Uncovering Nonlinearities with Regression Anatomy

Paul Bousquet*

February 12, 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. I thank Rachel Childers for providing 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, 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 α 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 LP 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 ε sufficiently large. Appropriate weighting justifies labeling $\beta_{i,j}$ with combinations of $i = \{\text{big}, \text{small}\}$ and $j = \{\text{positive}, \text{negative}\}$. 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, which also relates to efforts estimating IRFs from data. For the monetary policy application, there is an abundance of "puzzles" (Ramey, 2016), creating a noisy literature – one can find a well-cited paper suggesting variable x responds in y direction for all x and y. Looking at coefficient differences can help sidestep the puzzle rabbit hole by focusing on *relative* effects.

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).[1] 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, including selection criteria for the vast array of structural shock proxies. 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.

[1]"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) $g_h : \mathbb{R} \times \mathbb{R}^L \to \mathbb{R}$ for an outcome variable Y at time $t+h$

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

Here, ε_t is the structural shock of interest at time t and $\boldsymbol{S_{t+h}}$ is "everything else" in the system, which could for instance include the information set at time t as well as leads and lags of ε_t (and other shocks). Following Rambachan and Shephard (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 $\varepsilon_t \perp \boldsymbol{S_{t+h}}$ $\forall h \geq 0$. In that case, note that the conditional mean $\mathbb{E}[g_h(a, \boldsymbol{S_{t+h}})|\varepsilon_t = a]$ 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 $\left\{f_i(\cdot)\right\}_{i=1}^N$ and control set $\boldsymbol{W_t}$, suppose we regress Y_{t+h} on $\left\{1, \left\{f_i(\varepsilon_t)\right\}_{i=1}^N, \boldsymbol{W_t}\right\}$. The specification is

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

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

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

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

where $X_i^{\perp}$ is the residual from regressing the ith element of $\boldsymbol{X_t}$ on the remaining $N-1$ elements.[2] Thus, the estimands can be described as a weighted average of the data generating process' true marginal effects. In the Appendix, 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 ε_t. This of course will not hold if ε_t is not actually a shock (it will also depend on the control set $\boldsymbol{W_t}$). In addition, if we instead use a proxy z_t in place of ε_t (if ε_t is not observable), the weights still depend on ε_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.

[2] And a constant, in case ε_t is not mean 0 or some functions have a non-zero y-intercept. Also need $\left\{f_i(\varepsilon_t)\right\}_{i=1}^N$ s.t rank condition holds.

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)

$$Y_{t+h} = \alpha + \beta_1 \varepsilon_t + \beta_2 f(\varepsilon) + \gamma' \boldsymbol{W_t} + \epsilon_{t+h} \tag{5}$$

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

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

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

These results provide important clarity on past work (e.g., ε^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 ε_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 ε, $\beta_1 + 2\beta_2\varepsilon$ is not a consistent estimate for the average marginal effects of ε. This is because a corollary to their Proposition 1 is in specification (5), there must be negative weight placed on β_2 (see the Appendix for a proof). In general, specifications that include functions of ε 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 ε_t and also y_t on $\mathbb{1}_{\varepsilon_t \geq 0}$, both estimands are an average of the *same* object (marginal effects of ε_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 ε 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 this section 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 $\frac{\text{Cov}(\mathbf{1}_{\{a \leq \varepsilon_t\}}, X_i^\perp)}{\text{Var}(X_i^\perp)}$ is not obvious, so an example is useful. Under the following DGP[3]

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

note that $\mathbb{E}_t[\pi_{t+1}]$ will be a constant. So a regression of π_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: $$\pi_t = \alpha + \beta_1 y_t \cdot \mathbf{1}_{y_t > 0} + \beta_2 y_t \cdot \mathbf{1}_{y_t \leq 0} + \epsilon_t$$

[3]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.

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 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, $\widehat{\beta}_2$ 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.

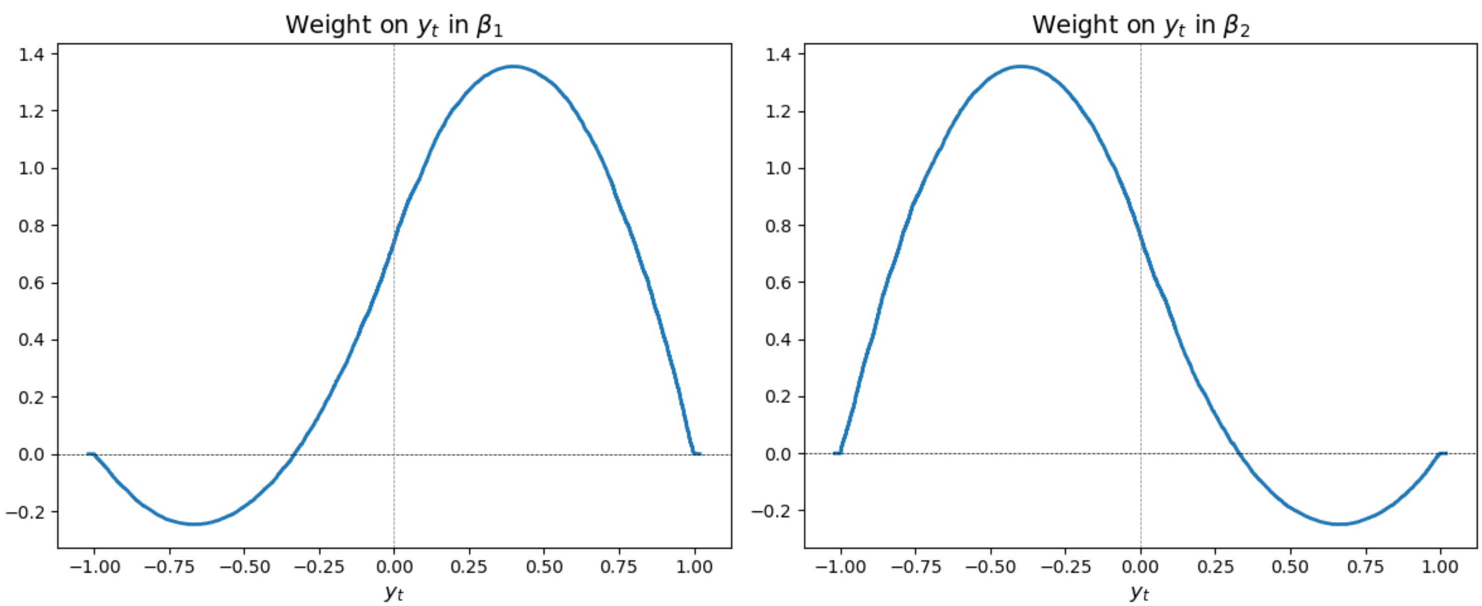

Figure 1: Uniform Shock with Example 1

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.[4] Even in the most aspirational scenario when shocks follow a well-behaved, symmetric distribution, the weights may be far from converging.[5]

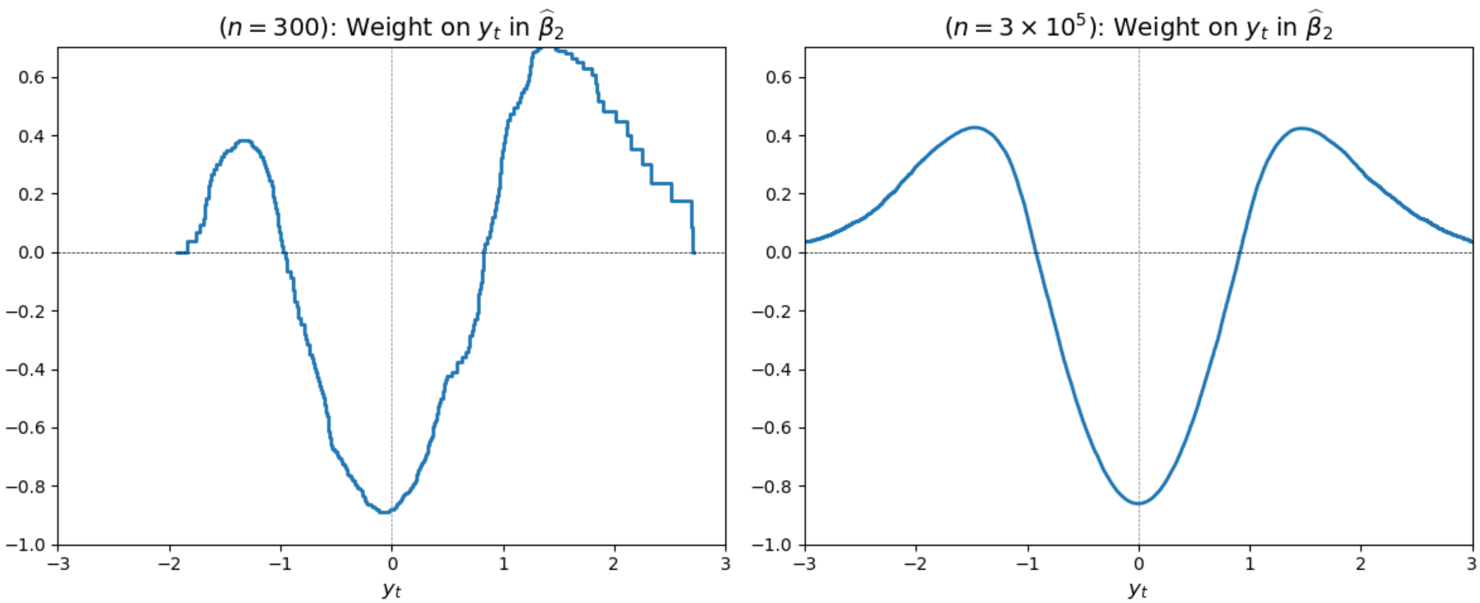

Figure 2: Standard Normal Shock with Selected Simulation of Example 2

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". The next section will show better approaches to accomplish the goal of nonlinearity detection. One cause of the lacking performance in Example 2 is the inclusion of the shock itself in the regression. As stated earlier, this guarantees the presence of negative weights.

[4]Here, $f(y)=\mathbb{1}_{y\geq\overline{y}}\cdot(y-\overline{y})+\mathbb{1}_{y\leq-\overline{y}}\cdot(y+\overline{y})$, where $\overline{y}$ is σ away from the mean (0). Results are similar for $f(y)=y^3$.

[5]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.

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. Sections 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 ε_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 ε_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 ≤ 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.[6]

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 limit 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 the results in Example 1 might give us pause. It turns out that interacting the indicator functions with the shock is the culprit, and just using the indicator functions themselves works well. To see this, we will re-do Example 2 with indicator functions. One really nice property of using regressions to detect nonlinearities is the implicit weighting is invariant to the

[6] 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.

outcome variable. So really, we don't need a model to evaluate the weights, just a time series for the shock process. So when we consider

Example 3: $\pi_t = \alpha - \beta_{\text{small, neg}} \cdot \mathbb{1}_{-y_t \in [.01,1.5]} - \beta_{\text{big, neg}} \cdot \mathbb{1}_{y_t < -1.5} + \beta_{\text{small, pos}} \cdot \mathbb{1}_{y_t \in [.01,1.5]} + \beta_{\text{big, pos}} \cdot \mathbb{1}_{y_t > 1.5} + \epsilon_t$

note that the weight plots in Figure 3 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 ($i = \{\text{big}, \text{small}\}$) and sign ($j = \{\text{positive}, \text{negative}\}$). 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 3 shows weight plots for this specification.[7] 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.

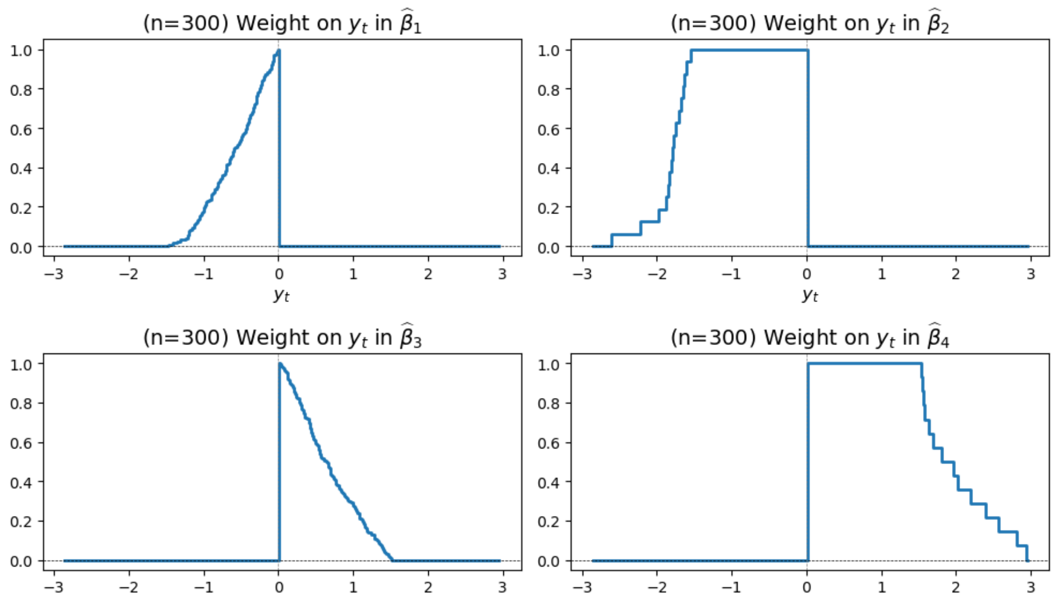

Figure 3: Standard Normal Shock with Example 3

So now to put Figure 3 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 achieve the same property under any shock

[7]Like Figure 2, this will vary across simulations, but the variance here is concentrated exclusively at the endpoints for the big shock weights.

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.

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

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

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

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

Proposition 1. *Suppose ε_t is a continuously distributed shock on $I \subset \mathbb{R}$ and Y_{t+h} follows a data generating process of the form (1) satisfying the conditions of Kolesár and Plagborg-Møller (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 $\boldsymbol{X}_t$ be their concatenation. If we project Y_{t+h} on $\boldsymbol{X}_t$ (and a constant and control set as in (2)), then $\beta_i = \beta_j\ \forall i, j$ if $m_h(\cdot)$ is linear in ε_t.*
Let $S_{ij} = O_0 \cup I_i \cup I_j$ for $i \neq j$. $\beta_i = \beta_j$ for $i \neq j$ if $m_h(\cdot)$ is linear in ε_t on $\left(\inf\{S_{ij}\}, \sup\{S_{ij}\}\right) \cap I$.

In plain terms: if the DGP is linear on the space where the weights on β_i and β_j are non-zero, then $\beta_i = \beta_j$. The statement of the result is a bit technical because of a couple subtle points. Notice that the total weight for big and small shocks of the same sign in Figure 3 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 the Appendix, 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

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

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 3, 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 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 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.

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. So thinking about the linear case where marginal effects are constant, the integral over a portion of the support will be the same no matter the portion. So when we compare two estimates, they will meaningfully differ only if there is nonlinearity. As with Proposition 1, the converse is not true; β_i, β_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 $n \to \infty$, a hypothesis test will always reject a null hypothesis that two estimates are the same even if the difference is marginal. Section 4 gives some ways to gauge if the rejection is consequential.

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

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

Suppose ε_t follows distribution F with support I and the collection $\{f_i\}_{i=1}^N$ corresponds to a partition 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. Even though a strict no overlap requirement is not one of the weight targets, it turns out 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

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

The next step is to find N functions, staying within this class, producing weights that are non-negative, relevant, and hump-shaped. The expression above shows a clear route to satisfaction. WLOG, consider the interval $I_i = [0,1]$. For a fixed $c \in (0,1)$, ε_t has probability mass $F(c) - F(0)$ on $[0, c]$ and mass $F(1) - F(c)$ on $[c, 1]$. Consider

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

This function abides by our constraint and targets:

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

There are some clear downsides, however. Recall from earlier that R_i denotes the region where it's permissible for weight to be placed. This structure allows for the possibility of weight overlap, which is ideal because we don't want to get married to qualitative descriptions for the partitioning of a shock's support (e.g., "$a = .99$ is a small shock, $a = 1.01$ is a big shock"). But 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 we will not know the distribution function. The procedure will work if instead use the empirical CDF, but we would much rather the functions we use not vary across repeated sampling. With these generated regressors, there would need to be a standard error correction, outlined later in this section and in the Appendix, which adds to 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

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

The Appendix shows the integral of the weights in β_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 β_1 represents a positively weighted average of marginal effects.[8] Equally, we need the analogous expression for β_2 to be positive. The simplest path to joint satisfaction is the functions are correlated with ε yet uncorrelated with each other. As the number of function grows, the potentially paradoxical paths become more unwieldy. For the second goal, we know from (4) the weights in β_1 will be large where ε_t has more density and $f_1(\varepsilon_t)$ is large (provided $\mathbb{1}_{a \le \varepsilon_t} = 1$).

All these "steps to success" contextualize the moderate success of disjoint indicator functions for the $N = 4$ case seen in Example 3. 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

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

[8] Though recall this is not a sufficient condition on its own; see Example 1 in the previous section.

For this case, one instance of training with standard normal shocks (which will be used in the next subsection to test assess performance against several DGPs) produces "small" functions resembling indicators and "big" functions that look like a ReLu. Their plots (Figure 9) roughly look like (chronologically)

$$f_1(x) = \mathbb{1}_{x>-0.5} - 1 \ \underline{\textbf{and}} \ f_2(x) = \min\{-.8x + 2, 0\}$$
$$f_3(x) = \mathbb{1}_{x>-0.1} - .1 \ \underline{\textbf{and}} \ f_4(x) = \max\{0, .8x - 2\}$$

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

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

In principle, a deep learning algorithm for the objectives (weighting targets) described at the beginning of this section is simple. Each iteration of training (epoch) will generate a candidate collection of functions $\left\{f_i(\cdot)\right\}_{i=1}^{4}$. 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 the next section and the Appendix. One complication 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 3. 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 irony is the methods developed to produce less false negatives failed to do so. This is because, while the weights look more appealing (see Figure 8 in the Appendix), 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 9 in the Appendix) 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 not autocorrelation). This is an indication that disjoint indicator functions should be the default method of choice, though in principle all methods jointly can be used as well.

One last point to discuss before moving on from DGP (6), that was not mentioned earlier, 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 raises an important limitation – an important nonlinearity of economic interest would be diminishing returns to scale of policy intervention. However, these are intrinsically harder to detect – the second derivative of the square-root function is essentially constant after moving away from 0. This is an unfortunate downside this framework is not as well-equipped to handle.

To get a more wholistic picture of performance, we can 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, which are described explicitly in the appendix. 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 the Appendix, 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 π_t and additional fundamentals $\boldsymbol{W}_t$. 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 π_t and randomly select other variables from the remaining categories for $\boldsymbol{W}_t$. 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

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

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 π_t and $\boldsymbol{W}_t$. For each choice of $g(\cdot)$, I run several local projections at horizon $h = 1$ for the different approaches for detecting nonlinearities. The LPs have 196 observations and include 4 lags of all variables except $\boldsymbol{W}_t$, which is not included at all to mimic the presence of omitted variable in practice. The results represent an average across 100 variations of (7) with 10,000 simulations each.

Figure 4 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 θ. 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 θ, 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 $\theta x^2 \cdot \mathbb{1}_{x \geq c} + x \cdot \mathbb{1}_{x<c}$. The shock is standardized (so c is set to 1.25) and across simulation follows 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. Again, this is reflective of the exacerbation of efficiency issues in local projection, and in general estimates will often not be significant. 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.

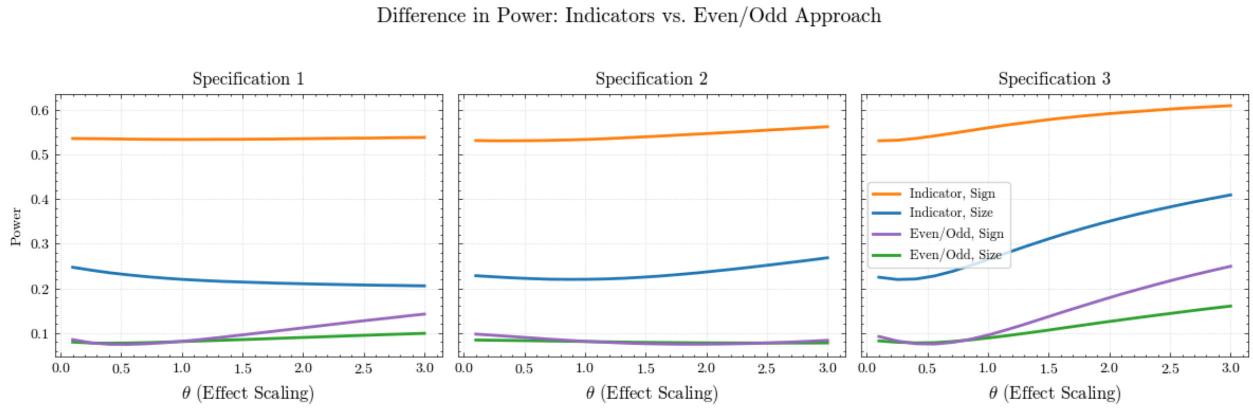

Figure 4

In every previous simulation, we have assumed the structural shock ε_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, heteroskedaticity, 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 5 plots the case of $z_t = \text{sgn}(x)(x+\epsilon)^2$ where ϵ 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.[9]

[9]This is another reason to use disjoint indicators. While it forces all coefficients to put some weight on the wrong sign, the values of ε_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"

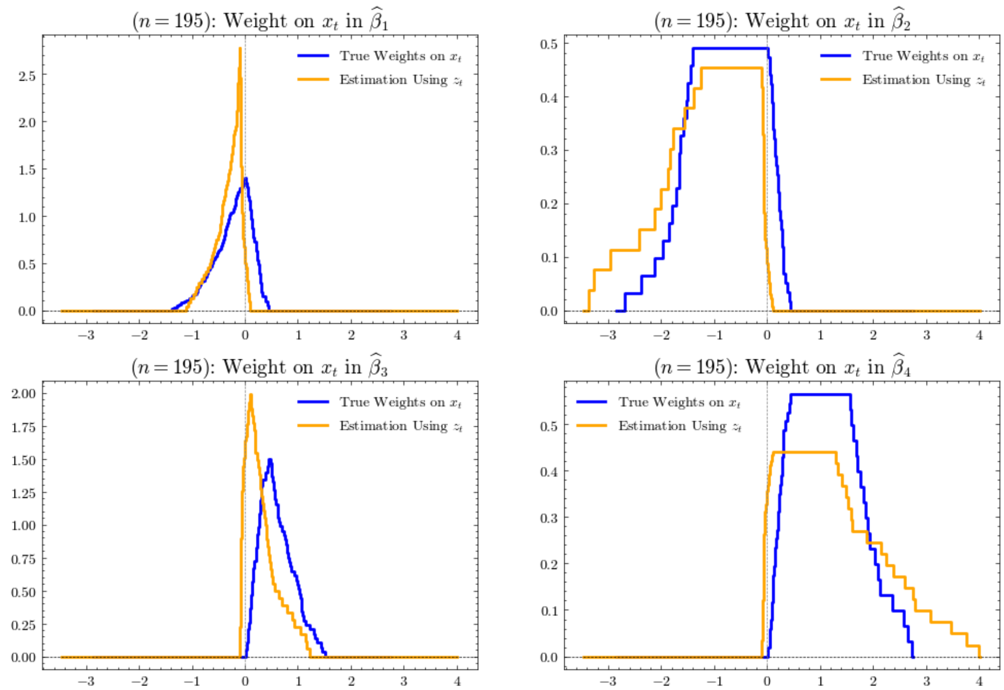

Figure 5: Selected Simulation of DGP (7) with Measurement Error

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 3 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 the actual structural shock of interest ε_t being used in the regression. In practice, we will have some proxy for the shock z_t. Kolesár and Plagborg-Møller (2024) note that if ε_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 the divergence between ε_t and z_t results from noise, even if this "measurement error" follows a complex process, 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 the special case of the previous section's DGPs where shocks enter linearly, point estimates in the regression of y_t on ε_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 $\varepsilon_t \approx z_t$.[10]

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. Findings include Newey-West standard errors can be problematic and using Huber-White with lagged control variables is sufficient (Herbst and Johannsen, 2024; Montiel Olea and Plagborg-Møller, 2021), point estimates should be adjusted for autocorrelation (Herbst and Johannsen, 2024), and "state-dependent" LPs are appropriate for average marginal effects, not state dependent impulse responses (Gonçalves et al., 2024). 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).[11] (Plagborg-Møller and Wolf, 2021) prove LP and VAR are asymptotically equivolent 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 since 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 chronology to implement this procedure is the following. First, select a shock series based on the confidence in approximating the underlying weights well and standardize it. Accumulate relevant control variables and run local projections with disjoint indicator functions representing your chosen partitioning of the shock's support. Because of sample size limitations, this paper has emphasized a choice of $N = 4$. In the regressions, include a healthy number of lags and use Huber-White Standard Errors for inference. The generated regressor and machine learning approach can complement the indicator function results, but may be limited in their ability to limit the amount of false negatives because the standard errors will be relatively large under typical sample sizes. 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 approach for the primary illustration of nonlinearities is to use penalized local projections. The one

[10]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.

[11]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.

drawback is inference is complicated by complications from regularization (and also possibly cross-validation). To be robust to the potential bias in estimating the variance-covariance matrix, Barnichon and Brownlees (2019) recommend computing using standard errors from an even more "under-smoothed" estimator. To go a step further, in my 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 β 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 roughly 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 my Application in Section 4.3, this translates to measuring the nonlinearity in terms of difference in percent change in outcome since the shock occurred.[12]

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 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

[12] 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 the Online Appendix.

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.[13] 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 forecasts in meeting notes.[14] But the actual values of these residuals 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 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, but suppose they had originally done

[13] There is risk of contamination in the few instances where there were shocks in the days before the formal announcement of the target.
[14] Aruoba and Drechsel (2024) argue these forecasts don't span the information set. They extend the methodology with text analysis.

this procedure in 2015. The median percent difference in shock magnitude between the original and "updated" series would be over 100%.[15]

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. 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 4.1 to be over-correct for any potential bias. I find evidence of nonlinearities in each variable. Figure 6 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. The Appendix 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 11). For sign effects, not much can be said about asymmetries for small shocks (Figure 12).

The generate regressor and deep learning approaches also point in the same direction. In particular, Figure 7 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 the Online Appendix, 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.

[15] 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).

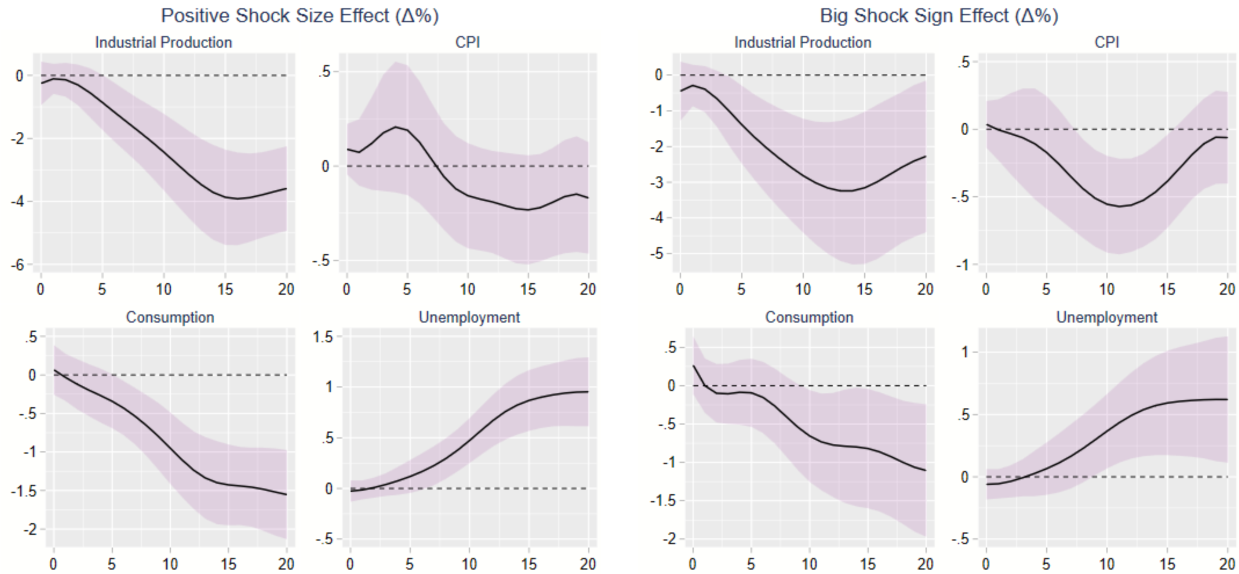

Figure 6: Indicator Approach with MP1 Shocks

Figure 7: 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 $\psi_p \to 0$. 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 the Appendix. Using the same sample period of US data, the Aruoba et al. (2017) extension of the downward-rigity 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, with results relegated to the Appendix, show 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.

Looking at the impulse response functions directly from the model (rather than running a LP) corroborates the above interpretations. Figure 8 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. The Appendix discusses various extensions to the model, like adding autocorrelated shocks, that ultimately don't help much.

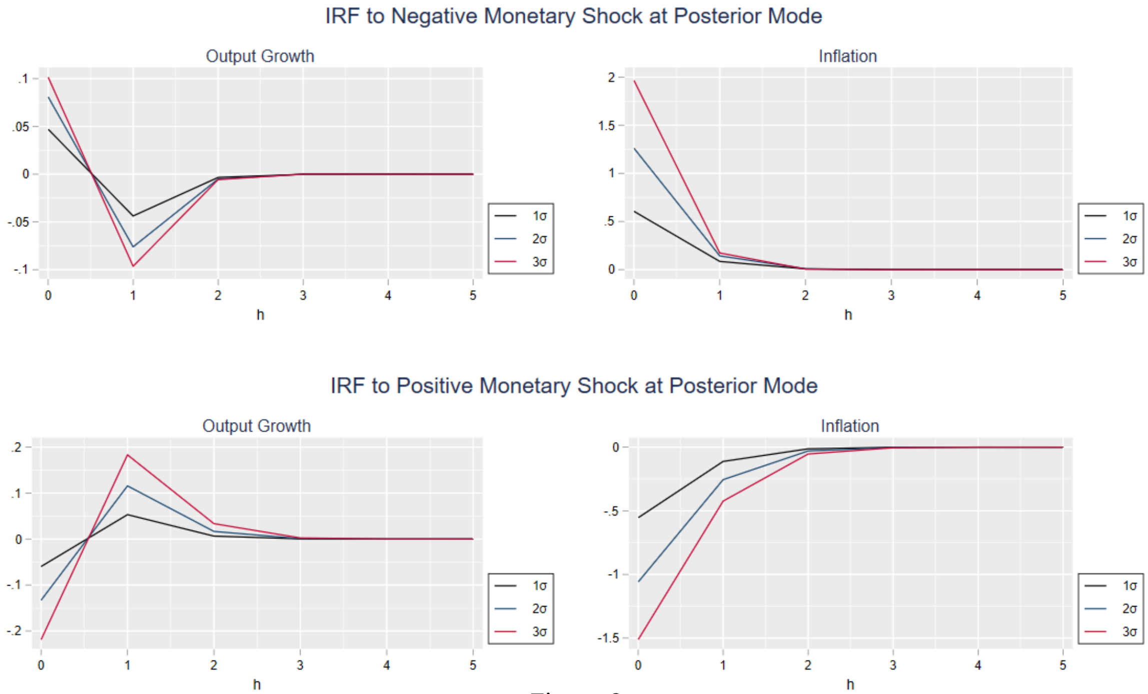

Figure 8

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.[16] Table 1 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. 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.

[16] Regardless of model, consecutive, large realizations of white noise innovations are unlikely, but the staying power of shocks can vary.

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, Regis, and Christian Brownlees. 2019. "Impulse Response Estimation by Smooth Local Projections." *The Review of Economics and Statistics*, 101(3): 522–530.

Barnichon, Regis, 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.

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

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

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.

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.

Yitzhaki, Shlomo. 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business & Economic Statistics*, 14(4): 478–486.

Appendix

Inherency of Negative Weight

If ε_t is a continuously distributed shock on $I \subset \mathbb{R}$, note that[17]

$$\int_I \mathrm{Cov}\big(\mathbf{1}_{\{a\geq\varepsilon_t\}}, f(\varepsilon_t)\big)\mathrm{d}a = \int_I \Big\{\mathbb{E}[\mathbf{1}_{\{a\geq\varepsilon_t\}} f(\varepsilon_t)] - \mathbb{E}[\mathbf{1}_{\{a\geq\varepsilon_t\}}]\mathbb{E}[f(\varepsilon_t)]\Big\}\,\mathrm{d}a$$

$$= \mathbb{E}\left[\big(f(\varepsilon_t) - \mathbb{E}[f(\varepsilon_t)]\big)\left(\int_{\underline{I}_t} \mathrm{d}a\right)\right] = \mathbb{E}\big[(f(\varepsilon_t) - \mathbb{E}[f(\varepsilon_t)])\varepsilon_t\big] = \mathrm{Cov}(f(\varepsilon_t), \varepsilon_t)$$

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

$$\int_I \mathrm{Cov}\big(\mathbf{1}_{\{a\geq\varepsilon_t\}}, f(\varepsilon_t)\big) - \frac{\mathrm{cov}(f(\varepsilon), g(\varepsilon))}{\mathrm{var}(g(\varepsilon))}\int_I \mathrm{Cov}\big(\mathbf{1}_{\{a\geq\varepsilon_t\}}, g(\varepsilon_t)\big) = \mathrm{cov}(f(\varepsilon_t), \varepsilon_t) - \frac{\mathrm{cov}(f(\varepsilon), g(\varepsilon))}{\mathrm{var}(g(\varepsilon))}\mathrm{cov}(g(\varepsilon_t), \varepsilon_t)$$

Recall from (4), the weight function on β_2 from (5) will follow $\omega_2(a) = \frac{\mathrm{Cov}(\mathbf{1}_{\{a\leq\varepsilon_t\}}, f(\varepsilon)^{\perp})}{\mathrm{Var} f(\varepsilon)^{\perp}}$. Since $f(\varepsilon)^{\perp}$ in this case is $f(\varepsilon) - \frac{\mathrm{cov}(f(\varepsilon),\varepsilon))}{\mathrm{var}(\varepsilon)}$, $\int_I \omega_2$ 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 ω_2 must take on negative values, stripping us of grounds to make causal claims.

Standard Errors for Generated Regressors

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

Specifically, for each $a \in I$

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

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

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

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

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

Because this estimand is formed with respect to a generated regressor, we need to adjust the standard errors.

[17] I largely follow Caravello and Martínez-Bruera (2024) for the derivations

Adjustment is done using the delta method. Differentiating

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

The adjustment takes the form of [18]

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

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

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

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

Unpacking the Weight Form

Recall the general form from (4) in Section 2

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

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

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

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

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

[18]Note that f_i is not differentiable at c_i

are relegated to the Online Appendix.[19] 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 ε_t is a shock. The weights $\omega_i(a)$ in β_i (corresponding to the i-th function) are

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

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.

Illustrations of Functions and Their Weights

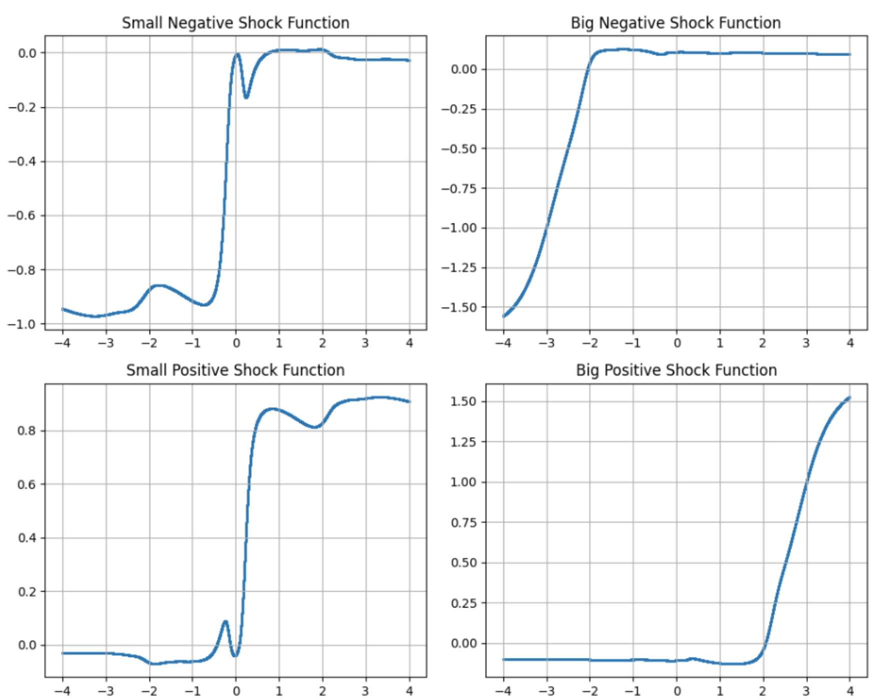

Figure 9: Neural Network Output with Standard Normal Shocks

[19]The Online Appendix can be found in the paper's GitHub repository https://github.com/paulbousquet/UncoveringNonlin

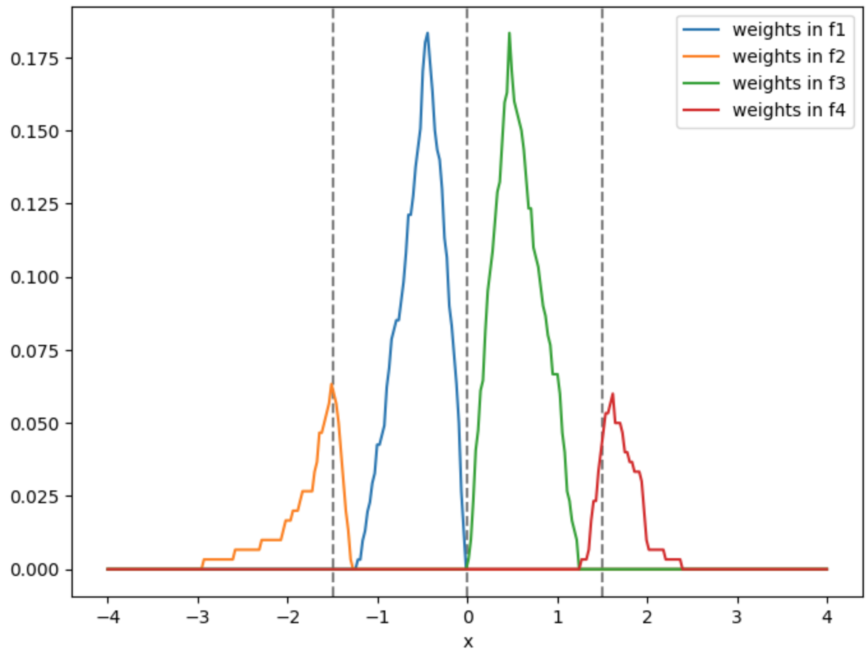

Figure 10: Generated Regressor Weights, Standard Normal Shocks

More Application Results

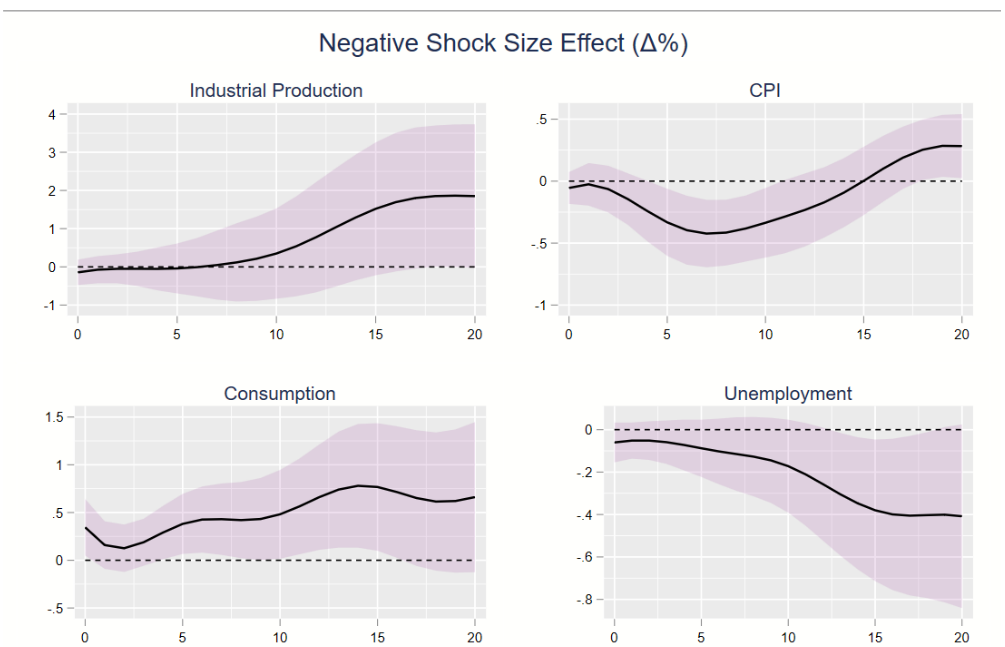

Figure 11: Indicator Function Approach with MP1 Shocks

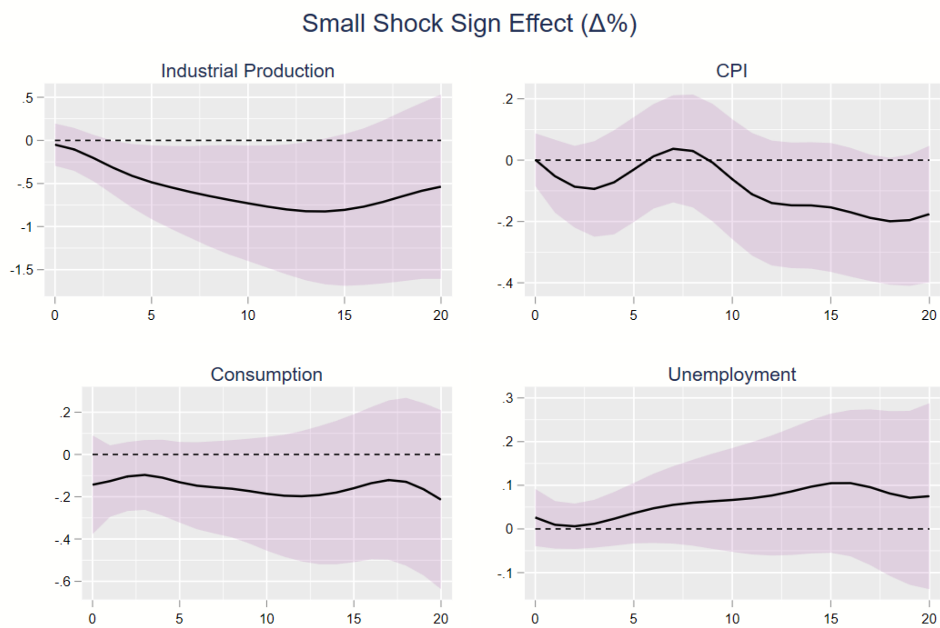

Figure 12: Indicator Function Approach with MP1 Shocks

Equilibrium Model Estimation Details

More details, in particular on the model, can be found in the Online Appendix.

For the set of draws that came out of our Metropolis- Hastings routine, I simulated data of 400 observations for each group of parameters to align with the US data sample size. The instrument in this case is $z_t = i_t - \mathbb{E}_{t-1}[i_t]$ to match the construction of monetary surprise measures in the literature.[20] Analogous control variables are included (lagged interest rates, zero lower bound, unemployment, output and interest rate variance) and plots are in terms of standard deviations to abstract away from any differences between model-simulated and US data.

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

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

[20] I use i to distinguish the model object R from its observation equation analogue.