrowing from [Kolesár and Plagborg-Møller \(2024\)](#)'s example, if all the weights on a government spending shock are being placed on positive values, what estimation is actually uncovering are the effects of spending buildups.

To conclude, the basic implementation chronology is the following. First, select and standardize a shock series  $z_t$  based on the confidence in approximating the underlying weights on  $\varepsilon_t$ . In the default case, the next step is to identify  $N$  regions of interest along the shock's support to create disjoint indicator functions. This paper emphasizes choice of  $N = 4$  because of sample size limitations. [Section 2](#) reveals that this procedure is not equivalent to including only 3 indicators and testing effects relative to the fourth, so instead define the indicators so that they exclude a small interval around 0. Further, [Proposition 1](#) emphasizes "normalizing" the indicator functions so that indicators on the negative real line enter the regression with a negative sign attached. For each  $f_i$  in the collection, define a new function  $g_i$ , where  $g_i = \alpha_i f_i$  and  $\alpha_i$  is the coefficient on  $f_i$  in a regression of  $z_t$  on  $\{f_i\}_{i=1}^N$  and a constant. Then run local projections of the outcome variable on  $\{g_i\}_{i=1}^N$ , a control set, and a constant. In the regressions, include a healthy number of lags and use Huber-White Standard Errors for inference, or a more involved estimation routine of choice (e.g., [Xu, 2023](#)). [Appendix A.1](#) shows the delta method correction for these generated regressors on the diagonal of the variance-covariance matrix is the square of the second stage coefficient divided by the first stage t-statistic, which will generally be negligible in practice. On the off-diagonals, the first stage covariance will also show up as  $\frac{\beta_i \beta_j}{a_i a_j} \text{Cov}(a_i, a_j)$ .

<sup>14</sup>The proxy should only be tied to one structural shock. If not accounted for, [Koo et al. \(2024\)](#) show inference will be incorrect. For proxies with many 0 values, finite sample correlation is inherent, see [Barnichon and Mesters \(2025\)](#) for discussion and a solution for narrative proxies.

<sup>15</sup>[Li et al. \(2024\)](#) confirm the "bias vs. efficiency" conjecture, but [Montiel Olea et al. \(2024\)](#) reveal the cost of efficiency gains can be prohibitive: VARs are comfortably insensitive to misspecification if and only if the relevant estimate has similar variance to its LP analogue.

The preceding paragraph outlines the "default" case, but there are other potentially useful considerations. For the initial function collection, [Section 3](#) showed two additional approaches (generated regressor and deep learning), but may be limited by the sample size. It should also be noted these two approaches are purely means of carrying out hypothesis testing, and a unit change in these variables does not have any economic interpretation. Because indicator functions do not face the same limitation, one option for the primary illustration of nonlinearities is to use penalized local projections. One drawback is inference is complicated by regularization (and possible cross-validation). To guard against potential bias in variance-covariance estimation, [Barnichon and Brownlees \(2019\)](#) recommend computing using standard errors from an even more "under-smoothed" estimator. To go a step further, in the application in [Section 4.3](#), I also fix the penalty parameter at a mild level before estimating and use 99% confidence intervals. Lastly, there's a point to be made about when nonlinearities actually matter. For example: suppose marginal effects for positive shocks are  $\beta$  and for negative shocks  $\beta + \varepsilon$ . A population hypothesis test will reject a null hypothesis of linearity, even though a linear model is appropriate. If the indicator functions are normalized so that their individual weights integrate to 1, coefficient differences can give some insight into whether the degree of nonlinearity matters because they have a reasonable interpretation as a difference in means. For the application in [Section 4.3](#), this translates to measuring the nonlinearity in terms of difference in percent change in outcome since the shock occurred.<sup>16</sup>

## 4.2 Selecting a Monetary Shock Series

To select a series for assessing possible nonlinearities in the transmission of U.S. monetary policy, we have to address a question for which there is surprisingly not a straightforward answer: what is a (structural) monetary policy shock? Unless one is willing to argue that central banks have a systematic way to set rates they decide arbitrarily to deviate from, which seems like a poor description of an institution like the Federal Reserve and its army of economists, monetary shocks are changes in policy unanticipated by private agents. This makes the high-frequency measures of forecast errors backed out from price changes in futures markets a natural choice.

Within the class of high-frequency measures, there are several options. [Bu et al. \(2021\)](#) is currently popular because of its ability to easily handle the zero lower bound period by creating a single measure to represent shocks across the entire yield curve. At the same time, their measure cannot be easily mapped into a candidate data generating process, so it's less clear would be estimated ([Brennan et al., 2024](#)). Another issue is that because private agents do not know perfectly the central bank's reaction function and there may not be a single information set for all agents, changes in futures markets may be representing combinations of multiple structural shocks, which is a challenge. There are several measures which look at changes to expected future interest rates, rather than the

---

<sup>16</sup>To avoid haggling over what constitutes meaningful nonlinearity, one option is to normalize by the linear estimate's standard deviations, so coefficient differences are still in units of effect sizes but in some sense have an interpretation similar to t-statistics (i.e., gesturing towards the likelihood parameters were drawn from the same distribution). Results are also more easily comparable to DSGE model output by minimizing unimportant scaling distortion from finite sample properties of time series and model-simulated data. Details are in Online Appendix B.6.



current period, and try to decompose them into "forward guidance" vs. "information shocks" (e.g., [Jarociński and Karadi, 2020](#)), but because these measures are estimation-specific, there is a risk that the deviation from structural shocks is systematic or sample-dependent, rather than pure noise. Instead, sacrificing performance at the zero lower bound and looking at changes to the Fed's expected change to its target in the current period seems to be the most practical option. All concerns about the possible tangling of effects from forward guidance, information, credibility, preferences, etc are moot when looking at the current period because once an action is announced, the adjustment is not function of ambiguity about any of those things because the Fed chair has essentially written the futures price correction in stone.<sup>17</sup> This leads to a selection of the [Jarociński \(2024\)](#) MP1 series, originally developed by [Kuttner \(2001\)](#), as the proxy of choice.

Before moving onto discussing other approaches more in-depth, it should be noted there are many concerns specific to the high-frequency series. When outcome variables are not also high-frequency, [Jacobson et al. \(2024\)](#) warn of temporal aggregation bias because the Federal Reserve's meeting calendar fluctuates and sometimes multiple shocks occur within the same month. Absent getting better data, the best response is likely to not put much stock in the results at the shortest horizons. [Casini and McCloskey \(2024\)](#) also point out that using a narrow observation is not actually a magic identification wand, though they show the [Nakamura and Steinsson \(2018\)](#) measure is relatively robust to the potential concerns. A final consideration is that these futures markets are not fully saturated with participants, particularly during the zero lower bound period, and past work has shown that there are arbitrage opportunities available from the apparent predictability of the high-frequency adjustments ([Miranda-Agrippino and Ricco, 2021](#); [Bauer and Swanson, 2023](#)). This concern has rightly been a focal point of the recent literature ([Acosta, 2023](#)), but the results may not be as damning as they seem. Leaving aside that these markets may be innately "inefficient", it seems more likely that was these finite sample results are showing are the effects of heteroskedasticity. When there is more movement in macro fundamentals, it is more likely for central banks to act, thus creating more variance for structural shocks. Heteroskedasticity will not distort the identification results and, as shown in [Section 3](#), does not significantly disturb the utility of proxies.

Another popular method in this literature is projection orthogonalization, or using the residuals from a linear regression. This is the basis for [Romer and Romer \(2004\)](#), who represent the change in interest rates unrelated to the Fed's information with the residuals in a regression of changes in the federal funds rate on Fed forecasts.<sup>18</sup> But the residuals themselves are extremely sensitive to the estimated coefficients, and we should not have faith that  $\varepsilon_t \approx z_t$  – [Cochrane \(2011\)](#) demonstrates this won't occur even in the simplest case where data generating process is linear (a basic New Keynesian model with a Taylor Rule). [Miranda-Agrippino and Ricco \(2021\)](#) and [Bauer and Swanson \(2023\)](#) use orthogonalization by residualizing existing measures of monetary policy shocks to guard against claims of predictability (see [Acosta \(2023\)](#) for a survey). These adjustments will likewise be sensitive to

---

<sup>17</sup>There is risk of contamination in the few instances where there were shocks in the days before the formal announcement of the target.

<sup>18</sup>[Aruoba and Drechsel \(2024\)](#) argue these forecasts don't span the information set. They extend the methodology with text analysis.

the realized OLS point estimates, which really has bite given the sample size. For instance, [Bauer and Swanson \(2023\)](#)'s shocks are based on a 1988-2023 monthly sample. If they had originally done this procedure in 2015, the median percent difference in shock magnitude between the original and "updated" series would be over 100%.<sup>19</sup>

### 4.3 Nonlinearities in the Effects of Monetary Policy Shocks

I look for evidence of nonlinearities in monetary policy transmission by applying the described procedure to the outcome variables of industrial production, consumer price index (CPI), consumption, and unemployment from November 1988 to January 2020 using the MP1 series.<sup>20</sup> I take (log) differences and cumulate them over future horizons so that the left hand side variable can be interpreted as "percent change since the shock occurred". The estimation is done with penalized Local Projections, with standard errors computed as described in [Section 4.3](#) to be over-correct for any potential bias. I find evidence of nonlinearities in each variable. Recall within this paradigm, we define size and sign effects in terms hypothesis tests that compare two coefficients. The visualizations can be thought of as showing the effect of one type of shock relative to the other. For example, a negative sign effect for big shocks and industrial production can be interpreted as the expansionary effect for big negative shocks (on IP) is smaller than the contractionary effect from big positive shocks.

[Figure 9](#) show size effects for positive shocks and sign effects for big shocks. With the exception of unemployment, larger positive shocks have a disproportionately more contractionary effect. Sign effects in this context can be interpreted through the lens of "pushing on a string" (see, e.g., [Fisher, 1935](#)), the idea that the effects of monetary policy are asymmetric because during a downturn there is little central banks can do to create an appetite for lending and spur broader economic activity. The sign effect plots show this narrative matches all variables but unemployment. [Appendix A.6](#) features more visualizations (as well as more details on replication). In particular, there is evidence big negative shocks have a more expansionary effect in the long-run than small shocks ([Figure 14](#)). For sign effects, not much can be said about asymmetries for small shocks ([Figure 15](#)).

The generate regressor and deep learning approaches also point in the same direction. In particular, [Figure 10](#) shows even the point estimates for the machine learning-based estimates are quite similar. It's important reiterate that a unit change in these two sets of functional regressors has no interpretation, even informally. They are merely concocted in a way so that we know the estimands represent a weighted average of marginal effects, but because the weights vary across approaches, they aren't directly comparable to the indicator approach. On the other hand, a unit change in an indicator function has a more direct interpretation, making it more appropriate to ascribe an interpretation of size and sign effects directly to looking at the difference in coefficients. More details on the output from the other are in [Online Appendix B.4](#), such as their corresponding weights. Overall, a picture is painted that is hard to square with standard models: nonlinearities that peak in the medium to long-run.

---

<sup>19</sup>[Sims \(1998\)](#) cautions against scrutinizing shock magnitudes in VARs, which are relative to a given information set. The concern here is distinct. Again,  $\varepsilon_t \approx z_t$  means the bias should be from systematic measurement noise (so shouldn't be mechanically sample-dependent).

<sup>20</sup>More details can be found in [Online Appendix B.4](#).

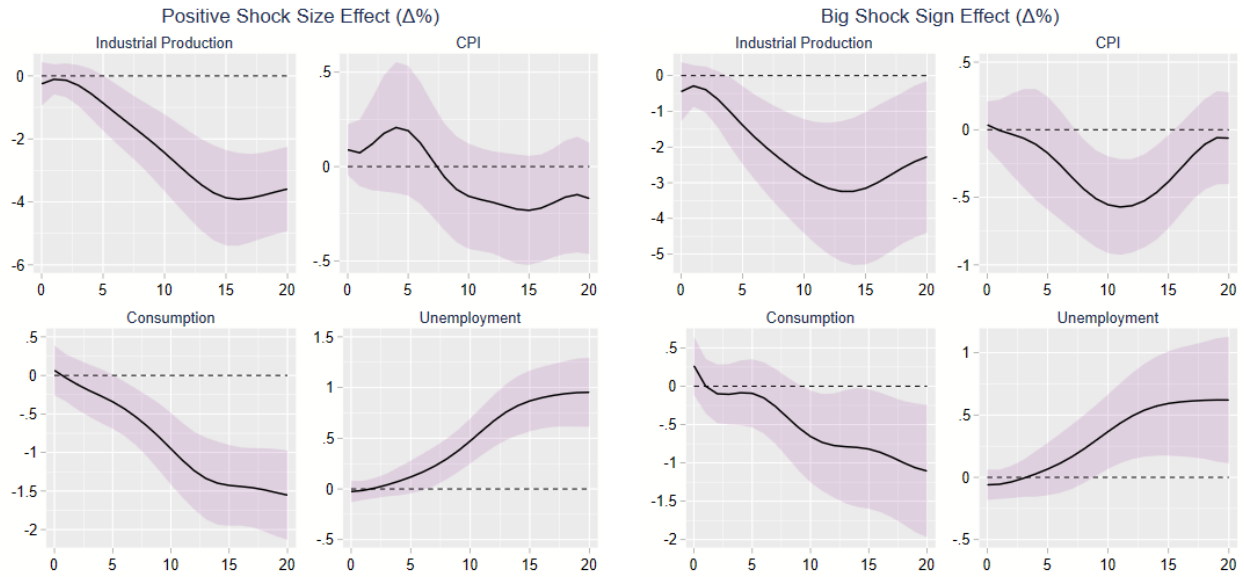


Figure 9: Indicator Approach with MP1 Shocks



Figure 10: Machine Learning Approach with MP1 Shocks

The next step after finding results like this is to try and explain them. To compare to the results from US data, a basic point of reference would be using a model that features meaningful nonlinearities to generate data and then run the same regressions. [Barnichon and Matthes \(2018\)](#) conjecture that sign effects which work in opposite directions for unemployment and inflation, which is what we observed in the last section, can be rationalized in a model with downward-rigid prices and wages ([Kim and Ruge-Murcia, 2009](#)). In this setting, firms seeking to

change its price at a rate different than steady-state inflation face an adjustment cost

$$\Phi_t^p(\pi_t) = \frac{\phi_p}{\psi_p^2} \left( e^{-\psi_p(\pi_t - \pi^*)} + \psi_p(\pi_t - \pi^*) - 1 \right)$$

For  $\psi_p > 0$ , it's more costly to decrease prices than raise them (downward-rigid), for  $\psi_p < 0$  prices are upward-rigid, and the function limits to symmetric adjustment costs as  $\psi_p \rightarrow 0$ . Nominal wage adjustment costs take on the same structure. Past estimation of this model have found evidence of downward rigidity in prices and wages, consistent with empirical evidence dating back to [Keynes \(1936\)](#) and [Tobin \(1972\)](#).

Since the relevance for this paper is largely motivation, I relegate most details about the model and the estimation to Online Appendix B.5. Using the same sample period of US data, the [Aruoba et al. \(2017\)](#) extension of the downward-rigidity model is estimated to second order via a standard random walk Metropolis-Hastings algorithm and particle filter ([Fernandez-Villaverde and Rubio-Ramirez, 2007](#)). I use the distribution of parameters generated by this exercise to simulate data and run the same local projections procedure to create a Bayesian analogs (i.e., using credible sets instead of confidence intervals) for the empirical results. These exercises show (full results in Online Appendix B.5) that this while the model can generate nonlinearities, in general the observed asymmetric effects for both size and sign occur on impact and then quickly dissipate. I also take the posterior mode of all parameters and then vary both asymmetry parameters (one at a time, in both directions, and then both at once in the same direction) while keeping everything else fixed, then simulate data and estimate for each combination. This exercise provides some clarity: on impact, certain combinations of the asymmetry parameters can generate any desired nonlinearities, but it cannot be sustained.

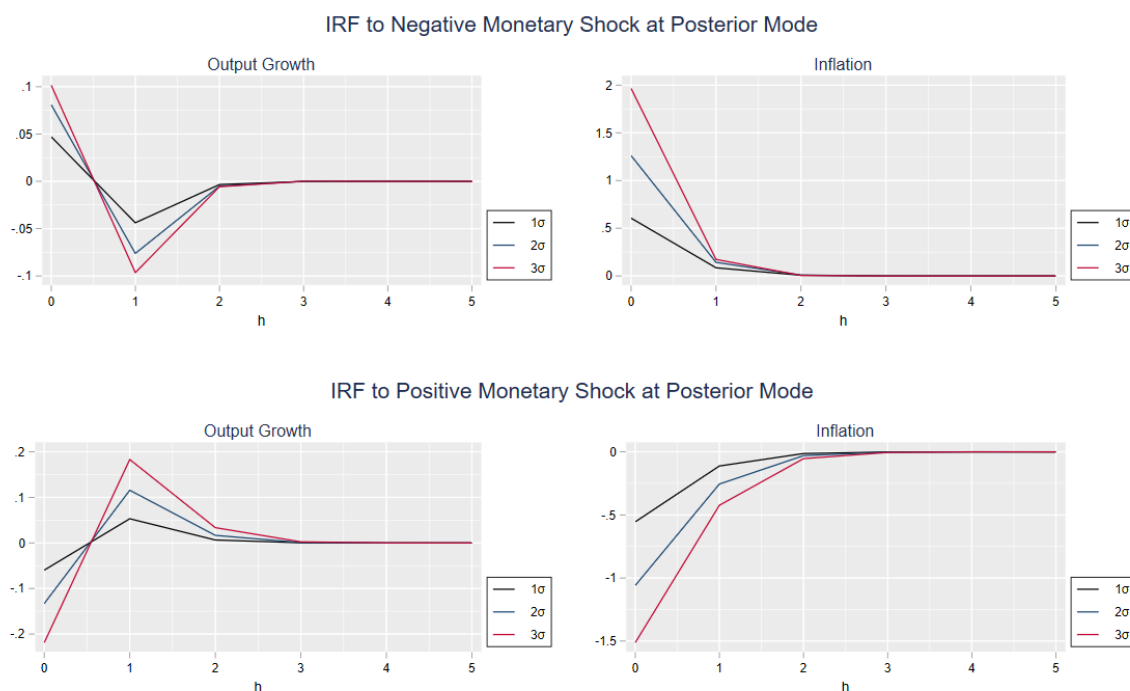


Figure 11

Looking at the impulse response functions directly from the model (rather than running a LP) corroborates the above interpretations. **Figure 11** shows impulse responses for both negative and positive shocks of different sizes. By a horizon of 5 periods after the shock, the magnitude of responses are near or below zero. Online Appendix B.5 discusses various extensions to the model, like adding autocorrelated shocks, that ultimately don't help much.

One reason why the effects of monetary shocks may not have a lasting effect is because of the lack of inertia in interest rate setting. Even though the Metropolis-Hastings produced draws with moderately high persistence in the Taylor Rule (posterior mode of  $\rho_r \approx .67$ ), an inspection of model simulated data reveals that whenever a large monetary shock takes the central bank away from its (nominal) target  $i^*$ , it generally doesn't take long to get back. Table 1 in Online Appendix B.5 shows the results of 10,000 simulations at the posterior mode. For each simulation, I take the median distance between the target interest rate and the current interest rate  $h$  periods after a big change in interest rates (magnitude greater than 10%) and then average across simulations. In periods in which the central bank heavily adjusts the interest rate, the target is relatively far away, but this is almost completely undone 2 periods later. There is also a large asymmetry on impact that quickly becomes less dramatic. Regardless of model, consecutive, large realizations of white noise innovations are unlikely, but the staying power of shocks can vary. These results suggest that the nonlinearities observed in data may not have an explanation in our standard class of models and warrant further explorations for channels in monetary policy transmission. There should also be some broader considerations added in model selection. Linearized general equilibrium models, appealing because of a reduction of analytical and computational complexity, can output sub-optimal normative prescriptions if the economy actually follows a data generating process with strong non-linear components.

## 5 Conclusion

This paper demonstrates a new method to test for nonlinearities in data exploiting properties of least squares regression that are consequences of assumptions about proxies for structural shocks that are commonly made in the applied macroeconomics literature. Three new approaches within this framework were characterized, but the simplest (disjoint indicator functions) seems to be the most useful in practice. While this seems to be yet another example of the power of OLS in spite of its simplicity, there are some limitations that point to more future work. There is a tension that emerges between making the weights appear in the desired places and the efficiency of estimates. The disjoint indicator functions are the most efficient option at the expense of having relatively dispersed weight. So while we can view coefficient differences as a good gauge of deviations from nonlinearity, we cannot interpret the estimands themselves as weighted averages of marginal effects on the areas the indicator functions are active (they are weighted averages over a larger region). It seems possible to expand along this dimension, but it's not immediately clear how. The procedure informed an application to monetary policy shocks, which showed results that are difficult to match even with a general equilibrium model that featured rich size and sign nonlinearities. Results like this can inform paths forward for better understanding the transmission of policy.

## References

- Acosta, Miguel.** 2023. "The Perceived Causes of Monetary Policy Surprises." February, Working paper.
- Angrist, Joshua D., and Jörn-Steffen Pischke.** 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.
- Angrist, Joshua, Kathryn Graddy, and Guido Imbens.** 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equation Models with an Application to the Demand for Fish." *Review of Economic Studies*, 67 499–527.
- Aruoba, S. Boragan, and Thomas Drechsel.** 2024. "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 Christian Brownlees.** 2019. "Impulse Response Estimation by Smooth Local Projections." *The Review of Economics and Statistics*, 101(3): 522–530.
- Barnichon, Régis, and Christian Matthes.** 2018. "Functional Approximation of Impulse Responses." *Journal of Monetary Economics*, 99 41–55.
- Barnichon, Régis, and Geert Mesters.** 2025. "Innovations meet Narratives -improving the power-credibility trade-off in macro." January, Working paper.
- Bauer, Michael D., and Eric T. Swanson.** 2023. "An Alternative Explanation for the "Fed Information Effect"." *American Economic Review*, 113(3): 664–700.
- Brennan, Connor M., Margaret M. Jacobson, Christian Matthes, and Todd B. Walker.** 2024. "Monetary Policy Shocks: Data or Methods?" finance and economics discussion series, Federal Reserve Board, Washington, D.C.
- Bu, Chunya, John Rogers, and Wenbin Wu.** 2021. "A unified measure of Fed monetary policy shocks." *Journal of Monetary Economics*, 118 331–349.
- Caravello, Tomás E., and Pedro Martínez Bruera.** 2024. "Disentangling Sign and Size Non-linearities." June.
- Casini, Alessandro, and Adam McCloskey.** 2024. "Identification and Estimation of Causal Effects in High-Frequency Event Studies." June, Working paper.
- Chen, Xiaohong, Han Hong, and Denis Nekipelov.** 2011. "Nonlinear Models of Measurement Errors." *Journal of Economic Literature*, 49(4): 901–937.

- Cochrane, John.** 2011. “Determinacy and Identification with Taylor Rules.” *Journal of Political Economy*, 119(3): 565–615.
- Cochrane, John.** 2024. “Expectations and the neutrality of interest rates.” *Review of Economic Dynamics*, 53 194–223.
- Fernandez-Villaverde, Jesus, and Juan F. Rubio-Ramirez.** 2007. “Estimating Macroeconomic Models: A Likelihood Approach.” *Review of Economic Studies*, 74(4): 1059–1087.
- Ferreira, Leonardo N., Silvia Miranda-Agrippino, and Giovanni Ricco.** 2024. “Bayesian Local Projections.” Forthcoming.
- Fisher, Irving.** 1935. *100% Money*.: Adelphi Company, , 1st edition 94.
- Friedman, Milton.** 1960. *A Program for Monetary Stability*.: Fordham University Press.
- Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár.** 2024. “Contamination Bias in Linear Regressions.” *American Economic Review*, 114(12): 4015–4051.
- Herbst, Edward P, and Benjamin K. Johansson.** 2024. “Bias in local projections.” *Journal of Econometrics*, 240(1): .
- Imbens, Guido W, and Joshua D. Angrist.** 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica*, 62(2): 467–475.
- Jacobson, Margaret M., Christian Matthes, and Todd B. Walker.** 2024. “Temporal Aggregation Bias and Monetary Policy Transmission.” March, Working Paper.
- Jarociński, Marek.** 2024. “Estimating the Fed’s unconventional policy shocks.” *Journal of Monetary Economics*, 144, p. 103548.
- Jarociński, Marek, and Peter Karadi.** 2020. “Deconstructing Monetary Policy Surprises - The Role of Information Shocks.” *American Economic Journal: Macroeconomics*, 12(2): 1–43.
- Jordà, Òscar.** 2005. “Estimation and Inference of Impulse Responses by Local Projections.” *American Economic Review*, 95(1): 161–182.
- Keynes, John Maynard.** 1936. *The General Theory of Employment, Interest and Money*. Cambridge: Macmillan Cambridge University Press.
- Kim, Jinill, and Francisco J. Ruge-Murcia.** 2009. “How much inflation is necessary to grease the wheels?” *Journal of Monetary Economics*, 56 365–377.

- Kolesár, Michal, and Mikkel Plagborg-Møller.** 2024. “Dynamic Causal Effects in a Nonlinear World: the Good, the Bad, and the Ugly.” Working Paper.
- Koo, Bonsoo, Seojeong Lee, and Myung Hwan Seo.** 2024. “What Impulse Response Do Instrumental Variables Identify?” April, Working paper.
- Kuttner, Kenneth N.** 2001. “Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market.” *Journal of Monetary Economics*, 47(3): 523–544.
- Li, Dake, Mikkel Plagborg-Møller, and Christian K. Wolf.** 2024. “Local Projections vs. VARs: Lessons From Thousands of DGPs.”
- Loria, Francesca, Christian Matthes, and Donghai Zhang.** 2025. “Assessing Macroeconomic Tail Risk.” *The Economic Journal*, 135(665): 264–284.
- Mallat, St’ephane.** 1999. *A Wavelet Tour of Signal Processing.*: Academic Press, i–xxiv, 1–637.
- Masten, Matt.** 2024. “Causality for the Cautious.” Unpublished textbook.
- Miranda-Agrippino, Silvia, and Giovanni Ricco.** 2021. “The Transmission of Monetary Policy Shocks.” *American Economic Journal: Macroeconomics*, 13(3): .
- 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.
- Montiel Olea, José Luis, Mikkel Plagborg-Møller, Eric Qian, and Christian K. Wolf.** 2024. “Double Robustness of Local Projections and Some Unpleasant VARithmetic.”
- Nakamura, Emi, and Jón Steinsson.** 2018. “High-Frequency Identification of Monetary Non-Neutrality: The Information Effect.” *The Quarterly Journal of Economics*, 133(3): 1283–1330.
- Plagborg-Møller, Mikkel, and Christian K. Wolf.** 2021. “Local Projections and VARs Estimate the Same Impulse Responses.” *Econometrica*, 89(2): 955–980.
- Rambachan, Ashesh, and Neil Shephard.** 2021. “When do common time series estimands have nonparametric causal meaning?” 10, Working paper.
- Ramey, Valerie.** 2016. “Macroeconomic Shocks and Their Propagation.” In *Handbook of Macroeconomics*. Chap. 2.
- Romer, Christina D., and David H. Romer.** 2004. “A New Measure of Monetary Shocks: Derivation and Implications.” *American Economic Review*, 94(4): 1055–1084.
- Sims, Christopher A.** 1998. “Comment on Glenn Rudebusch’s “Do Measures of Monetary Policy in a VAR Make Sense?”.” *International Economic Review*, 39(4): 933–941.



- Small, Dylan S., Zhiqiang Tan, Roland R. Ramsahai, Scott A. Lorch, and M. Alan Brookhart.** 2017. "Instrumental Variable Estimation with a Stochastic Monotonicity Assumption." *Statistical Science*, 32(4): .
- Stock, James H., and Mark W. Watson.** 2018. "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments." *Volume 128, Issue 610*, 128(610): 917–948.
- Tobin, James.** 1972. "Inflation and unemployment." *American Economic Review*, 62 1–18.
- White, Halbert.** 1980. "Using Least Squares to Approximate Unknown Regression Functions." *International Economic Review*, 21(1): 149–170.
- Xu, Ke-Li.** 2023. "Local Projection Based Inference under General Conditions."
- Yitzhaki, Shlomo.** 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business & Economic Statistics*, 14(4): 478–486.

## Appendix<sup>21</sup>

### A.1 Proof and Discussion of Proposition 1

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  $0 \in O_0$ ,  $O_0 \cup (\cup_{i=1}^N I_i) = \mathbb{R}$ , and  $O_0 \cap (\cup_{i=1}^N I_i)$  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  $\mathbf{X}_t^f$  has full rank,  $x \in I_i \iff f_i(x) \neq 0$ , 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:  $f_i^\perp(x_t)$  are 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  $\{f_i(x_t)\}_{i=1}^N$  on a sign partition  $\{I_i\}_{i=1}^N$ , 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  $\{f_i(x_t)\}_{i=1}^N$  (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  $\{f_i(x_t)\}_{i=1}^N$  on a sign partition  $\{I_i\}_{i=1}^N$ . 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  $b_0 = \mathbb{E}[f_1] - \sum_{k=2}^N a_{k-1} \mathbb{E}[f_k]$ . The other estimands solve  $b_0 \mathbb{E}[f_k] = -b_{k-1} \mathbb{E}[f_k^2]$ , which leads to  $a_{k-1} = -b_0 \text{sign}(I_k)$ , where  $k \geq 1$  and  $\text{sign}(I_k)$  is equal to  $\text{sign}(i_k)$  for any  $i_k \in I_k$  (given our definition of normalized collection and sign partition). Define  $\mu_1 = \mathbb{P}(x \in I_1)$  and  $\mu_0 = \mathbb{P}(x \in O_0)$ . Substituting into the equation for the constant, we get that  $\text{sign}(I_1) \mu_1 = b_0 (\mu_1 + \mu_0)$ . Therefore

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

<sup>21</sup>The Online Appendix can be found in the paper's GitHub repository <https://github.com/paulbousquet/UncoveringNonlin>

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  $\varepsilon_t$  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

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

where  $F(\cdot)$  is the distribution function of  $\varepsilon_t$ . 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 shows that the weights represent the remaining mass  $\varepsilon_t$  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-weighted parts are on  $O_0$ , so moving along the real line towards  $\infty$  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.<sup>22</sup> Combined with the extensions of [Kolesár and Plagborg-Møller \(2024\)](#) shown in (3), we have shown these coefficients represent positively weighted sum 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

<sup>22</sup>Note if there are gaps in the support  $I \subset \mathbb{R}$ , the behavior is the same just in a discontinuous fashion.

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  $\beta_i$  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  $\beta_1$  and  $\beta_2$ . If  $\int_I \omega_1(a)da < 1$  and  $\int_I \omega_2(a)da = 1$ , then even if  $m_h(\cdot)$  is linear in  $\varepsilon_t$ ,  $\beta_1 \neq \beta_2$ . 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.

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  $\alpha_i$  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  $\int_I \omega_i^g(a)da = 1$ . 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). This is easily estimatable by OLS. For the standard error correction, the projection estimands for  $g_i$  defined implicitly in terms of all the  $\alpha_i$ . 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  $\left(\frac{\tilde{\beta}_i}{\partial \alpha_i}\right)^2 \text{Var}(\alpha_i) = \frac{\tilde{\beta}_i^2 \text{Var}(\alpha_i)}{\alpha_i^2}$ , 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  $\varepsilon_t$  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  $I_1, \dots, I_{n^-}$  are the group of negative intervals and  $I_{n^-+1}, \dots, I_N$  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  $\beta_i$  will non-zero for  $a \in (I_{i-1} \cup I_i)$ . For if  $i = 1, n^- + 1$ , there is non-zero weight for  $a \in (O_0 \cup I_i)$ . 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  $\beta_i$  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  $\beta_i$  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.

## A.2 Inherency of Negative Weight

If  $\varepsilon_t$  is a continuously distributed shock on  $I \subset \mathbb{R}$ , note that<sup>23</sup>

$$\begin{aligned} \int_I \text{Cov}(\mathbf{1}_{\{a \geq \varepsilon_t\}}, f(\varepsilon_t)) da &= \int_I \left\{ \mathbb{E}[\mathbf{1}_{\{a \geq \varepsilon_t\}} f(\varepsilon_t)] - \mathbb{E}[\mathbf{1}_{\{a \geq \varepsilon_t\}}] \mathbb{E}[f(\varepsilon_t)] \right\} da \\ &= \mathbb{E} \left[ (f(\varepsilon_t) - \mathbb{E}[f(\varepsilon_t)]) \left( \int_{I_t} da \right) \right] = \mathbb{E}[(f(\varepsilon_t) - \mathbb{E}[f(\varepsilon_t)]) \varepsilon_t] = \text{Cov}(f(\varepsilon_t), \varepsilon_t) \end{aligned}$$

where  $I_t = \{x \in I : x \leq \varepsilon_t\}$ . Also notice for a generic  $f(\cdot)$  and  $g(\cdot)$

$$\int_I \text{Cov}(\mathbf{1}_{\{a \geq \varepsilon_t\}}, f(\varepsilon_t)) - \frac{\text{Cov}(f(\varepsilon), g(\varepsilon))}{\text{Var}(g(\varepsilon))} \int_I \text{Cov}(\mathbf{1}_{\{a \geq \varepsilon_t\}}, g(\varepsilon_t)) = \text{Cov}(f(\varepsilon_t), \varepsilon_t) - \frac{\text{Cov}(f(\varepsilon), g(\varepsilon))}{\text{Var}(g(\varepsilon))} \text{Cov}(g(\varepsilon_t), \varepsilon_t)$$

Recall from (4), the weight function on  $\beta_2$  from (5) will follow  $\omega_2(a) = \frac{\text{Cov}(\mathbf{1}_{\{a \leq \varepsilon_t\}}, f(\varepsilon)^\perp)}{\text{Var}(f(\varepsilon)^\perp)}$ . Since  $f^\perp(\varepsilon)$  in this case is  $f(\varepsilon) - \frac{\text{Cov}(f(\varepsilon), \varepsilon)}{\text{Var}(\varepsilon)}$ ,  $\int_I \omega_2(a) da$  will be proportional to the above result when  $g(\varepsilon) = \varepsilon$ , which is 0. Thus, if  $\omega_2(a) \neq 0$  for any  $a$ , there must be both  $\omega_2$  must take on negative values, stripping us of grounds to make causal claims.

## A.3 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 B.3, as well as derivations for the case that was initially presented in Section 3.3 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)  $\{\varepsilon_t\}_{t=0}^T$ . 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  $I_i$ .<sup>24</sup> 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  $\widehat{\beta}_i^2 / n_i$ .

<sup>23</sup>This holds in the interior of  $I$ , see Kolesár and Plagborg-Møller (2024) Lemma 3 and Caravello and Martínez Bruera (2024) Lemma 1.

<sup>24</sup>The same adjustment described in Proposition 1 can be performed to let the weights integrate to 1

## A.4 Unpacking the Weight Form

Recall the general form from (4) in Section 2

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

where  $X_i^\perp$  is the residual from regressing  $X_i$  on the other elements in  $\mathbf{X}_t$ . We can unpack this definition to get things solely in terms of covariances and variance of terms of  $X_t$ , 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  $X$  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  $X$ . 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  $X$  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. The details are tedious, so they are relegated to Online Appendix B.1. Those details allow us to explicitly write out the  $N = 4$  special case of interest. Recall that the setting of interest is including functions  $\{f_i(\varepsilon_t)\}_{i=1}^4$  in a regression, where  $\varepsilon_t$  is a shock. The weights  $\omega_i(a)$  in  $\beta_i$  (corresponding to the  $i$ -th function) are

$$\omega_i(a) = \frac{C_{\mathbb{1},i} - \sum_{j \neq i} C_{\mathbb{1},j} \frac{C_{i,j} - \sum_{k \neq j} \frac{C_{i,k} C_{j,k}}{V_k}}{V_j - \sum_{k \neq j} \frac{C_{j,k}^2}{V_k}}}{V_i - \sum_{j \neq i} \frac{C_{\mathbb{1},j}^2 - 2C_{\mathbb{1},j} \sum_{k \neq j} \frac{C_{i,k} C_{j,k}}{V_k} + \sum_{k \neq j} \frac{C_{i,k}^2 C_{j,k}^2}{V_k^2}}{V_j - \sum_{k \neq j} \frac{C_{j,k}^2}{V_k}}}$$

where  $C_{i,j}$  denotes the covariance between  $f_i$  and  $f_j$ ,  $C_{\mathbb{1},i}$  is the covariance between  $\mathbb{1}_{(\varepsilon_t \geq a)}$  and  $f_i$ , and  $V_i$  is the variance of  $f_i$ . This  $N = 4$  case is actually simple compared to the sprawling recursions of the general case. The representation above also implicitly assumes the functions are mean 0, which need not be the case.

As made explicit at the beginning of Section 3, the goal is to pick functions so that  $\omega_i(a)$  are non-negative, relevant (don't put weight where we don't want), and hump-shaped. The inscrutable form above makes deep learning a natural solution to the complex function search in the case where we allow the functions to potentially be correlated.

## A.5 Illustrations of Functions and Their Weights

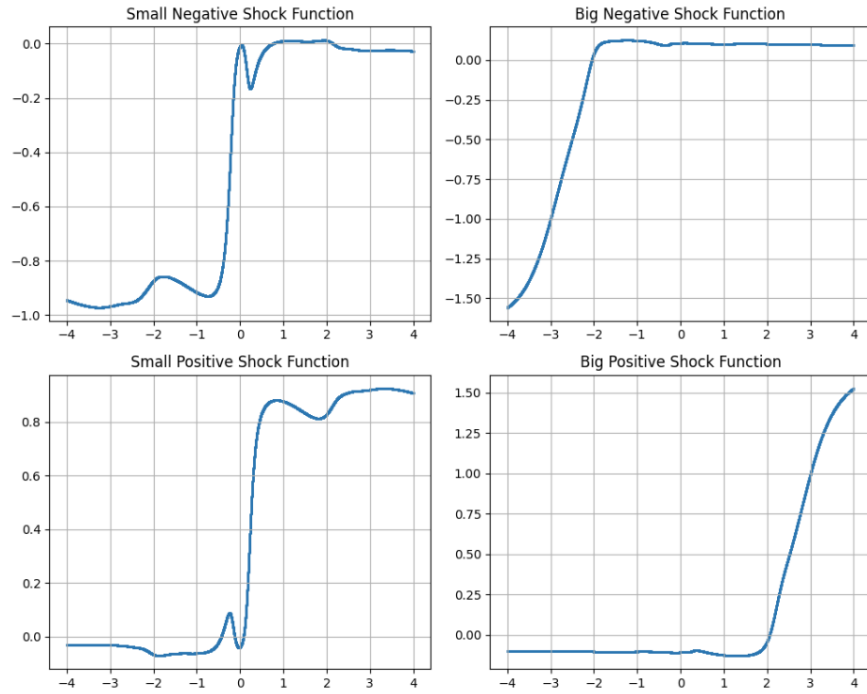


Figure 12: Neural Network Output with Standard Normal Shocks

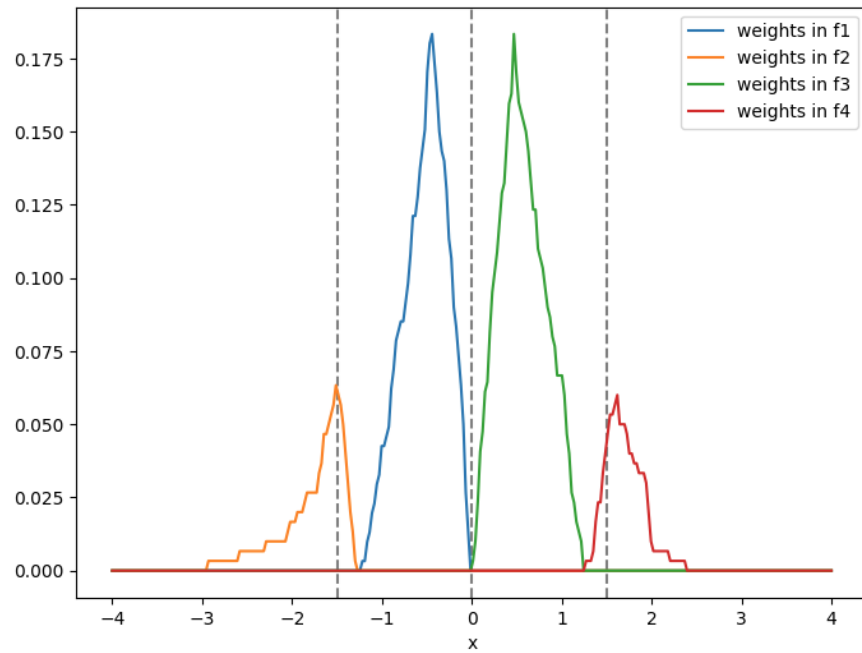


Figure 13: Generated Regressor Weights, Standard Normal Shocks



## A.6 More Application Results

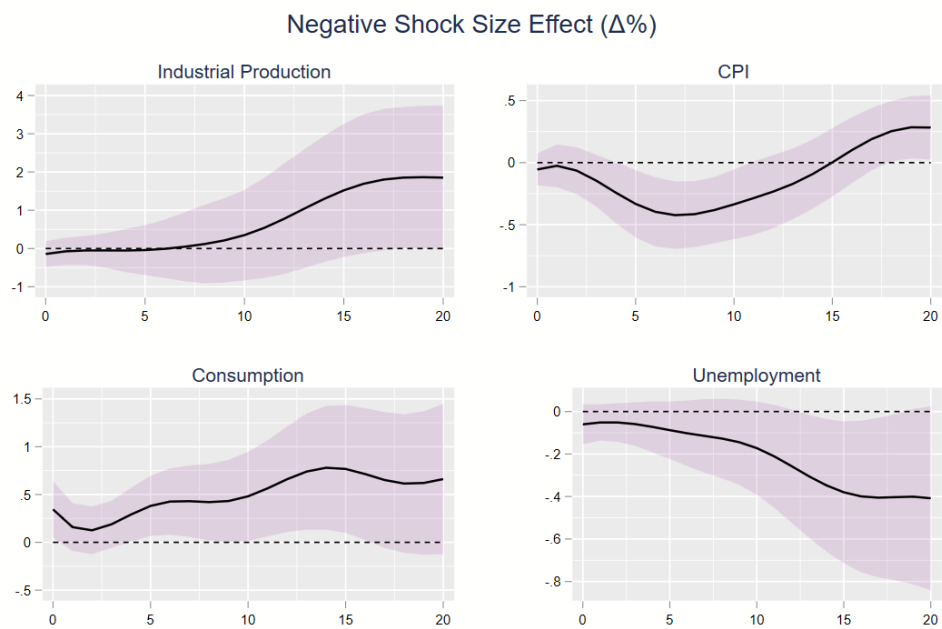


Figure 14: Indicator Function Approach with MP1 Shocks

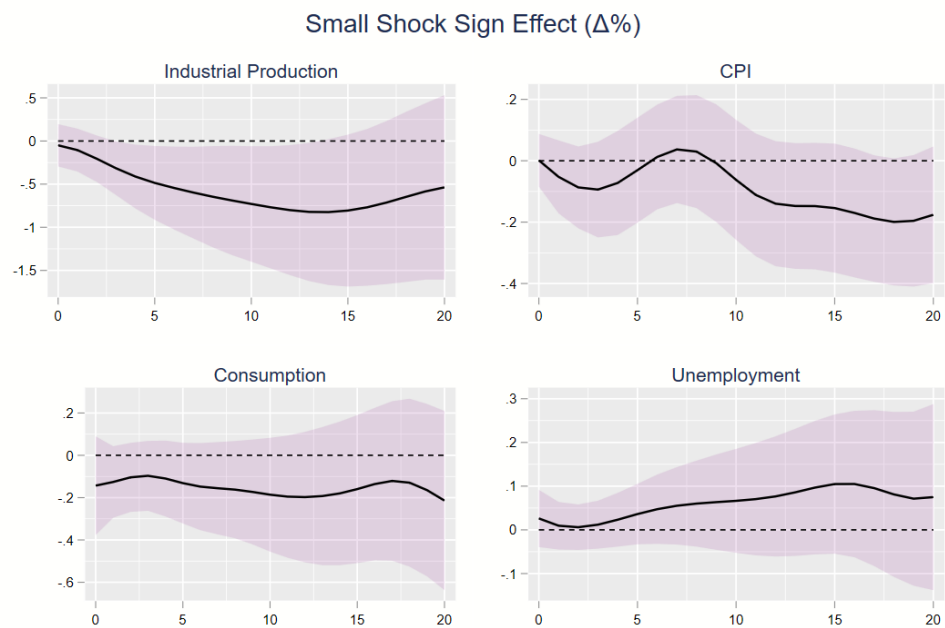


Figure 15: Indicator Function Approach with MP1 Shocks