8-2-2022

Embedding Rational Expectations in a Structural VAR: Internal and External Instruments for Set Identification

Zhengyang Chen

Victor Valcarcel

Embedding Rational Expectations in a Structural VAR: Internal and External Instruments for Set Identification*

Zhengyang Chen† Victor J. Valcarcel‡

First Draft: March 18, 2022
This Draft: Aug 2, 2022

Abstract

We propose a novel approach that embeds Rational Expectations (RE) into a low-dimensional structural vector autoregression (SVAR). We establish an instrumental variable procedure internal to the SVAR founded on a purely theoretical framework, which does not rely on any mapping strategy to a reduced form. Alternatively, a separate strategy considers data external to the SVAR to aid in the identification of structural shocks on a purely empirical basis. We report *clouds* of responses from an RE-consistent theoretical model as well as regions of plausible responses from the empirical approach. We conclude that a Taylor Rule characterization of monetary policy shocks remains relevant when the theoretical RE-SVAR is properly augmented with information from fluctuations—or momentous events—in markets that garnered increased attention since 2008, such as reserves and various money markets.

JEL Classification: E3, E4, E5

Keywords: Monetary Policy, rational expectations VAR, external variable instruments, sequential identification, event-based shock identification, regime-dependent monetary transmission

*Acknowledgments deferred for review.

†St. Cloud State University. Email: zhengyang.chen@stcloudstate.edu

‡University of Texas at Dallas. Email: victor.valcarcel@utdallas.edu

1 Introduction

Far from a typical economic disruption, the Great Financial Crisis (GFC) of 2007 proved to have lasting ramifications for the conduct and transmission of monetary policy shocks. The conventional view for these shocks over the previous two decades called for a characterization based on the Taylor Rule. In accordance with this rule, the Federal Reserve lowered the target federal funds rate to its effective lower bound (ELB) of zero in response to deteriorating economic conditions in late 2008.

The following decade saw the Federal Reserve turning to what became generally known as unconventional monetary policy (UMP) comprising: (i) liquidity injections into the financial system through novel facilities, (ii) engaging in large-scale asset purchases and balance sheet unwinds, and (iii) adopting forward guidance in its own communications to the public so as to better anchor market expectations.

In this historical context, structural vector autoregression (SVAR) models of monetary policy should address (i) the important role of expectations in this period and (ii) the role a Taylor Rule retains in a protracted near-zero interest rate environment. On the first point, the revolutionary work of Lucas (1972) has had lasting implications on macroeconomic modeling. Whereas dynamic stochastic general equilibrium (DSGE) models are essentially founded on rational expectations (RE), SVAR models' connection with RE has typically been more ephemeral. Regarding the second point, an important question to address is whether SVARs can preserve the relevance of a Taylor Rule—when accompanied by a Phillips curve (PC) and investment-savings (IS) equations—for this period.

This paper proposes a novel SVAR approach that directly embeds expectations into a low-dimensional "consensus model," which we refer to as "textbook." Alternatively, we consider a separate data-driven strategy that looks to information external to the SVAR to aid identification. This permits comparison between a purely theoretical and purely empirical identification of monetary policy shocks through the lens of theoretical clouds of impulse response functions versus empirical sets of responses. Finally, we combine both approaches and conclude that a theoretical model—when augmented with identification strategies surrounding events that took place during the GFC and UMP periods—salvages or at least sustains the relevance of the Taylor Rule for characterizing monetary policy shocks in a modern sample.

The rest of this paper is organized as follows. Section 2 provides some background on modeling RE in DSGE and SVAR models. Section 3 establishes some notation to distinguish the theoretical and the empirical methodologies we advance. Section 4 describes our procedure for embedding RE directly into an SVAR and presents results from this purely theoretical approach. Section 5 proposes overlaying various (size/event/external variable) restrictions to the RE-SVAR of the previous section. Section 6 discusses results from this augmentation of the purely theoretical model by overlaying the added restrictions proposed. Section 7 outlines the purely empirical approach imposing the same overlaying restrictions applied to the theoretical RE-SVAR. Section 8 presents results from the empirical approach and Section 9 concludes.

2 A Fundamental Disconnect in the Modeling of Rational Expectations: DSGEs vs. SVARs

Beginning with Blanchard and Khan (1980), there has been a large literature on the methodology for solving linear rational expectation models. Binder et al. (1995), Binder and Pesaran (1997), Klein (2000), Sims (2002) among others, paved the way for incorporating RE into DSGE models. These models typically begin with a canonical multivariate single equation relating variables of interest (say, x_t) to their expected future paths $E_t x_{t+1}$, as well as other exogenous variables and shocks (ϵ_t). Importantly, variables in x_t may not generally be observable. This is a broad and general framework capable of accommodating high-dimensional models. Therefore, the RE groundwork that underwrites the DSGE approach has great potential for describing a dizzying array of economic dynamics.

Conversely, given that VAR models are inherently "backward-looking," SVARs are generally ill-equipped to accommodate RE.[1] Partly, in an effort to address the backward-lookingness of VARs, a relatively small literature has engaged in looking for conditions under which the state-space framework that underwrites many DSGEs can be mapped into a VAR/VARMA representation (see Fernández-Villaverde et al. (2007), Ravenna (2007), Morris (2016), Morris (2017), and Martínez-García (2020) as a non-exhaustive list.) This is an admirable pursuit, which continues to grow. However, this mapping approach never scales from a VAR with expectations into a DSGE. Instead, these methodologies largely begin from a higher-dimensional RE-DSGE to a lower-dimensional VAR *representation* that

[1] A notable exception is Keating (1990).

would not accommodate RE unless: a moving average process were also appended, lags were truncated, or some algebraic mechanism for dimension reductionality were advanced.

A prototypical n-variable SVAR begins with the data from which reduced-form residuals are extracted. Subsequently, the researcher appeals to economic theory, some market mechanism(s) or empirical regularities about the data for help to pin down a mapping matrix connecting the residuals to some semblance of economic interpretation. Responsible researchers will not want to conclude their assumptions[2] so dynamics that are imposed by construction are often de-emphasized in favor of results that are driven by the combination of the researcher's identifying restrictions and the data. Thus, the data hopefully yields insight on economic dynamics and the identifying restrictions yield a credible identification of economic (a.k.a structural) shocks.

Differences of opinion among researchers may arise on various market mechanisms, theoretical models, or even the importance of empirical regularities. Therefore, far from incontrovertible, identification schemes are always debatable. An important aspect of the suitability of a given identifying restriction scheme rests on its ability to render orthogonalization of the innovations in the system. Thus, much of the SVAR analysis centers on satisfying arguments that propitiate a lack of contemporaneous correlations among shocks. Importantly, the orthogonality of shocks is a necessary but not a sufficient condition for the identification of VAR innovations as structural shocks.[3]

Some approaches discipline the SVAR identification with a theoretical model that is estimated a priori. Conclusions drawn from the theory are then incorporated ex-post as a rationalization for a plausible parameterization of the requisite elements in, say, some impact matrix. Other common practices include calibrating or estimating a DSGE model and subsequently looking for restrictions that match (or minimize the distance between) the impulse responses of a VAR to those of the DSGE, as well as finding conditions under which a DSGE can be represented as an SVAR (see various references cited above on this point).

[2]See Uhlig (2005) for a cautionary tale surrounding sign restrictions in this context.

[3]For example, if a restriction scheme had credibly identified innovations from a bivariate VAR as structural—say nominal versus real—shocks, appending a white noise process as a third shock would almost surely guarantee orthogonality. Yet attributing an economic interpretation to this third shock and considering it structural for the economy would strain credulity.

We propose an identification strategy founded on RE that is *directly* derived—rather than loosely mapped—from a theoretical construct. Incorporating RE in a VAR setting often involves mapping a theoretical model through a restriction strategy of a reduced-form (statistical) VAR. In a first, we insert the theoretical model directly into a VAR construct without the need for pinning down the $n(n-1)/2$ requisite elements in a matrix that connects statistical innovations to structural shocks. In other words, our approach does not involve the common practice of estimating a statistical VAR and then imposing a restriction scheme. Instead, we never deal with a reduced-form VAR. Our VAR construct begins with structural shock identification out of the gate. Thus, it is structural from the outset.[4] The most important benefit we can see is that it allows us to model forward-looking behavior directly within the VAR. Forward-looking agents are a typical assumption in theoretical macroeconomic models, but standard VARs are (mathematically speaking) backward-looking. Most analyses require clever identification strategies as well as external proxies/instruments from survey data to overlay "forward-lookingness" of the theoretical model to the "backward-lookingness" of the statistical model. Our approach allows us to circumvent this disconnect.

Another advantage of our approach—over many DSGEs—for identifying RE-consistent structural shocks is that our RE-SVAR method does not require adding unobservables to the information set. Therefore, there is no need for casting the model into state space. If the setup does not require unobserved state variables, there is no need to specify a transition equation. The ensuing measurement equation is fully observed, which accommodates low-dimensional modeling in a way that DSGEs cannot typically achieve.

[4]The validity of a reduced-form VAR is rarely called into question for assessing the suitability of a model. Rather, the scrutiny typically rests on the restriction strategy itself. Thus, if the restriction scheme is doubted, one may call into question the mapping of the innovations to the structural shocks. Our approach, however, does not rely on such mapping. The scrutiny here must rest on the suitability of our theoretical construct. For example, if one did not believe in a consensus $AS\text{-}IS\text{-}MP$ model, or if one did not believe in the particular characterization of the Taylor Rule we employ, then our approach would be a non-starter.

3 A Unified Identification Framework: Motivating Theoretical and Empirical Restrictions

Consider the following n-variable statistical VAR of order p:

$$x_t = \sum_{i=1}^{p} B_i x_{t-i} + e_t \tag{1}$$

with reduced-form parameters stacked into $B = (vec(B_1)'...vec(B_p)')'$ and statistical inovations $e_t \sim (0, \Omega_e)$. The covariance matrix of residuals can be decomposed according to $\Omega_e = PP'$, where P is the unique lower triangular Cholesky factor with nonnegative diagonal elements. Let $\boldsymbol{A}$ denote an invertible matrix that relates the reduced-form innovations $\boldsymbol{e_t} = (e_t^1, e_t^2, ..., e_t^n)'$ to a set of structural shocks $\boldsymbol{\epsilon}_t = (\epsilon_t^1, \epsilon_t^2, ..., \epsilon_t^n)'$. SVAR models provide a simple framework that enables researchers to quantify macroeconomic effects of disturbances without a full characterization, from first principles, of all the theoretical laws of motion involved. The prototypical n-variable SVAR analysis rests on attaining sufficient information about the matrix $\boldsymbol{A}$ that uniquely maps a set of reduced-form statistical innovations e_t, stemming from the data, to a set of mutually uncorrelated economic shocks ε_t:

$$e_t = \boldsymbol{A}\varepsilon_t \tag{2}$$

The $\hat{B}_i$ coefficients can be estimated from the reduced-form VAR and the sample residuals $\hat{e}_t(\hat{B})$ in (1) are consistent estimates of e_t. Given that the $\boldsymbol{A}$ matrix contains n^2 elements and the covariance matrix of the residuals $\hat{e}_t$ estimated in (1) only provides information from $n(n+1)/2$ elements, identification of all the shocks in the system requires placing assumptions on $n(n-1)/2$ elements, leaving the rest of the $\boldsymbol{A}$ matrix to be estimated from the data.

We depart from this standard approach by considering two cases. One case involves imposing fewer than $n(n-1)/2$ restrictions for a system that cannot guarantee point identification of responses. We refer to this as a purely empirical identification scheme. The other case is based on writing a fully specified theoretical model. In principle, this approach could potentially be consistent with full knowledge of the mapping matrix. We denote this theoretic mapping matrix as $\boldsymbol{A}^+$ to differentiate it from the $\boldsymbol{A}$ matrix, which only requires knowledge of a portion of the total number of its elements. We refer to this as the purely theoretical identification scheme and we formulate it to be consistent with a particular specification of an RE model.

We denote restriction schemes founded on theoretical constructs as $\bar{\boldsymbol{F}}$ to distinguish them from other types of restrictions collectively represented as $\bar{\boldsymbol{G}}$. Importantly, the theoretical construct is conditioned on what would be consistent with an identification strategy based on full knowledge of the mapping matrix $\boldsymbol{A}^{+}$. Thus, our proposed strategy conditioned on a full specification of the structural parameters guarantees a unique realization of impulse responses stemming from $\bar{\boldsymbol{F}}(\boldsymbol{A}^{+})$. We compare our approach to restrictions schemes $\bar{\boldsymbol{G}}(\boldsymbol{A})$ that do not yield point identification.[5] These latter types of restrictions obtain non-unique solutions giving rise to large sets of impulse responses that require further winnowing to be informative. Thus, we consider a spectrum from a purely theoretical identification restriction ($\bar{\boldsymbol{F}}(\boldsymbol{A}^{+})$) to a purely statistical one ($\bar{\boldsymbol{G}}(\boldsymbol{A})$). We offer some elaboration on trade-offs between the two approaches below.

When combining this expression (2) with the data, implicit regularity conditions about the nature of covariances yield information on $n(n+1)/2$ elements of the n x n matrix A. Therefore, insofar as this (2) mapping is informative, it acts as a restriction, albeit mild, of the parameter space—a point emphasized by Ludvigson et al. (2021) who refer to this as a covariance restriction. For a given identification scheme Z, let $\bar{\boldsymbol{G}}_{Z}(\boldsymbol{A})$ denote this mild covariance restriction as follows:

$$\bar{\boldsymbol{G}}_{Z}(\boldsymbol{A}) \equiv vech(\hat{\Omega}_{e}) - vech(\boldsymbol{A}\boldsymbol{A}') = 0 \tag{3}$$

where the operator $vech(\bullet)$ stacks the lower-triangular elements of a symmetric n x n matrix into a single vector of length $n(n+1)/2$. This covariance restriction is not enough for structural identification as there can be infinitely-many solutions that satisfy $\bar{\boldsymbol{G}}_{Z}(\boldsymbol{A})$=0. Let this set of empirical solutions be collected into:

$$\hat{\mathbb{A}}_{E} = \{\boldsymbol{A}, diag(\boldsymbol{A}) \geq 0, \bar{\boldsymbol{G}}_{Z}(\boldsymbol{A}) = 0\} \tag{4}$$

Similarly, let a set of theoretical solutions be collected into:

$$\hat{\mathbb{A}}_{T} = \{\boldsymbol{A}^{+}, diag(\boldsymbol{A}^{+}) \geq 0, \bar{\boldsymbol{F}}(\boldsymbol{A}^{+})\} \tag{5}$$

Both the theoretical and empirical solutions may yield large sets of impulse responses, which will subsequently be disciplined further with a set of overlaying restrictions. These overlaying restrictions could be useful if they help narrow down the response sets. Conversely, these added restrictions may be uninformative if: (i) they do not serve to winnow out the response region enough for a qualitative assessment of the direction of response, or (ii) if the

[5] See Ludvigson et al. (2021) for a similar setup.

winnowing is so extreme that it yields an empty set of possible responses. Importantly, a restriction scheme that sufficiently thins out the region of responses does not guarantee that the restriction itself is sensible. Solutions that satisfy the constraints can be found to exist and still yield nonsensical conclusions. Any overlaying response that serves to narrow out the set of solutions to $\mathbb{A}_E$ (or $\mathbb{A}_T$) can still be debatable on empirical or theoretical grounds.

We first discuss the theoretical solutions $\hat{\mathbb{A}}_T$ by elaborating on a specification based on $\bar{\boldsymbol{F}}(\boldsymbol{A}^+)$. For notational ease, we simply refer to the theoretical framework as $\bar{\boldsymbol{F}}$ with the understanding that this identification strategy is based on the full knowledge of the system. Later in the paper we discuss the approach for the empirical solutions $\hat{\mathbb{A}}_E$.

4 A Structural VAR Framework Consistent with Rational Expectations

Our restriction strategy $\bar{\boldsymbol{F}}$ is premised on a theoretical construct described in equations 4.7—4.10 below. We consider a VAR specification, which consists of four variables, measured at monthly frequencies, stacked in $x_t = [i_t, \pi_t, y_t, b_t,]'$ where i_t is the monthly average of the Wu and Xia (2016) shadow effective federal funds rate, π_t is the annualized inflation rate from the personal consumption expenditures (PCE) index, y_t is the natural log of industrial production, and b_t is the Gilchrist and Zakrajšek (2012) excess bond premium (EBP).[6]

4.1 A Purely Theoretical $\hat{\mathbb{A}}_T$ Identification Framework

We propose an identification strategy that pins down shocks to the monetary policy indicator with a method that does not rely on a Cholesky decomposition. Instead, our proposed technique orders the policy function first in the system. We leverage a simple application of the RE methodology to find a statistical relationship between the reduced-form innovations in the policy equation and the associated structural shocks. Once we have isolated the monetary policy shock, we can obtain dynamic effects for the short-term rate, inflation, output, and excess bond premia. These shocks may subsequently be used as an instrument to identify other structural parameters from our model, which in turn allows us to pin down

[6] We report results from a monthly sample encompassing 1988:m10—2020:m2.

other structural shocks.[7]

This setup poses a novel (internal) instrumental variable methodology for modeling rational expectations directly in a structural VAR setting. The typical approach estimates a reduced-form VAR first, and then advocates a mapping to the structural shock of interest. In a way congruent with the persuasive descriptions in Arias et al. (2019), the approach in our paper (as in many other VAR applications) must deal with the joint problem of VAR modeling: *statistical uncertainty* and *model uncertainty*. In essence, our VAR construct takes model uncertainty off the table. We begin with the assumption that the structural model we generate responses from is appropriate (if it is not, it renders the whole enterprise a non-starter). Contingent on this "accepted" structure, we generate impulse responses by directly imposing values on the structural parameters in what becomes a pseudo-calibration exercise.

Each response we report is a separate realization of a distinctly identified structural VAR. This allows us to produce a "cloud" of structural responses, each of which is unique[8] for a given value of the structural parameters.

Consider the following structural VAR

$$A_0 x_t = \sum_{i=1}^{p} A_i x_{t-i} + \varepsilon_t \tag{4.6}$$

where p is the number of lags and where $\mathrm{E}(\varepsilon_t \varepsilon_t')$ is a diagonal covariance matrix of the structural shocks.

[7]Our methodology describes a way to identify monetary policy (MP) shocks first. We use some of that information to subsequently identify investment-savings (IS) shocks. Then, information derived from the identification of MP and IS can be leveraged to identify the third innovation as an aggregate supply (AS) shock. Finally, armed with shocks in MP, IS, and AS, we identify bond risk (BR) shocks. While this paper limits the application of our approach only to MP shocks, identification of the remaining shocks remains crucial for the various overlaying restrictions schemes to monetary policy that we subsequently consider.

[8]This voids the need for the construction of confidence bounds. While ours is a frequentist approach, Inoue and Kilian (2020) argue against constructing confidence bounds around median responses in Bayesian VARs. They also advance the notion of reporting clouds of responses to denote a credible set.

The structural model consists of four equations:

$$i_t = \phi_\pi \mathbb{E}_t \pi_{t+h_\pi} + \phi_y \mathbb{E}_t y_{t+h_y} + A^{MP}(L)x_{t-1} + \varepsilon_t^{MP} \tag{4.7}$$

$$y_t = \mathbb{E}_t y_{t+1} - \alpha_1(i_t - \mathbb{E}_t \pi_{t+1}) + A^{IS}(L)x_{t-1} + \varepsilon_t^{IS} \tag{4.8}$$

$$\pi_t = \alpha_2 \mathbb{E}_t \pi_{t+1} + \alpha_3 y_t + A^{AS}(L)x_{t-1} + \varepsilon_t^{AS} \tag{4.9}$$

$$b_t = \alpha_4 \mathbb{E}_t b_{t+1} + \alpha_5(i_t - \mathbb{E}_t \pi_{t+1}) + \alpha_6 \mathbb{E}_t y_{t+1} + A^{BR}(L)x_{t-1} + \varepsilon_t^{BR} \tag{4.10}$$

where h_π and h_y refer to the number of forward-looking horizons in the policy reaction function to inflation and output, respectively, and A^{MP}, A^{IS}, A^{AS}, and A^{BR} are the autoregressive matrices containing the structural parameters.

Let the reduced-form VAR in companion form be given by $X_t = \beta X_{t-1} + De_t$ where $X_t = [x_t', x_{t-1}', ..., x_{t-p-1}']'$ is $np \times 1$, and $D = (I_n, 0_n, \ldots, 0_n)'$ is $np \times n$ and

$$\beta = \begin{pmatrix} B_1 & B_2 & \cdots & B_{q-1} & B_q \\ I_n & 0_n & \cdots & 0_n & 0_n \\ 0_n & I_n & \cdots & 0_n & 0_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_n & 0_n & \cdots & I_n & 0_n \end{pmatrix}.$$

is $np \times np$. The matrix I_n is an $n \times n$ identity matrix and the matrix 0_n is an $n \times n$ matrix of zeros.

Define a selection vector S_v such that

$$S_v X_t = v_t \tag{4.11}$$

where v_t is some component of X_t (such as i_t, π_t, y_t, or b_t, in our model above). We can forecast, or rationally expect the movement of a given variable using the VAR:

$$\mathbb{E}_t v_{t+j} = \mathbb{E}_t S_v X_{t+j} = S_v \beta^j X_t \tag{4.12}$$

Equation (4.12) follows from the fact that, via recursive substitution, it can be shown that $X_{t+j} = \beta^j X_t + \beta^{j-1} e_{t-1} + \beta^{j-2} e_{t+2} + \cdots + e_{t+j}$. Given equation (4.12), along with the companion form of the reduced-form VAR, and the assumption that $\mathbb{E}_t e_{t+i} = 0$ for $i > 0$, the following equation holds:

$$\mathbb{E}_{t-1} v_{t+j} = \mathbb{E}_{t-1} S_v X_{t+j} = S_v \beta^j \mathbb{E}_{t-1} X_t = S_v \beta^j \beta X_{t-1},$$

it is, then, straightforward to show the expectational, or forecast, revision is given by:

$$\mathbb{E}_t v_{t+j} - \mathbb{E}_{t-1} v_{t+j} = S_v \beta^j X_t - S_v \beta^j \beta X_{t-1} = S_v \beta^j (X_t - \beta X_{t-1}) = S_v \beta^j De_t \tag{4.13}$$

We use this general result to identify the structural VAR model defined above. We will use a sequential instrumental variable estimation strategy.

4.1.1 Monetary Policy Shocks

Taking a stand on the coefficients in the first equation, we can derive the monetary policy shocks by expressing them as a linear combination of the reduced form residuals without estimating any structural parameters. The policy feedback rule is given by:

$$i_t = \phi_\pi \mathbb{E}_t \pi_{t+h} + \phi_y \mathbb{E}_t y_{t+h_2} + A^{MP}(L) x_{t-1} + \varepsilon_t^{MP} \tag{4.14}$$

Rewriting the (4.14) equation in expectational difference form and subtracting the expectation of the policy rule at time $t-1$ from (4.14) yields:

$$i_t - \mathbb{E}_{t-1} i_t = \phi_\pi \left(\mathbb{E}_t \pi_{t+h} - \mathbb{E}_{t-1} \pi_{t+h} \right) + \phi_y \left(\mathbb{E}_t y_{t+h_2} - \mathbb{E}_{t-1} y_{t+h_2} \right) + \varepsilon_t^{MP} \tag{4.15}$$

We can then use equation (4.13) to find expressions in the expectational difference in the above equation, and solving for the structural shock in the interest rate feedback rule obtains the following:

$$\varepsilon_t^{MP} = e_t^i - (\phi_\pi S_\pi \beta^h De_t)' - (\phi_y S_\pi \beta^{h_2} De_t)' \tag{4.16}$$

4.1.2 IS Shocks

Similar to the analysis above, taking $\mathbb{E}_{t-1}$ of the IS equation and subtracting it from the IS equation renders the following expectational difference:

$$y_t - \mathbb{E}_{t-1} y_t = \mathbb{E}_t y_{t+1} - \mathbb{E}_{t-1} y_{t+1} - \alpha_1 ((i_t - \mathbb{E}_{t-1} i_t) - (\mathbb{E}_t \pi_{t+1} - \mathbb{E}_{t-1} \pi_{t+1})) + \varepsilon_t^{IS} \tag{4.17}$$

We can then use Equation (4.13) to find expressions for the expectational differences in the above equation:

$$e_t^y = (S_y \beta De_t)' - \alpha_1 (e_t^i - (S_\pi \beta De_t)') + \varepsilon_t^{IS} \tag{4.18}$$

We can rewrite this as a linear equation with slope coefficient α_1:

$$e_t^y - (S_y \beta De_t)' = -\alpha_1 (e_t^i - (S_\pi \beta De_t)') + \varepsilon_t^{IS} \tag{4.19}$$

Given that the error term in this equation may be correlated with e_t in general, and e_t^i in particular, OLS estimates of α_1 will generally be biased. However, leveraging (4.16), we can use ε_t^{MP} as an instrument (correlated with e_t^i but uncorrelated with ε_t^{IS}) to gain unbiased estimates of α_1. Once this is done, a time series for ε_t^{IS} can be recovered.

4.1.3 *AS* Shocks

Repeating the analysis yet again, we have the following expectational difference for the *AS* equation:

$$\pi_t - \mathbb{E}_{t-1}\pi_t = \alpha_2(\mathbb{E}_t\pi_{t+1} - \mathbb{E}_{t-1}\pi_{t+1}) + \alpha_3(y_t - \mathbb{E}_{t-1}y_t) + \varepsilon_t^{AS}$$

Once more, we can apply equation (4.13) to find expressions in the expectational difference in the above equation:

$$e_t^{\pi} = \alpha_2(S_{\pi}\beta De_t)' + \alpha_3 e_t^y + \varepsilon_t^{AS} \tag{4.20}$$

This is a linear equation with slope coefficients α_2 and α_3. Again, OLS estimates of α_2 and α_3 from this equation will generally be biased, but we can use ε_t^{MP} and ε_t^{IS} as instruments to gain unbiased estimates of the coefficients. Once this is done, a time series of ε_t^{AS} can be obtained.

4.1.4 *BR* Shocks

Finally, we can apply equation (4.13) to ultimately arrive at an equation for the structural shock to the excess bond premium as follows:

$$\varepsilon_t^{BR} = e_t^b - \alpha_4(S_b\beta De_t)' + \alpha_5(e_t^i - S_{\pi}\beta De_t)' - \alpha_6(S_y\beta De_t)' \tag{4.21}$$

Given the possible bias in the estimates of α_4, α_5, and α_6 that may result from the endogeneity of e_t^b, e_t^i and e_t more generally, we conduct two-stage least squares using the identified structural shocks to the previous equations as instruments. In the first stage, we estimate three regressions, one for each of the three terms in brackets in equation (4.21), and derive three OLS coefficients. In the second stage, we regress the residuals e_t^b from the fourth equation on the combinations of the first-stage estimates and the structural shocks to obtain sample estimates for $\hat{\alpha_4}$, $\hat{\alpha_5}$, and $\hat{\alpha_6}$. A time series for ε_t^{BR} can then be generated by replacing the coefficients α_4, α_5, and α_6 with their corresponding sample estimates.

4.1.5 Structural Shocks to the System

This framework for sequentially constructing structural shocks in this system provides a way to specify an SVAR that is consistent with rational expectations (RE-SVAR). Our methodology allows us to produce impulse response functions to all four shocks in the system. While we focus our attention exclusively on monetary policy shocks, recovering the full set of structural shocks remains crucial—particularly when we consider overlaying restrictions that exploit the empirical characteristics of these other shocks as a way to augment our identification strategy with empirical restrictions.

4.2 Operationalizing the RE-SVAR Framework

Our restriction strategy $\bar{\boldsymbol{F}}$ is premised on a theoretical construct described in equation (4.16), which shows a way to construct a series of structural (ε_t^{MP}) shocks. Rather than estimating values of the structural parameters in equation (4.14), we opt for a pseudo calibration approach. Given a time series construction of ε_t^{MP} as described in the previous section, along with given values for ϕ_π, ϕ_y, and a given h_π and h_y, we can compute a unique realization of the responses of variables in x_t to shocks in ε_t^{MP}. Throughout the analysis, we consider an exogenous standard deviation increase in the federal funds rate, substituting it with the Wu and Xia (2016) shadow rate for those periods when the ELB binds.

We proceed as follows. We produce a response for each variable of interest to an ε_t^{MP} shock by imposing a value for ϕ_π, ϕ_y, h_π, and h_y. We record the response and repeat the analysis. We iterate over a relatively fine grid search of values for these parameters. We let the ϕ_π coefficient cycle between values of zero and five in (1/15) increments. We pose a similar treatment on the output gap coefficient, with a different response for each $\phi_y = 0, 0.0667, 0.1333, 0.2, 0.2667, \ldots, 4.9333, 5$. This grid search is motivated by robust evidence in the empirical literature of structural change in the Federal Reserve's systematic response to economic fluctuations. Coibion et al. (2012) conduct a similar search over a range of values for ϕ_π but in the context of welfare gains. Additionally, the conventional wisdom in the empirical literature is that the Federal Reserve has become more forward-looking regarding inflation in modern times. Quantifying this degree of forward-lookingness is a difficult proposition at best. We address this concern in an imperfect way by also letting the horizon (h_π) of inflation expectations—as well as output expectations (h_y)—in the policy feedback rule (4.14) take on values between zero and 12 months.

Based on the restriction scheme $\bar{\boldsymbol{F}}$, we produce impulse responses to ε_t^{MP} shocks from a total of **976,144 different structural VAR specifications** (comprising a combination of 76 possible values of ϕ_π, with another 76 possible values for ϕ_y, 13 potential horizons ($h_\pi = 0, .., 12$) for inflation expectations, the first expectation term in equation (4.16), and another 13 ($h_y = 0, .., 12$) for output in the second term.[9] All of these (976K) permutations of parameters render clouds of uniquely identified structural responses.[10]

We tally up the incidence of puzzling responses with the following arbitrary heuristic.

Following a contractionary shock to the federal funds rate (ordered first in our specifications), we count as a puzzle any industrial production or PCE inflation response that shows a positive value at any time within the first year following the shock.

Any industrial production response at any, or all, horizons between impact and the twelfth period post-shock, we designate an output puzzle. We keep a separate count for the incidence of PCE inflation puzzles. There may often be realizations that show both of these puzzling responses. Therefore, we also report the sets of *surviving* responses that show neither puzzle. We call these the "*no (joint) puzzles responses,*" which exclude the incidence of either or both puzzles according to our criterion described above. In addition, we register the ϕ_π and ϕ_y values of the surviving responses.

4.3 Clouds of $\bar{\boldsymbol{F}}$-Rendered Impulse Responses

Figure 1 contains eight charts organized in two columns. Beginning with the left column, the Northwest (top left) column chart contains the responses of the policy indicator variable to its own shock. The solid-dotted yellow line is the median response across the 976K specifications, each of which is represented by a solid black line. This creates a cloud of 976K distinct responses. Some abnormally large responses dominate the scaling, creating the incorrect impression that the median response is zero. The remaining charts in the left column show clouds of puzzling responses (according to our heuristic above) for each of the variables in the system. Similarly, all charts from the second row to the end of the right

[9] AIC selected a lag of six for a VAR specification containing our variables based on a Cholesky ordering for our sample. For comparability, we fix this lag length across our 976K specifications.

[10] A few combinations of these hyperparameters yield outsized shapes of responses, which dominate the scaling but in all cases constitute a minute proportion. Dominated by these infrequent large-scale responses, there are often hundreds upon hundreds of responses (including the median responses of each set) that are superimposed (or nearly so) on each other.

column show clouds of, what we term, *surviving* responses—responses for each variable that do not show a puzzle in either inflation and/or output.

Each solid black line in the chart located in the second row of the left column shows a puzzling response (containing positive values at any time during the first year) of the PCE inflation rate. This chart also produces the median response across this black cloud of puzzling responses denoted by the solid-dotted red line. The chart includes a count of these puzzling responses expressed as a percent of the overall number of specifications—for this specification, 50.9% of the 976,144 responses show a positive response of inflation at any time during the first year to an exogenous increase in the federal funds rate. Finally, this graph also includes a solid-dotted green line denoting the median response of the nonpuzzling responses. For example, if the black cloud of these puzzling responses constituted 50.9% of the 976K total responses, the solid-dotted red line shows the median response across these, while the solid-dotted green line shows the median response of the remaining nonpuzzling 49.1% responses (not shown in this specific subchart).

The next subchart down the left column repeats the analysis for industrial production with a cloud of responses, each of which showing at least one positive value in the first 12 months. The red dotted line is the median response across these, whereas the green dotted line shows the median response in the cloud of sensible (non-puzzling) output responses (not shown in the subchart). The incidence of an output puzzle stands at a small 2.1% of the total 976K responses. Some of these puzzle responses, however, could overlap with an inflation puzzle showing a *joint* puzzle—specifications that show both an inflation *and* an output puzzle. Every one of the output responses we report in the small cloud of output puzzles stems from a specification that also shows an inflation puzzle. Thus, we find 50.9% of responses show an inflation puzzle, 2.1% of responses show an output puzzle, and the total percent of responses that show *neither* puzzle is 49.07% of the total 976K responses.

The subchart at the bottom of the left column shows a cloud of responses of the excess bond premium that stem from a specification that yields an inflation puzzle, an output puzzle, or both. This is a cloud containing 50.9% of these *joint puzzle* responses.

We now describe the right column. Each blue line in the subchart located in the second row shows an inflation response for a given specification that did not incur a puzzle in either variable. The label on the y-axis counts the number of these blue lines in this cloud, which corresponds to the responses that *"survive"* any puzzling behavior as we define it above—a

total of 478,952 responses constituting 49.1% of the total combinations. The solid-dotted green line shows the median of these sensible responses, which for our model, is roughly comparable with the solid-dotted green line on the corresponding chart on the left column. The next subchart down the column repeats the analysis for industrial production and the bottom right column subchart shows a blue cloud of responses of the excess bond premium that stem from a *no (joint) puzzle* specification. The bar chart at the top of the right column takes the number of surviving responses and collates them according to the values of ϕ_π and ϕ_y. For example, if we have a 49.1% survival rate, then the bar chart would distribute the resulting 478,952 sensible responses as follows. If the x-axis at zero for ϕ_y shows a light colored bar at 1,700, this would mean that 1,700 out of the 478K sensible responses had a value of zero for the output coefficient (in combination with possible values of $0 \leq \phi_\pi \leq 5$ and $h_\pi/h_y = 0, ...12$). Similarly, if the dark-colored bar at the six value of the x-axis for ϕ_π showed 478,952, there would be no other bars in the bar chart and it would mean all the sensible responses had a value of $\phi_\pi = 5$ (in combination with $0 \leq \phi_y \leq 5$ and $h_\pi/h_y = 0, ...12$).

Figure 1 shows that the Wu and Xia (2016) shadow rate yields clouds of responses for inflation, output, and EBP that can be roughly bisected into qualitatively puzzling (the black cloud of responses on the left column) and qualitatively sensible responses (the blue clouds on the right). We estimate a cloud of 478,952 sensible responses for each of these variables drawn in blue along with a median response (across the 479K for each step of the impulse response function) drawn as a green dotted-solid line. Focusing on these responses, the highest incidence of responses shows a value of ϕ_y between 0.5 and 1, with the incidence of values higher than 1 decreasing somewhat monotonically from 1 through 5. Whereas virtually no response shows a value of ϕ_π less than 0.5, with the incidence of values increasing monotonically from 0.5 through 5. Over 90% of the sensible responses show values of ϕ_π consistent with the Taylor Principle.

Overall, these results are highly stylized. They are conditioned on an economic structure—that is assumed *a priori* appropriate—in order to generate time series of structural shocks that are, subsequently, fed through a range of values of the structural parameters in a Taylor-type policy reaction function. Letting the data reveal the dynamics of the responses of interest is a useful approach to obtaining not merely theoretic-consistent responses from a mapping scheme, but responses that are derived directly from the theory. Another advantage of the approach is that it allows us to directly model, from the observables in the system,[11] the forward-lookingness of a monetary feedback rule—insofar as the Federal Re-

[11]Our approach to modeling forward-looking expectations differs from the traditional DSGE approach,

serve's forward-lookingness be encapsulated in the horizons of output and inflation.

Ours is a different methodology to the standard approach, which typically begins with a statistical VAR and subsequently imposes a restriction strategy for a plausible mapping between statistical innovations and structural shocks. The standard approach may be founded on relatively weak or rather uncontroversial restrictions. But they can also be prescriptively governed by overly restrictive mapping schemes. Implictly, our restriction strategy for generating clouds of responses from a theoretical solution set $\hat{\mathbb{A}}_T$ stems from within the system. Beginning from this RE-SVAR specification, we now turn to overlaying further restrictions from within our VAR system. But this time we restrict the realizations of the structural shocks themselves, rather than the laws of motion of the system. We will then append further restrictions from information external to the VAR.

5 Overlaying Restrictions

In deference to the notion that monetary policy may have experienced substantial regime changes throughout the period we investigate, we consider three separate classes of empirical restrictions. First, we examine the incidence of large shocks throughout our sample, which can be characterized as *shock size constraints*, and we denote these as $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$. We, then, focus on a portion of our sample, a period when UMP was prevalent. This second type of restriction involves *event constraints* where some *ex post* knowledge of a particular period may be suggestive of some feature of the structural shock of interest. Specifically, we focus on the *Quantitative Easing* (QE) and *Quantitative Tightening* (QT) events that took place in the aftermath of the GFC and beyond. We will denote these restrictions as $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ and $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$. Finally, we consider external variables to the VAR and impose a priori assumption for how they should correlate to the VAR variables. These *external variable constraints* are denoted by $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$.

5.1 Size Restrictions

We now inspect the structural shocks that our purely theoretical restriction strategy $\bar{\boldsymbol{F}}$ yields and look for periods when those shocks had realized large values. The idea is that if

which requires the addition of unobservables though a state equation onto a comparatively more restrictive search of the parameter space.

an identified shock is large, particularly around a period known to be fraught with instability associated with the shock of interest, one may more reasonably assume the shock has a material effect not spuriously introduced by a questionable identification strategy. This idea is introduced by Antolín-Díaz and Rubio-Ramírez (2018) and refined by Ludvigson et al. (2021).

Figure 2 registers the distribution of maximum values for each realization of the four structural shocks of the RE-SVAR. At first glance, none of these shocks exhibit Gaussianity in their distribution of maximum values. The distributions of ε_t^{MP} and ε_t^{IS} exhibit some skewness. Interestingly, a majority of the maximum values of the first three shocks occur in Sep-Oct of 2008 (94% of the maximum values of the first shock, 96% of the second shock, and 54% of the third shock are found on these dates). This is a period typically associated with high uncertainty surrounding the GFC when Lehman Brothers filed for bankruptcy and when the DJIA index fell roughly 20%. At 33%, the highest incidence of values for the shock to the excess bond premium is found in July 2013, which is roughly around a reactionary panic (commonly known as the *Taper Tantrum*) that triggered a spike in U.S. Treasury yields following a Federal Reserve announcement that it would slowly begin unwinding the QE program.

Motivated by these findings from the RE-SVAR structural shocks, we consider the following $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ restriction set:

$$\bar{\boldsymbol{G}}_1(\boldsymbol{A}) : \varepsilon^{AS}_{\{2008:10\}} \geq \tau^2 \quad \wedge \quad \varepsilon^{IS}_{\{2008:10\}} \geq \tau^3 \quad \wedge \quad \varepsilon^{BR}_{\{2013:7\}} \geq \tau^4 \tag{5.22}$$

where $\tau^2 = 0.313$, $\tau^3 = 0.805$, and $\tau^4 = 0.110$ which constitute the 75th percentile values of each respective shock on the corresponding date. As mentioned earlier, by construction these restrictions may not in any way add to the cloud of responses generated by $\bar{\boldsymbol{F}}$. The $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ restriction scheme may be informative if it helps in thinning out the cloud generated by $\bar{\boldsymbol{F}}$ and uninformative if it does not narrow down the set of responses. Every realization of the four shocks is obtained from a given combination of the hyperparameters outlined in the previous sections. Therefore, if a given realization of, say, ε_t^{AS} does not satisfy the restriction in $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$, not only does that particular realization of ε_t^{AS} exit the solution set $\mathbb{A}_T$ but, importantly, the accompanying ε_t^{MP}, ε_t^{IS}, and ε_t^{BR} for that same realization of the structural parameters are removed as well.[12]

[12]We could have also imposed the restriction scheme $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ on the first shock in the system ε_t^{MP} directly. However, to remain conservative given our interest in the dynamic responses to structural disturbances in the first variable, we opt not to impose the restriction on the first shock.

5.2 Event Restrictions: *QE* and *QT*

There is demonstrative work that, following 2008, the tools of monetary policy changed. See work by Gagnon et al. (2011); Krishnamurthy and Vissing-Jorgensen (2011); D'Amico et al. (2012); Wright (2012); D'Amico and King (2013); Carpenter et al. (2015); Ihrig et al. (2018); Swanson (2018); Swanson (2020); Bundick and Smith (2020); Vayanos and Vila (2021) and Christensen and Gillan (2022). This could have led to changes in transmission mechanisms of monetary policy. Our RE-SVAR is founded on a specification substantiated by a consensus model of the macroeconomy with a PC curve, an IS curve, and a monetary policy rule.

However, it could be argued that the relevance of the consensus model might have changed after GFC—hence the need for augmentation with added restrictions. Importantly, while there is ample evidence that the excess bond premium is useful and important in empirical work, there is a dearth of theoretical work on how to incorporate it into the consensus macroeconomic model (before and after GFC). This is highlighted by our choice to place the EBP in fourth place in our RE-SVAR, along with a milder restriction in equation (4.10) relative to that of equation (4.7). Therefore, we now turn to restriction schemes that mostly bind to the shock to the fourth variable (ε_t^{BR}) surrounding well-known events during the (post-2008) UMP period. We impose two restriction sets: $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ dealing with the Taper Tantrum, and $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$ dealing with QE episodes.

Following nearly half-a-decade-long accommodation in response to the GFC, the Federal Reserve began a gradual normalization effort. Table 1 shows two announcements by, then, Chairperson Bernanke, on the Fed's intent to eventually taper the pace of purchases, which led to a financial market reaction known as the *Taper Tantrum*. Later on, Table 1 shows that a plan for a gradual normalization is hinted at in May 2014 and begins in earnest in September 2014. This became known as the QT period, which extended from September 2014 to about August of 2019. There were two distinct subperiods of QT. From September 2014 through September 2017, the Federal Reserve reinvested proceeds of maturing securities. This *Full Reinvestment* phase resulted in declining reserves without a commensurate decline in asset holdings. Then, beginning in September 2017 until August 2019, the Federal Reserve purchased fewer assets than were maturing. This *Asset Runoff* phase resulted in declines in both reserves and the Fed's asset holdings.

Smith and Valcarcel (2020) outline a strong asymmetric effect of both phases of QT on financial interest rates relative to the earlier QE periods that took place prior to 2014.

Especially after the *Taper Tantrum* episode the FOMC engaged in a concerted effort to divorce expectations of future rate increases from unwinding the balance sheet. Therefore, while QE announcements typically contained a large signaling component, signaling effects were mostly absent from QT announcements. Consequentially, those authors show a significant response of various financial rates (treasury yields, corporate bond rates, MBSs, and Eurodollar) within a two-day window of the Taper Tantrum announcement, whereas QT announcements elicited no significant financial market response. Motivated by these conclusions, we consider the following $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ restriction scheme:

$$\begin{aligned}\bar{\boldsymbol{G}}_2(\boldsymbol{A}) : std(\varepsilon^{BR}_{\{2013:05-2013:07\}}) \geq 2 \times std(\varepsilon^{BR}_{\{2014:05-2014:07\}}) \quad \wedge \ldots \\ \wedge \quad std(\varepsilon^{BR}_{\{2013:05-2013:07\}}) \geq 2 \times std(\varepsilon^{BR}_{\{2017:09-2017:11\}})\end{aligned} \tag{5.23}$$

This restriction requires variation in the structural shock ε^{BR}_t around the June 2013 announcement related to the Taper Tantrum to be much larger than the variance around the first phase of the QT period in June 2014 and during QT's Asset Runoff phase in October 2017.

More broadly, the finding in Smith and Valcarcel (2020) that QT announcement effects on financial markets were negligible facilitates a comparison to a relatively large literature (see e.g. Gagnon et al. (2011), Krishnamurthy and Vissing-Jorgensen (2011), Bauer and Rudebusch (2014) among others) that finds important financial effects from QE announcements. However, the various QE episodes may have had materially different effects on interest rates of various financial markets. In a detailed review of the QE literature, Kuttner (2018) highlights an overall agreement that QE1 announcements had very large, negative effects on interest rates, whereas the effects of subsequent QE programs on yields were materially smaller. Thus, we consider the following $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$ restriction scheme:

$$\begin{aligned}\bar{\boldsymbol{G}}_3(\boldsymbol{A}) : std(\varepsilon^{BR}_{\{2008:11-2009:01\}}) \geq 3 \times std(\varepsilon^{BR}_{\{2010:08-2010:11\}}) \quad \wedge \ldots \\ \wedge \quad std(\varepsilon^{BR}_{\{2008:11-2009:01\}}) \geq 3 \times std(\varepsilon^{BR}_{\{2012:09-2012:11\}})\end{aligned} \tag{5.24}$$

This restriction requires variation in the structural shock ε^{BR}_t around the QE1 announcement to be much larger than the variance around the QE2 and QE3 periods.

5.3 External Variable Restrictions During the UMP Period

Substantial research on structural identification has often incorporated information that is external to the SVAR. Some of the information may stem from observable variables that augment the VAR with: a very large panel of real and financial factors for the US as in Bernanke et al. (2005), a panel of international factors as in Mumtaz and Surico (2009), or a small set of money market factors for the US as in Chen and Valcarcel (2021). Other approaches have used a narrative approach from unobservables. This has often involved constructing shock series from historical readings of political and economic events to be used as external instrumental variables as in Baker and Bloom (2013), Auerbach and Gorodnichenko (2013), Mertens and Ravn (2014), among others.

Much of this external variable instrument literature achieves point identification by assuming that the instruments have a zero correlation with some shocks (the traditional orthogonality assumption required for identification) and a nonzero correlation with others (an assumption on the informational content of the instrument that is relevant to the question at hand). Importantly, our overlaying restrictions $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$—$\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ may provide finer sifting of the clouds of structural responses (set identification) but do not necessarily guarantee point identification. Thus, we follow the reasoning by Ludvigson et al. (2021) in stipulating that our choice of external variables is not required to be valid exogenous instruments that have zero correlations with some of our identified shocks. Instead, we only require the milder condition that the random processes driving our chosen external variables be determined outside of the VAR system, while allowing the variables themselves to be partially correlated with some of the variables in the system. This facilitates further winnowing of our sets of structural responses.

We consider three variables external to our system: the log of bank reserves (R_t), the spread between the 10-year and 1-year treasury rates (TR_t), and the 1-month federal funds futures rate (ff_t). Data on federal funds futures has been prominent as a plausible instrument of policy shocks since the seminal work of Friedman and Kuttner (1992), Kuttner (2001) and, more recently, Gertler and Karadi (2015). The 10-to-1 year treasury interest rate spread is a popular measure of the term structure theory of interest rates and its relation to monetary policy shocks. Finally, a strong relationship between reserves balances and monetary policy has been established by: (i) Strongin (1995) during the Great inflation period, (ii) Carpenter et al. (2012) in the aftermath of the GFC, and (iii) Smith and Valcarcel (2020) during the first QT period of the late 2010s. We impose the following restriction

$$
\begin{aligned}
\bar{\boldsymbol{G}}_4(\boldsymbol{A}) : corr(\varepsilon_t^{MP}, R_t)_{\{2017:10-2019:09\}} \leq corr(\varepsilon_t^{MP}, R_t)_{\{2013:06-2017:10\}} \quad \wedge \ldots \\
\wedge \quad corr(\varepsilon_t^{BR}, R_t)_{\{2017:10-2019:09\}} \leq corr(\varepsilon_t^{BR}, R_t)_{\{2013:06-2017:10\}} \quad \wedge \ldots \\
\wedge \quad corr(\varepsilon_t^{BR}, ff_t)_{\{2008:04-2008:12\}} \leq 0 \quad \wedge \ldots \\
\wedge \quad corr(\varepsilon_t^{BR}, TR_t)_{\{2008:04-2008:12\}} \geq 0
\end{aligned}
\tag{5.25}
$$

where the first two inequalities relate to the QT period and the latter two bind to the GFC. The first inequality restricts the correlation between log reserves and monetary policy shocks to be more deeply negative during the Asset runoff phase of the QT period than during the previous phase of QT—which occurred between 2013:06, the month of the taper tantrum, and 2017:10 the month when the Fed begins the active phase of its balance sheet unwind. This choice is motivated by the finding in Smith and Valcarcel (2020) of a stronger liquidity effect at work during the second phase of the QT period than during the previous *Full Reinvestment* period (see Figure 3).[13] The second inequality restricts the correlation between log reserves and the BR shock to follow the same pattern as that of the MP shock. Again Smith and Valcarcel (2020) find overwhelming evidence of a tightening of financial conditions taking place across many financial markets during the asset runoff stage of QT with little indication of this dynamic occurring during the *Full Reinvestment* phase of QT.

The third inequality in (5.25) imposes a negative correlation between the BR shocks and the federal funds rate futures in the months between JP Morgan Chase purchase of the failing Bear Stearns and the Fed's onset of the ELB period. This is a period of financial turmoil when the EBP experienced substantial hikes even as markets were pricing a high probability that the Fed would conduct a massive expansionary policy. Finally, the fourth inequality restriction imposes a positive correlation between the term premium and the BR shock over this key period, when investors were likely flocking to the safety of treasuries as most financial condition indicators (including the EBP measure) were spiking.

[13] More recently, Sengupta et al. (2022) predict a more aggressive post-COVID-19 QT period, than the pre-COVID-19 period Smith and Valcarcel (2020) study, may have more portentous effects on financial conditions.

6 Evidence from an Empirically Augmented RE-SVAR Identification

Figures 4 – 6 show responses from our RE theoretically motivated scheme $\bar{\boldsymbol{F}}$, placed on the northwest corner of each figure. Essentially these are the 976,144 responses that were already discussed in Figure 1. The rest of the charts in each figure correspond to: *(i)* the restriction on the size of structural shocks ($\bar{\boldsymbol{G}}_1(\boldsymbol{A})$), *(ii)* the event restrictions surrounding the QT period ($\bar{\boldsymbol{G}}_2(\boldsymbol{A})$), *(iii)* the event restrictions surrounding the QE period ($\bar{\boldsymbol{G}}_3(\boldsymbol{A})$), and *(iv)* external variable restrictions described by ($\bar{\boldsymbol{G}}_4(\boldsymbol{A})$). The operative shock throughout these results is a contractionary exogenous standard deviation increase in the federal funds rate (augmented with the Wu and Xia (2016) shadow rate during the ELB period).

The grey areas in those figures are formed from all the applicable impulse response function estimates that conform to each identification strategy. The left chart on the top row of each figure shows the responses from all the 976K combinations of parameters in the RE-SVAR—the $\bar{\boldsymbol{F}}$ restriction scheme. The remaining charts in each figure show the responses from overlaying each of the $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ – $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ restriction strategies on to $\bar{\boldsymbol{F}}$. Thus, each $\bar{\boldsymbol{G}}(\boldsymbol{A})$ scheme will either present the same number of responses—if completely uninformative—or a reduced number of responses from the 976K in $\bar{\boldsymbol{F}}$. If the overlaying restrictions effectively provide a winnowing effect, the resulting figures will be delimited by a smaller region of responses.

Figure 4 shows inflation responses to a contractionary shock in the federal funds rate for the RE-SVAR $\bar{\boldsymbol{F}}$ model and the four overlaying restrictions. The left chart on the top row of the figure shows the 976K responses congruent with the $\bar{\boldsymbol{F}}$ scheme. The chart shows a relatively small (grey) region of responses that lay above zero within the first 12 months post shock. A much larger region shows a sensibly negative response of inflation to a contractionary monetary policy shock in the first year. The black and blue lines show a reduction in inflation following a monetary contraction and a return to zero about one-and-a-half years after the shock. This chart also reports the median values across all responses for the Taylor Rule coefficients and the median horizon in the expectation for each. The median values for ϕ_π and ϕ_y are 2.5 and 4.3, respectively, with the median horizons for both at 6 months. Our RE-SVAR specification is highly indicative of an active and forward-looking Fed, strongly following the *Taylor Principle* in its response to inflation.

The middle chart of the top row shows the inflation responses that—starting from RE-

SVAR described by the $\bar{\boldsymbol{F}}$ scheme—overlay the $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ restriction strategy based on the size of the shocks described earlier. This restriction winnows the 976K responses from $\bar{\boldsymbol{F}}$ down to 710,731 responses that satisfy the shock size restrictions. The responses from the $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ scheme show an even smaller region of puzzling responses in the first 6 months. The Taylor Rule coefficients for the median response across the 710K responses are still large and consistent with the Taylor Principle, with the horizon for inflation still at 6 months, but now the horizon for output is 10 months ahead. The mean, median and mode responses all show qualitatively sensible dynamics.

The top right chart shows the inflation responses for the $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ restriction schemes surrounding the QT events outlined earlier. At 23,540 responses that satisfy this overlying restriction strategy, $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ proves restrictive and informative as it serves to winnow out a very large number off of the original 976K responses. The region of inflation puzzles that stem from this scheme is virtually nil. The mean, median and mode responses show expected and sensible dynamics. The median values for ϕ_π and ϕ_y are 3.3 and 1.0, respectively, with the median horizons for inflation settling at two-months-ahead, and the median horizon for output at 11. In many respects, the QT restriction scheme in $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ might be the most sensible.

The bottom left chart shows the inflation responses for the $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$ restriction schemes surrounding the QE events outlined earlier. This restriction scheme vastly reduces the region of permissible responses from 976K to 81,117. However, the median and mode responses do show a puzzling response where inflation increases for the first four months following the contractionary shock. In addition, at 0.8, the median value for ϕ_π is too low for consistency with the Taylor Principle. Finally, the bottom right chart for the $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ restriction schemes based on information from the external variables, provides the largest reduction in responses from 976K down to merely 18,924. This scheme yields a negligible region of puzzling responses within the first year as well as sensible responses for the mean and median across the set. However, this scheme suggests no forward-lookingness in the rule for output stabilization ($h_y = 0$), and ϕ_π at a low value of 0.73.

Figure 5 shows the responses of industrial production (as our output measure) to contractionary policy shocks. The 976K RE-SVAR ($\bar{\boldsymbol{F}}$) responses on the left of the top row show a smaller region of industrial production lying above than below zero. All mean, median, and mode responses display the correct sign in response to a contractionary monetary policy shock. The middle chart on the top row shows the $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ size restriction set. There is substantial winnowing down to 71K responses, which successfully eliminate all those responses

that laid above zero in the RE-SVAR responses of the previous chart. The next chart for the QT scheme $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ reduces the region substantially, down to 23,540. And the mean and median responses look sound. However, the restriction seems to mostly thin out the sensible negative region from $\bar{\boldsymbol{F}}$ so that zero now bisects the set of responses roughly equally. The bottom left chart for the QE-motivated $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$ restrictions shows the vast majority of the 81,117 responses to be sensible. Finally, the bottom right chart for the $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$ restrictions based on external variables show mostly sensible, but with a nonnegligible positive portion of responses.

Figure 6 shows the excess bond premium responses to exogenous standard deviation increases in the federal funds rate for the RE-SVAR (at the top left) and the overlaying restrictions $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ – $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$. Here, there seems to be overwhelming qualitative agreement on the positive EBP response across restriction strategies. Perhaps the worst performing chart is that of the QT-motivated $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ restriction scheme where the region of responses within the first year is more or less evenly divided across zero, which suggests $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ is relatively uninformative for the EBP response to a federal funds rate hike within the first year.

Our RE-SVAR $\bar{\boldsymbol{F}}$ scheme elicits informative responses for all the variables in the system—insofar as the ensuing ranges of responses tend to fall either mostly north, or mostly south, of zero. The $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ scheme based on the size of the AS, IS, and BR shocks proves to be the least restrictive in that it reduces the number of responses produced by $\bar{\boldsymbol{F}}$ down by 27%. Still, this restriction scheme exhibits desirable properties in terms of the qualitative nature of the responses, the degree of forward-lookingness, and the adherence to the Taylor Principle in the (4.7) policy rule.

The remaining overlaying restrictions $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$, $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$, and $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ look to be far more restrictive as the cloud of responses satisfying these restrictions reduce the original cloud by an order of magnitude. The $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ scheme bearing on the BR shocks on dates surrounding the QT event reduces the number of responses from 976,144 down to 23,540. The ensuing response for inflation looks sensible regarding shape, coefficient and horizon values. The output (i.e. industrial production) response is, however fairly uninformative. While the mean and median responses across the 25K responses look *textbook*, there are many responses that predict a positive output reaction to a monetary contraction. Indeed the zero line seems to mostly bisect the set of responses, rendering them somewhat ambiguous. $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$, which operates on the BR shock surrounding events during the QE period shows some mixed results from the resultant 81,117 responses. For the most part, the regions of inflation and output

responses appropriately lie below zero. However, the mean and median inflation responses lie above zero for the first four months and the median value for $\phi_\pi = 0.8$ that is too low to satisfy the Taylor Principle. Finally, the external variable $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ strategy provides the most restrictive scheme with the smallest number of responses at 18,924. The dynamic responses look sensible but $\phi_\pi = 0.73$ is again too low.

Overall, most of the inflation and output responses to a monetary contraction are found to lie within a sensible range. And, by and large, the EBP responses are qualitatively robust across restriction strategies. This concludes the treatment that stems from the theoretical solution scheme $\mathbb{A}_T$, which consisted of overlaying empirical restrictions ($\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ – $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$) onto a purely theoretical ($\bar{\boldsymbol{F}}$) methodology from the RE-SVAR. We next turn to the purely empirical solution approach $\mathbb{A}_E$, where we begin with an empirical scheme $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ and overlay the same $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ – $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ restrictions.

7 A Purely Empirical $\hat{\mathbb{A}}_E$ Restriction Strategy

For many macroeconomic applications, requisite restrictions for $n(n-1)/2$ elements of $\boldsymbol{A}$ that guarantee the unique mapping in equation (2) may still be debatable, contentious, or otherwise simply unavailable. One way of circumventing the uniqueness issue and having to defend a theoretically-motivated restriction is advanced by Ludvigson et al. (2021). Theirs is an innovative approach that persuasively turns the identification paradigm on its head by generating large sets of candidate solutions for $\boldsymbol{A}$ that satisfy $\hat{\mathbb{A}}_E$.

There are trade-offs to the approach. A major advantage is that this method generates orthogonal shocks without having to defend any particular mapping (2), which is itself unobservable. Another advantage is that by beginning with a desirably weak restriction scheme (a large set of candidate solutions), their approach provides added flexibility to overlay further restriction schemes (as we do in our theoretical model) premised on empirical characteristics of the generated shocks—based on information inside and outside of the VAR system. A disadvantage of their framework is that it does not allow for point identification of the restriction scheme—a point emphasized in the paper. Since no mapping is uniquely identified, the technique does not produce point estimates of impulse response functions. That is, the mapping advanced is set-identified (rather than point-identified) rendering response regions that would contain uncountably many, *and equally likely*, potential combinations of line responses. Importantly, while the shocks generated from the solution set of $\boldsymbol{A}$ are orthogonal,

they may or may not have an economic interpretation.

The approach to solving for $\hat{\mathbb{A}}_E$ begins by initializing the $\boldsymbol{A}$ matrix with a lower-triangular Cholesky factor of $\hat{\Omega}_e$ with non-negative diagonal elements $\hat{\boldsymbol{P}}$. Then, a matrix of n x n random variables $\boldsymbol{M} \sim N(0,1)$ is drawn from which an ensuing orthonormal $\boldsymbol{Q}$ matrix from the $\boldsymbol{QR}$ decomposition of $\boldsymbol{M}$ such that $\boldsymbol{A} = \hat{\boldsymbol{P}}\boldsymbol{Q}$. This procedure is then repeated an arbitrarily large number of times (r) to collect a set of possible solutions:

$$\hat{\mathbb{A}}_E = \{\boldsymbol{A} = \hat{\boldsymbol{P}}\boldsymbol{Q} : \boldsymbol{Q} \in \mathbb{O}_n, diag(\boldsymbol{A}) \geq 0, vech(\hat{\Omega}_e) - vech(\boldsymbol{A}\boldsymbol{A}') = 0\} \tag{7.26}$$

where $\mathbb{O}_n$ is a set of n x n orthonormal matrices. Given the mapping in (2), r-many generated values for $\boldsymbol{A} \in \mathbb{A}_E$—constructed from r-many rotations of the matrix $\boldsymbol{Q}$—yield r-many unconstrained values of $\varepsilon_t^r(\boldsymbol{A}) = (\hat{\boldsymbol{P}}\boldsymbol{Q})^{-1}\hat{e}_t(\hat{\boldsymbol{B}})$ for $t = 1, 2, ...T$. As Ludvigson et al. (2021) point out, the set of (r) possible solutions of $\hat{\mathbb{A}}_E$ do not allow for point identification of $\boldsymbol{A}$ toward a unique identification of structural shocks $\varepsilon_t(\boldsymbol{A})$. Instead, a set of r-many shocks $\varepsilon_t^r(\boldsymbol{A})$ equally likely over the sample $t = 1, 2, ..., T$ can be generated. Subsequently, further restrictions can be overlayed to reduce the set of plausible $\varepsilon_t^r(\boldsymbol{A})$ responses.[14]

Thus, the first set of possible solutions to the mapping matrix $\mathbb{A}_E$ conforms to a very mild restriction of (7.26) consistent with the nonnegativity of the covariance matrix of the VAR. This is the starting point from where the region of possible responses cannot increase.[15] Subsequently, we consider further restrictions to potentially narrow down the potential sets of responses.

Following the QR factorization described in equation (7.26) we generate a total of **r=1 million rotations** of the $\boldsymbol{Q}$ matrix, which allows us to construct a set of one million shocks $\varepsilon_t^r(\boldsymbol{A})$ equally likely over the sample $t = 1, 2, ..., T$ for each of the four variables. Importantly, these shocks are not motivated by any theoretical assumption, no matter how innocuous. Instead, the ensuing shocks stem from the weak restriction $\bar{\boldsymbol{G}}_Z(\boldsymbol{A})$ corresponding to equation 3. As mentioned earlier, this covariance restriction is enough to generate large sets of empirical $\mathbb{A}_E$ solutions but not enough for structural identification. We refer to this starting point when the identification scheme $(Z = 0)$ is mildest as $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$—which is the analog empirically restricted starting point to our theoretically restricted $\bar{\boldsymbol{F}}$ from the RE-SVAR.

[14] See Ludvigson et al. (2021) and Antolín-Díaz and Rubio-Ramírez (2018) for restrictions schemes based on the size of shocks during certain historical events.

[15] See Ludvigson et al. (2021) for details on this feature.

Therefore, we begin our search of possible restrictions $\bar{\boldsymbol{G}}_Z(\boldsymbol{A})$ of the parameter space with the nonnegative covariance matrix restriction imposed in (7.26). $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ should provide the widest and most uninformative regions of responses over the one million rotations. We then overlay other restriction schemes one at a time so that each must always satisfy the $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ covariance condition *and* the added size/event/external restrictions and, hopefully, further constrain the set of plausible responses.

Importantly, we overlay the same restriction strategies $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$, $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$, $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$, and $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ onto our $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ for the pure empirical solution set $\mathbb{A}_E$ as we placed earlier (onto $\bar{\boldsymbol{F}}$) for the purely theoretical solution set $\mathbb{A}_T$. We then bootstrap the rotated sets of shocks for each variable and generate the regions of responses that satisfy each restriction scheme. While the following figures are reminiscent of the cloud of responses reported earlier from the RE-SVAR, they are substantially different in that they are not clouds of uniquely identified lines (point estimate identification). Instead, they represent regions of responses where point identification is not feasible. Each region represents a continuum of equally likely values for a given response.

8 The Full Empirical Scheme $\bar{\boldsymbol{G}}_0(\boldsymbol{A}) - \bar{\boldsymbol{G}}_4(\boldsymbol{A})$

Figure 7 shows the regions of least-restrictive $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ responses to a contractionary shock in the federal funds rate. We append to each chart the corresponding response to a Cholesky factorization on the $x_t = [i_t, \pi_t, y_t, b_t,]'$ and display it as a red line. With the exception of the federal funds rate response to its own shock, the rest of responses look largely uninformative. This is highlighted by the fact that zero bisects the sets of responses for each of the three variables. The inflation response looks to be the most symmetrically divided above and below the zero line for the first 24 months post shock. It is only at longer horizons that a larger portion of the response falls in the negative territory. There are even odds that industrial production responds either positively or negatively to a contractionary policy shock during the first year. From the second year on, however, the response looks statistically positive. The EBP response looks uninformative as well, particularly for the first 16 months after the contractionary shock. These $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ responses starkly contrast those of the $\bar{\boldsymbol{F}}$ RE-SVAR in Figures 4 – 6. The inflation response is largely negative within 12 months from the incidence of the shock in $\bar{\boldsymbol{F}}$ and by and large uninformative in $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$. A large majority of industrial production responses are sensibly negative in $\bar{\boldsymbol{F}}$ and uninformative for $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$—and the Cholesky response shows a puzzingly positive output response in the first

four months following the contractionary shock. After 20 months post shock, $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ and $\bar{\boldsymbol{F}}$ show the most agreement for a positive EBP response. However, at short horizons, the RE-SVAR predicts a largely positive EBP response, whereas the EBP response in $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ remains ambiguous.

In a similar analysis to what we conducted earlier, we now inspect the one million shocks from the purely empirical restriction strategy $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ and look for periods when those shocks had realized large values. Figure 8 shows the distribution of maximum values for each realization of the four shocks. At first glance, none of these shocks exhibit Gaussianity in their distribution of maximum values. The distributions of the first three shocks do exhibit a relatively lower degree of skewness than those of the RE-SVAR in Figure 2. Perhaps a function of the relative lack of informative responses from $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$, maximum values are found to be much more diffused throughout the rotations, suggested by the lower percentages in $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ relative to $\bar{\boldsymbol{F}}$ for each shock. The empirical $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ restriction finds 28% of the one million rotations of the first shock are found to have a maximum value in October 1989—whereas $\bar{\boldsymbol{F}}$ finds 94% of the 976K plausible combinations of the first shock are found in September 2008. There is more agreement about the maximum value of the inflation shock. Both find the maximum value of the second shock in October 2008 (31% for the one million in $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ and 96% for the 976K in $\bar{\boldsymbol{F}}$). Coincidentally both schemes find 54% of the generated shocks to industrial production variable find a maximum value (again in October 1989 for $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ and in October 2008 for $\bar{\boldsymbol{F}}$). Finally, at 37% of the one million rotations of $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$, the maximum value of the EBP shock is found in October 2008, and at 33% of the 976K in $\bar{\boldsymbol{F}}$, the highest incidence of values for this fourth shock is found in July 2013. Overall, the highest incidences of maximum values for shocks in the federal funds rate, as well as industrial production, are found in October 1989 and the highest values for inflation and the EBP are found in October 2008.

We now overlay the first set of restrictions $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ using the exact same heuristics we employed in our RE-SVAR, but now based on the profile for the size of the shocks from Figure 8. Furthermore, we consider the same restrictions—$\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ and $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$—based on the same events from the QT and QE periods, respectively, that we appended previously onto our RE-SVAR. Finally, we impose the same external variable restrictions $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ as well. We chiefly overlay these same $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ – $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ restrictions in the same fashion (one at a time) as we previously did for the RE-SVAR.[16]

[16]For example, we do not comprehensively mount the restriction schemes onto each other. The $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ notation for the RE-SVAR, involves overlaying a size restriction onto $\bar{\boldsymbol{F}}$ and, alternatively, $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ involves

Column (a) of Figure 9 shows the response regions from overlaying the size shock constraint $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ onto the nonnegative covariance restriction $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$. This restriction seems effective in winnowing the region or responses from the one million original rotations in $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$ down to 200,101 rotations which survive the $\bar{\boldsymbol{G}}_1(\boldsymbol{A})$ constraint. However, the ensuing regions remain largely inconclusive, where an exogenous increase in the federal funds rate seems to exert a roughly even region of responses falling in the positive and the negative territory of inflation, industrial production, and the EBP. Column (b) of Figure 9 shows results from imposing the events surrounding the QT period $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$. There are 65,285 rotations (from the original one million) that satisfy these restriction schemes. The performance of the inflation response worsens. While zero still bisects the region, we now see a larger portion of the set falling in the positive territory consistent with an inflation puzzle. This stands in sharp contrast to the inflation response from the RE-SVAR with the same $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ restriction. Conversely, the $\bar{\boldsymbol{G}}_2(\boldsymbol{A})$ industrial production response marginally improves, from that of column (a), as it shows a slightly larger region falling below zero. Finally, the EBP response here looks more informative as well, now that a larger portion looks negative. However, this contradicts the RE-SVAR evidence across all specifications, along with the Cholesky prediction that EBP likely increases following a contractionary policy shock.

Figure 10 repeats the analysis for the QE restriction scheme—for $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$ displayed on the left column—and for the restriction that brings external variable information from reserves, treasury spreads, and data on federal funds futures—collected into $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ on the right column of the figure. The QE-motivated $\bar{\boldsymbol{G}}_3(\boldsymbol{A})$ restrictions winnow down the regions from the original one million to 215,543 rotations. However, the shapes of the regions remain largely inconclusive for inflation, industrial production, and the EBP. The contrast in the performance is salient for the external variable restrictions strategy. First, $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ whittles down the permissible number of rotations by more than one order of magnitude. Out of the original one million rotations, only 77,407 rotations satisfy these restrictions. More important, the shape of the inflation and industrial production responses are now more informative. Perhaps the largest improvement is for the inflation response. This restriction scheme seems to perfectly resolve any incidence of the inflation puzzle, at least for the first six months following the shock. While a portion of the response set turns positive between six and 40

appending a QT event restriction—*instead of the size restriction*—onto $\bar{\boldsymbol{F}}$. We hold this for our purely empirical strategy as well, where, say, $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ overlays the external variable restriction alone onto $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$. An alternative approach would be to cumulate the restrictions so that $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ would impose size *plus* event *plus* external variable restrictions onto $\bar{\boldsymbol{F}}$, or $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$. This would of course be a far more restrictive approach.

months post shock, this remains a relatively small region vastly dominated by the negative response.

It is worth comparing this chart to the inflation response of the RE-SVAR for the same $\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ restriction (the bottom right chart in Figure 4). The RE-SVAR shows a very small positive region of inflation responses within the first four months, whereas the purely empirical response here guarantees no positive response in that time frame. Both specifications show a relatively small region of positive responses at longer horizons. Finally, a point of sharp contrast between the two approaches is that the negative region of the purely empirical response proves to persist at longer horizons with a region that remains starkly negative even at two years post shock. The negative region shrinks four years post shock. We would need to extend the horizons of our impulse response functions to elucidate whether this negative response eventually dies down. Conversely, the negative cloud of inflation responses from the RE-SVAR returns to zero much quicker with virtually no negative inflation responses remaining past 18 months. Further inspection of the (b) column of Figure 10 reveals the industrial production response mostly lies below zero for the first 18 months post shock. This looks to be the best performing industrial production response in the entire empirical $\bar{\boldsymbol{G}}_0(\boldsymbol{A})$—$\bar{\boldsymbol{G}}_4(\boldsymbol{A})$ identification schemes. Finally, the EBP response from the external variable restriction scheme continues to be largely ambiguous.

9 Conclusion

We provide a framework for the identification of structural shocks that is grounded on a purely theoretical foundation consistent with a rational expectations mechanism. An advantage of the approach is that it enables the modeling of forward-lookingness strictly from observables within a reasonably small-scale system. Modeling forward-looking behavior has typically been accomplished primarily through larger dimensional models buttressed with information from unobservables. These theoretical methodologies have typically been the province of DSGE modeling. And there has been a relatively large literature on what conditions are requisite for representing a medium/large scale forward-looking DSGE as an SVAR, which is mostly a backward-looking modeling mechanism. Our RE-SVAR offers a compromise between the highly prescriptive large-dimensional approach of most DSGEs and the loosely restricted search of the parameter space by most SVAR modeling—which are often motivated, but not directly generated, by theory.

We find our RE-SVAR provides a preponderance of sensible responses for inflation, industrial production, and EBP to a contractionary monetary policy shock. Of course, our framework is directly and strictly constructed from a theoretical model. Therefore, we also consider a purely empirical approach, which does not allow for point identification but rather set identification of responses. We find, perhaps unsurprisingly, that an extremely weak restriction of the parameter space provides largely uninformative regions of responses. The purely empirical approach is contingent on quite general and mild conditions for the generation of a shock from the data. The purely theoretical is far more restrictive but it does allow for the construction of impulse response functions based on point estimate, rather than regions. Consequently, we consider a middle ground by overlaying added empirical restrictions—both to the purely theoretical RE-SVAR strategy and the purely empirical one. We find that our shock size restrictions help further refine conclusions of the RE-SVAR but are largely uninformative for the empirical restriction scheme. Similarly, our event restrictions surrounding QT and QE periods seem useful for our RE-SVAR but less so for the purely empirical restriction scheme. Finally, restrictions from external data on reserves, treasury spreads, and federal funds futures improves the performance of the purely empirical restriction scheme. Still, the RE-SVAR response, when combined with these external variable restrictions, dominates that of the purely empirical strategy for the totality of the variables in the system. Overall, we find our RE-SVAR restriction scheme performs rather well and yields a relatively low incidence of output and inflation puzzles—particularly when augmenting it with other empirical restrictions. These same restriction schemes perform less well on a purely empirical approach.

Both the conduct and the transmission of monetary policy have likely experienced important regime changes over time. The advent of the UMP period makes this observation even more salient. This may lead to the conclusion that a Taylor rule-based consensus model may be less applicable in the decade and a half since the GFC. Our paper shows results largely at odds with that notion. We conclude that disciplining a VAR with a direct theoretical scaffolding based on plausible parameterizations of the Taylor Rule may prove desirable. Particularly, if it is properly augmented with information from markets whose relevance increased during the UMP period, such as the reserves or various money markets.

References

Antolín-Díaz, J. and J. F. Rubio-Ramírez (2018): "Narrative Sign Restrictions for SVARs," *American Economic Review*, 108, 2802–29.

Arias, J. E., D. Caldara, and J. F. Rubio-Ramirez (2019): "The systematic component of monetary policy in SVARs: An agnostic identification procedure," *Journal of Monetary Economics*, 101, 1–13.

Auerbach, A. J. and Y. Gorodnichenko (2013): "Output spillovers from fiscal policy," *American Economic Review*, 103, 141–46.

Baker, S. R. and N. Bloom (2013): "Does uncertainty reduce growth? Using disasters as natural experiments," Tech. rep., National Bureau of Economic Research.

Bauer, M. and G. D. Rudebusch (2014): "The signaling channel for Federal Reserve bond purchases," *International Journal of Central Banking*.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005): "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach*," *The Quarterly Journal of Economics*, 120, 387–422.

Binder, M., H. M. Pesaran, et al. (1995): "Multivariate rational expectations models and macroeconomic modelling: a review and some new results," *Cambridge Working Papers in Economics*.

Binder, M. and M. H. Pesaran (1997): "Multivariate linear rational expectations models: characterization of the nature of the solutions and their fully recursive computation," *Econometric Theory*, 13, 877–888.

Blanchard, O. J. and C. Khan (1980): "The Solution of Linear Difference Models Under Rational Expectations, Econometrica 48," *1311*, 1305.

Bundick, B. and A. L. Smith (2020): "The dynamic effects of forward guidance shocks," *Review of Economics and Statistics*, 102, 946–965.

Carpenter, S., S. Demiralp, J. Ihrig, and E. Klee (2015): "Analyzing Federal Reserve asset purchases: From whom does the Fed buy?" *Journal of Banking & Finance*, 52, 230–244.

Carpenter, S. B., J. E. Ihrig, E. Klee, A. Boote, and D. Quinn (2012): "The Federal Reserve's balance sheet: a primer and projections," *Available at SSRN 2193900*.

Chen, Z. and V. J. Valcarcel (2021): "Monetary transmission in money markets: The not-so-elusive missing piece of the puzzle," *Journal of Economic Dynamics and Control*, 131, 104214.

Christensen, J. H. and J. M. Gillan (2022): "Does quantitative easing affect market liquidity?" *Journal of Banking & Finance*, 134, 106349.

Coibion, O., Y. Gorodnichenko, and J. Wieland (2012): "The optimal inflation rate in New Keynesian models: should central banks raise their inflation targets in light of the zero lower bound?" *Review of Economic Studies*, 79, 1371–1406.

D'Amico, S., W. English, D. López-Salido, and E. Nelson (2012): "The Federal Reserve's large-scale asset purchase programmes: rationale and effects," *The Economic Journal*, 122, F415–F446.

D'Amico, S. and T. B. King (2013): "Flow and stock effects of large-scale treasury purchases: Evidence on the importance of local supply," *Journal of Financial Economics*, 108, 425–448.

Fernández-Villaverde, J., J. F. Rubio-Ramírez, T. J. Sargent, and M. W. Watson (2007): "ABCs (and Ds) of understanding VARs," *American economic review*, 97, 1021–1026.

Friedman, B. M. and K. N. Kuttner (1992): "Money, income, prices, and interest rates," *The American Economic Review*, 472–492.

Gagnon, J., M. Raskin, J. Remache, B. Sack, et al. (2011): "The Financial Market Effects of the Federal Reserve's Large-Scale Asset Purchases," *International Journal of Central Banking*, 7, 3–43.

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

Gilchrist, S. and E. Zakrajšek (2012): "Credit spreads and business cycle fluctuations," *American economic review*, 102, 1692–1720.

Ihrig, J., E. Klee, C. Li, M. Wei, and J. Kachovec (2018): "Expectations about the Federal Reserve's balance sheet and the term structure of interest rates," *International Journal of Central Banking*.

Inoue, A. and L. Kilian (2020): "Joint Bayesian Inference about Impulse Responses in VAR Models," Working Papers 2022, Federal Reserve Bank of Dallas.

Keating, J. W. (1990): "Identifying VAR models under rational expectations," *Journal of Monetary Economics*, 25, 453–476.

Klein, P. (2000): "Using the generalized Schur form to solve a multivariate linear rational expectations model," *Journal of economic dynamics and control*, 24, 1405–1423.

Krishnamurthy, A. and A. Vissing-Jorgensen (2011): "The Effects of Quantitative Easing on Interest Rates: Channels and Implications for Policy," *Brookings Papers on Economic Activity*, 43, 215–287.

——— (2013): "The ins and outs of LSAPs," in *Proceedings-Economic Policy Symposium-Jackson Hole*, Federal Reserve Bank of Kansas City, 57–111.

Kuttner, K. N. (2001): "Monetary policy surprises and interest rates: Evidence from the Fed funds futures market," *Journal of monetary economics*, 47, 523–544.

——— (2018): "Outside the Box: Unconventional Monetary Policy in the Great Recession and Beyond," *Journal of Economic Perspectives*, 32, 121–46.

Lucas, R. E. (1972): "Expectations and the Neutrality of Money," *Journal of economic theory*, 4, 103–124.

Ludvigson, S. C., S. Ma, and S. Ng (2021): "Uncertainty and business cycles: exogenous impulse or endogenous response?" *American Economic Journal: Macroeconomics*, 13, 369–410.

Martínez-García, E. (2020): "A Matter of Perspective: Mapping Linear Rational Expectations Models into Finite-Order VAR Form," *Globalization Institute Working Paper*.

Mertens, K. and M. O. Ravn (2014): "A reconciliation of SVAR and narrative estimates of tax multipliers," *Journal of Monetary Economics*, 68, S1–S19.

Morris, S. D. (2016): "VARMA representation of DSGE models," *Economics Letters*, 138, 30–33.

——— (2017): "DSGE pileups," *Journal of Economic Dynamics and Control*, 74, 56–86.

Mumtaz, H. and P. Surico (2009): "The Transmission of International Shocks: A Factor-Augmented VAR Approach," *Journal of Money, Credit and Banking*, 41, 71–100.

Ravenna, F. (2007): "Vector autoregressions and reduced form representations of DSGE models," *Journal of monetary economics*, 54, 2048–2064.

Sengupta, R., A. L. Smith, et al. (2022): "Assessing Market Conditions ahead of Quantitative Tightening," *Economic Bulletin*, 1–4.

Sims, C. A. (2002): "Solving linear rational expectations models," *Computational economics*, 20, 1.

Smith, A. L. and V. J. Valcarcel (2020): *The Financial Market Effects of Unwinding the Federal Reserve's Balance Sheet*, Federal Research Bank of Kansas City Working Paper.

Strongin, S. (1995): "The identification of monetary policy disturbances explaining the liquidity puzzle," *Journal of Monetary Economics*, 35, 463–497.

Swanson, E. T. (2018): "The Federal Reserve Is Not Very Constrained by the Lower Bound on Nominal Interest Rates." *Brookings Papers on Economic Activity*, 555–573.

——— (2020): "Measuring the effects of Federal Reserve forward guidance and asset purchases on financial markets," *Journal of Monetary Economics*, Forthcoming.

Uhlig, H. (2005): "What are the effects of monetary policy on output? Results from an agnostic identification procedure," *Journal of Monetary Economics*, 52, 381–419.

Vayanos, D. and J.-L. Vila (2021): "A Preferred-Habitat Model of the Term Structure of Interest Rates," *Econometrica*, 89, 77–112.

Woodford, M. (2012): "Methods of policy accommodation at the interest-rate lower bound," in *Proceedings-Economic Policy Symposium-Jackson Hole*, Federal Reserve Bank of Kansas City, 185–288.

Wright, J. H. (2012): "What does monetary policy do to long-term interest rates at the zero lower bound?" *The Economic Journal*, 122, F447–F466.

Wu, J. C. and F. D. Xia (2016): "Measuring the macroeconomic impact of monetary policy at the zero lower bound," *Journal of Money, Credit and Banking*, 48, 253–291.

Table 1: Quantitative Tightening Announcements

Date	Announcement	Description
May 22, 2013[a]	Taper	Bernanke says tapering could begin "in the next few meetings"
Jun 19, 2013[b]	Taper	Bernanke states that tapering could be appropriate "later this year"
May 21, 2014[c]	Unwind	Minutes signal beginning of balance sheet normalization planning
Jul 9, 2014[c]	Unwind	Minutes discuss gradual approach to ceasing asset reinvestments
Aug 20, 2014[c]	Unwind	Minutes offer details on balance sheet normalization planning
Sep 17, 2014[c]	Unwind	FOMC releases *Policy Normalization Principles and Plan*
Jan 12, 2017[d]	Unwind	Three Fed speeches discuss normalizing the balance sheet
Apr 5, 2017[c]	Unwind	Minutes signal phasing out reinvestments "later this year"
May 24, 2017[c]	Unwind	Minutes detail plan for phasing out reinvestment
Jun 14, 2017[b]	Unwind	FOMC releases asset runoff plan, announces that runoff will begin "this year"
Sep. 20, 2017[b]	Unwind	FOMC announces that asset runoff will begin next month

[a] Source: The Economic Outlook Congressional Hearings, 113th Congress, Joint Economic Committee.

[b] Source: FOMC Meeting Meeting calendars, statements, and minutes (2016-2021).

[c] Source: Federal Reserve History of the FOMC's Policy Normalization Discussions and Communications.

[d] Source: Ben Bernanke's Brookings Blog Shrinking the Fed's balance sheet.

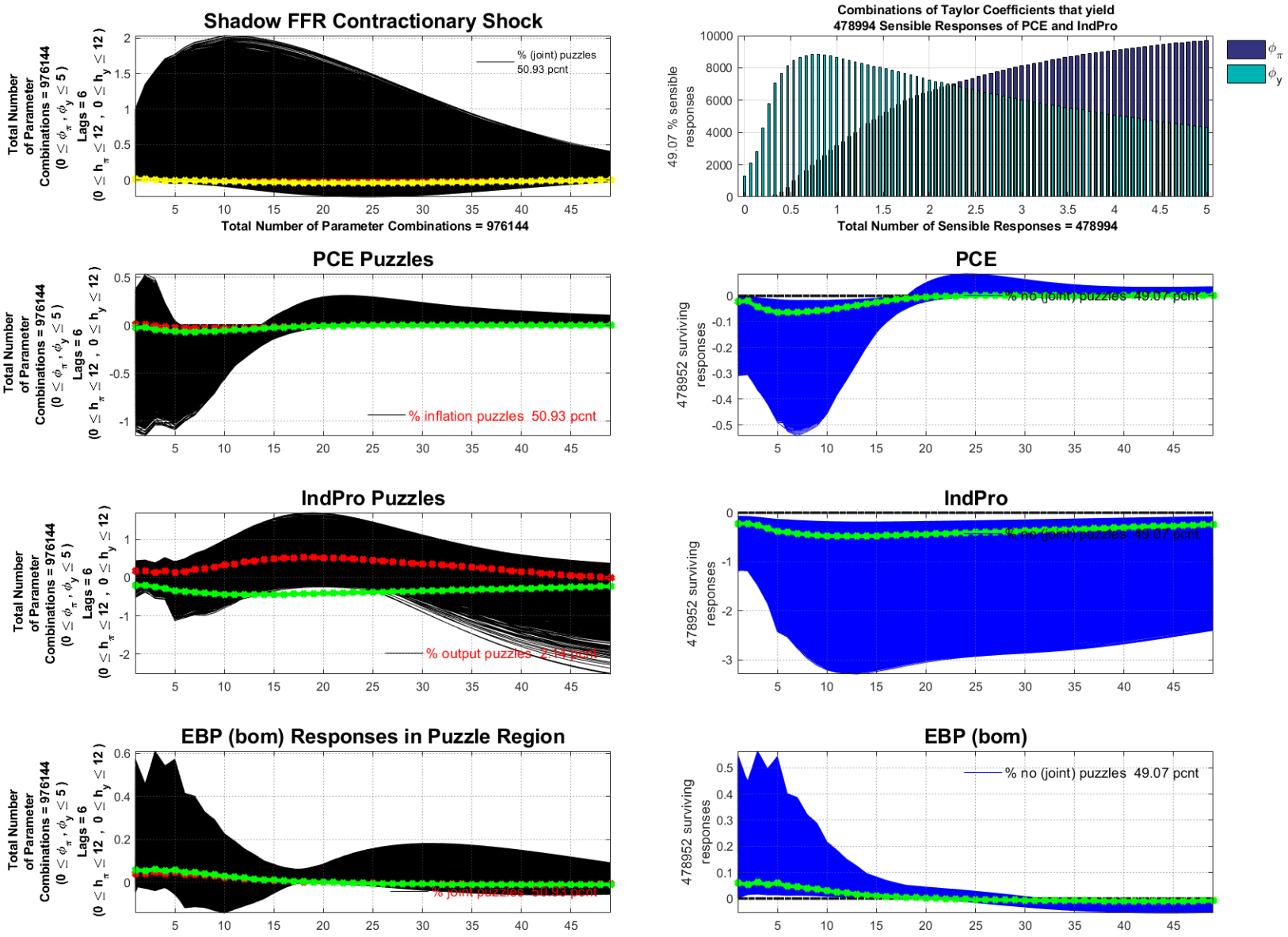
Table 2: Quantitative Easing Announcements

Date	Announcement	Event	Source
Nov 25, 2008[a]	QE 1	FOMC Meeting	Krishnamurthy and Vissing-Jorgensen (2011)
Dec 1, 2008[a]	QE 1	Speech	Krishnamurthy and Vissing-Jorgensen (2011)
Dec 16, 2008[a]	QE 1	FOMC Meeting	Krishnamurthy and Vissing-Jorgensen (2011)
Jan 28, 2009[a]	QE 1	FOMC Meeting	Krishnamurthy and Vissing-Jorgensen (2011)
Mar 18, 2009[a]	QE 1	FOMC Meeting	Krishnamurthy and Vissing-Jorgensen (2011)
Aug 10, 2010[a]	QE 2	FOMC Meeting	Krishnamurthy and Vissing-Jorgensen (2011)
Sep 21, 2010[a]	QE 2	FOMC Meeting	Krishnamurthy and Vissing-Jorgensen (2011)
Sep 21, 2011[b]	MEP	FOMC Meeting	Krishnamurthy and Vissing-Jorgensen (2013)
Sep 13, 2012	QE 3	FOMC Meeting	Krishnamurthy and Vissing-Jorgensen (2013)

[a] See also Woodford (2012) for a description of these events.

[b] Note: MEP denotes Maturity Extension Program.

Figure 1: 900K+ Specifications of a Rational Expectations Structural VAR



Note: The top chart in the left column shows the cloud of 976,144 federal funds rate responses to their own shock (with the median response in yellow). The rest of that column shows black clouds of puzzling responses of inflation, industrial production, and EBP respectively. The red dotted lines show the median responses of these clouds of puzzles. The right column shows blue clouds corresponding to the sets of sensible responses for each variable. The green dotted lines are the median values over the sensible set for each horizon.

Figure 2: Distribution of Structural Shock Realizations from the Rational Expectations SVAR

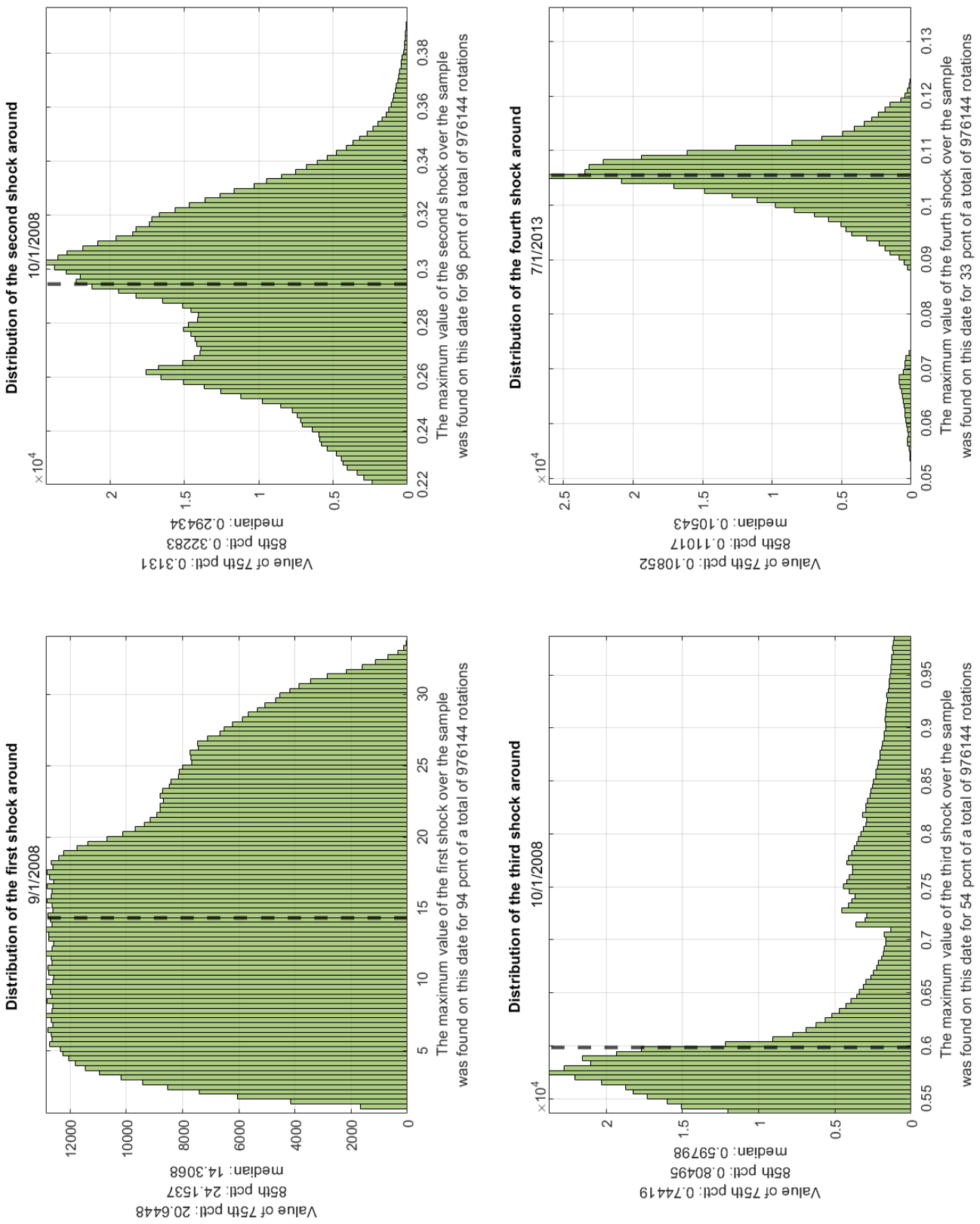
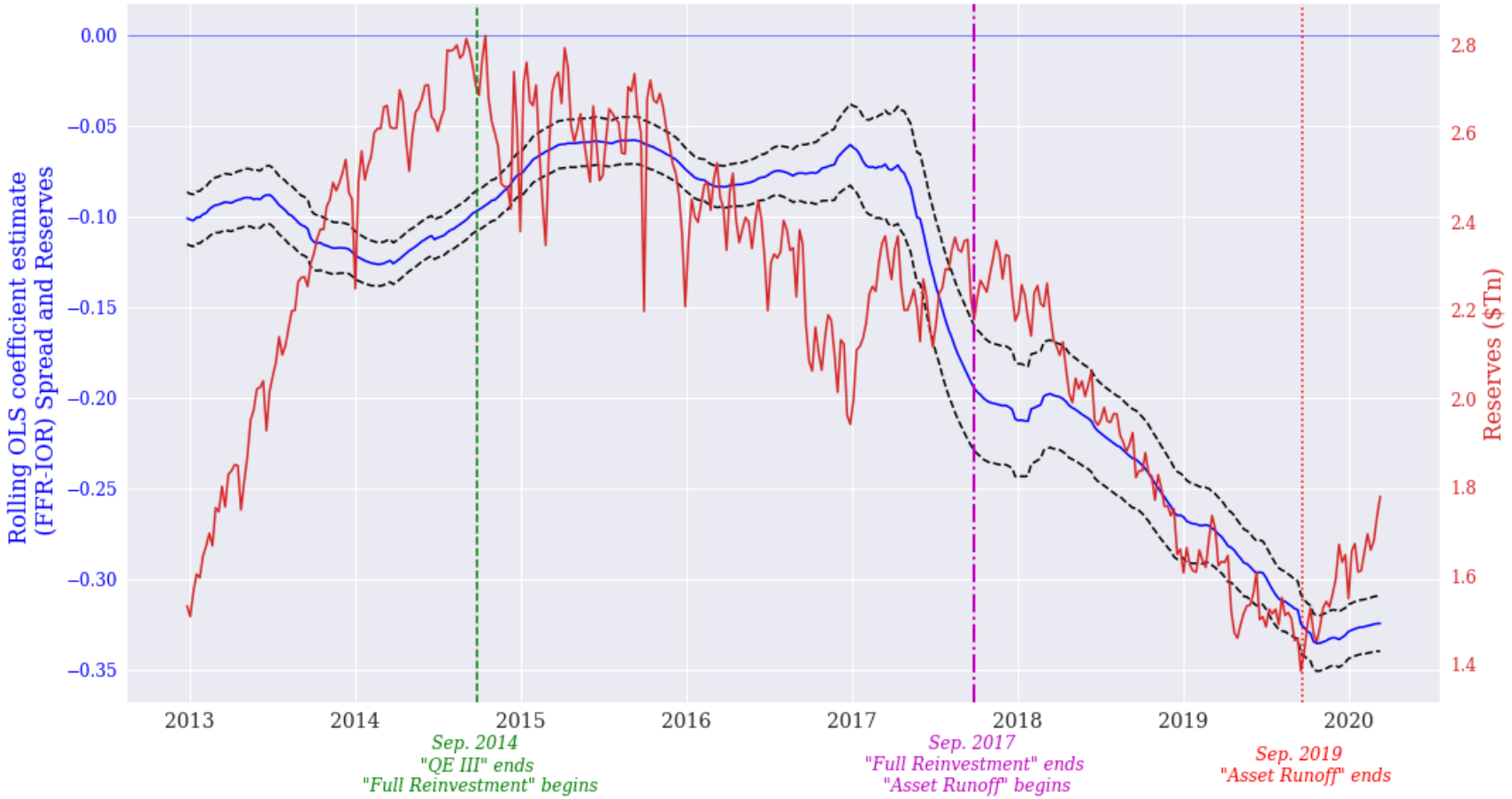


Figure 3: Total Bank Reserves and the Short-Term Interest Rate



Source: Smith and Valcarcel (2020). The (solid) blue line denotes the rolling regression estimate of the liquidity effect obtained by regressing the spread between the federal funds rate and the interest rate paid on reserves on a constant and the natural log of reserve balances. This estimate is flanked by a 90% confidence interval. For this regression estimate, the date on the x-axis denotes the end point of a 208-week rolling window. The first vertical (dashed) line corresponds with the end of the *QE III* period and the beginning of the *Full Reinvestment* phase of the balance sheet unwind period (2014-Q3). The (dashed-dotted) vertical line in the middle of the chart acts as a line of demarcation between the *Full Reinvestment* and *Asset Runoff* phases within the normalization period. The rightmost (dotted) vertical line denotes the end of the balance sheet normalization period.

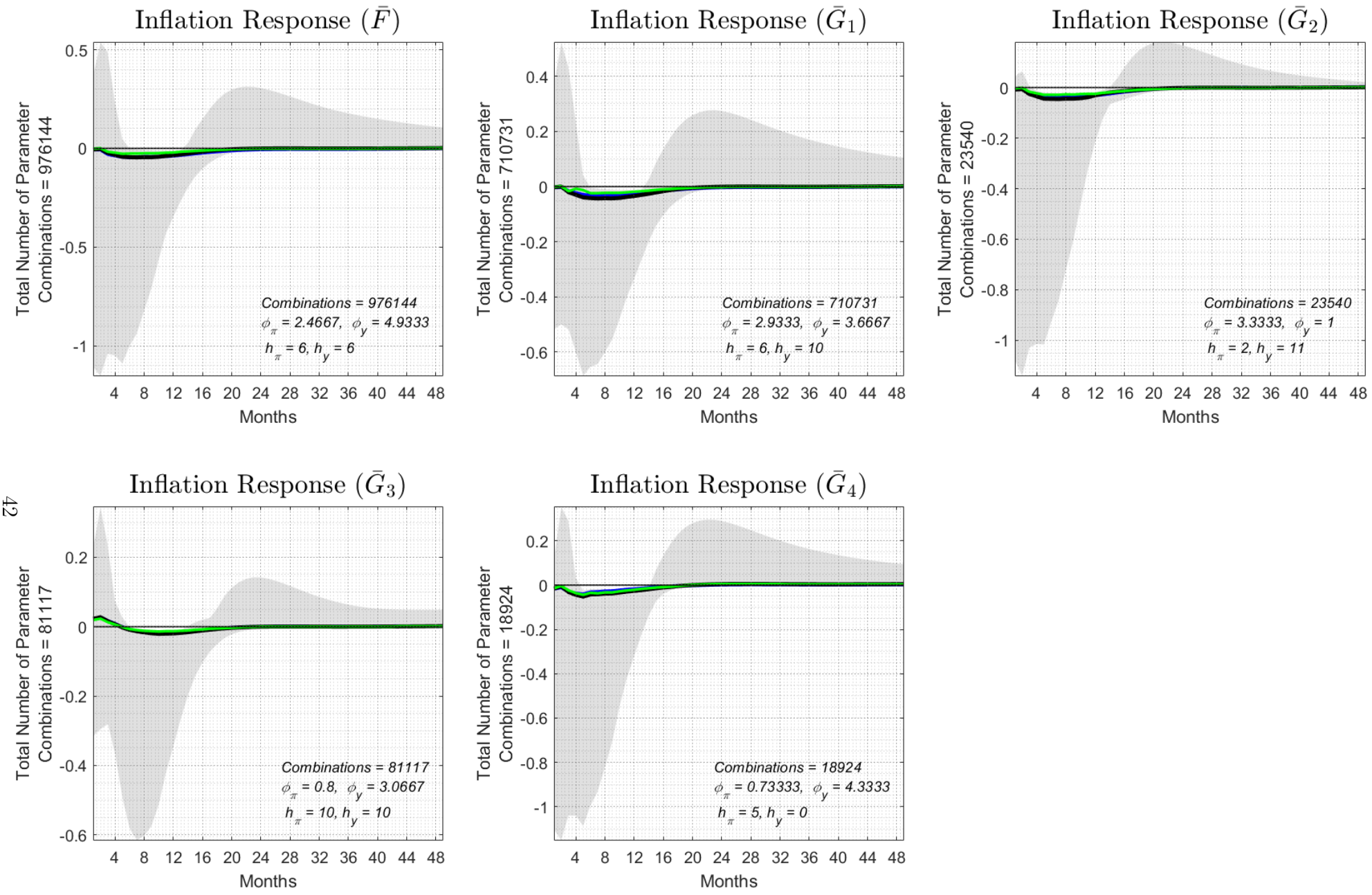


Figure 4: The solid black and blue lines represent the median and mean response across each response set. The green lines connect the modal values across all responses for each horizon.

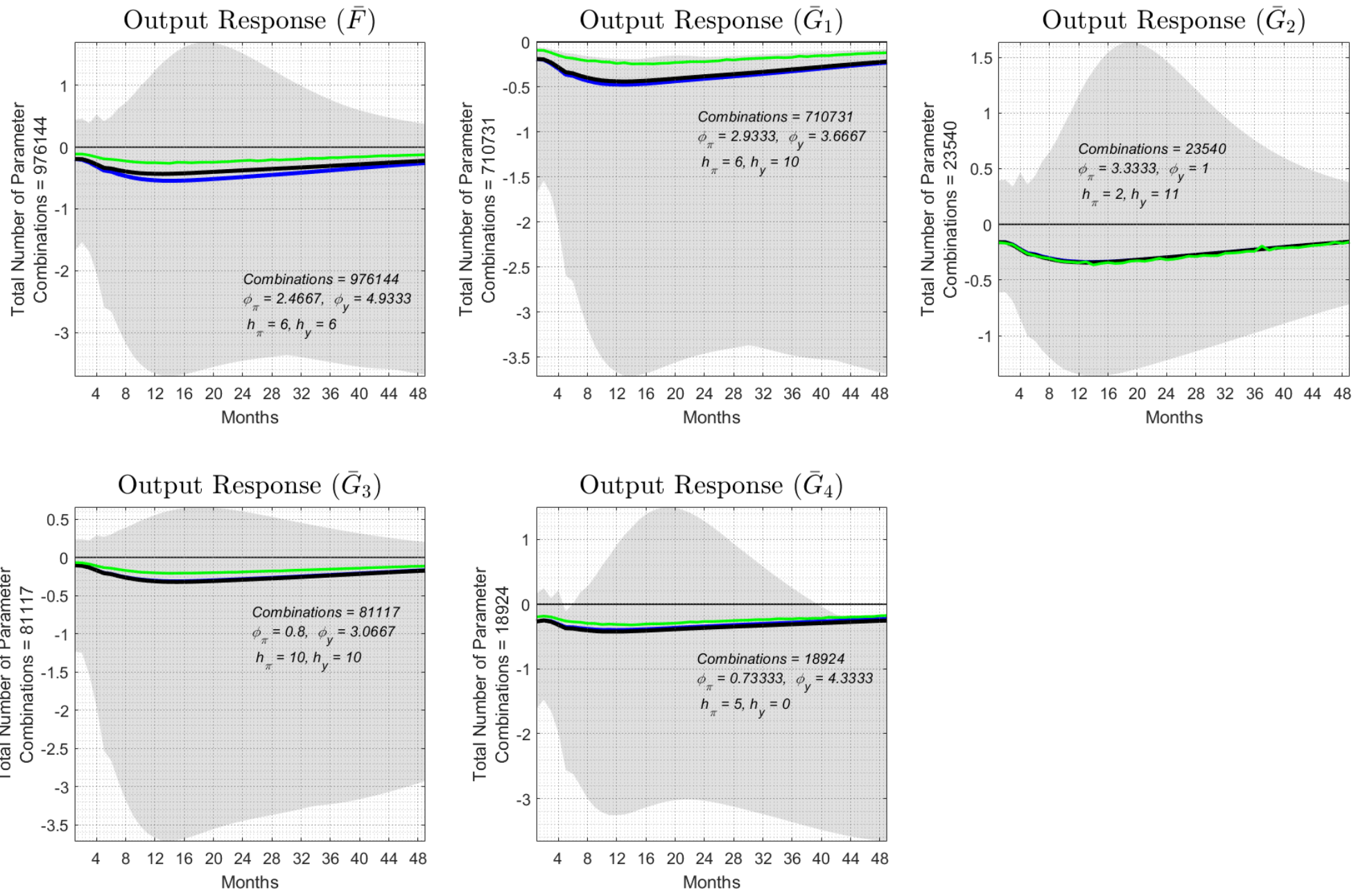


Figure 5: The solid black and blue lines represent the median and mean response across each response set. The green lines connect the modal values across all responses for each horizon.

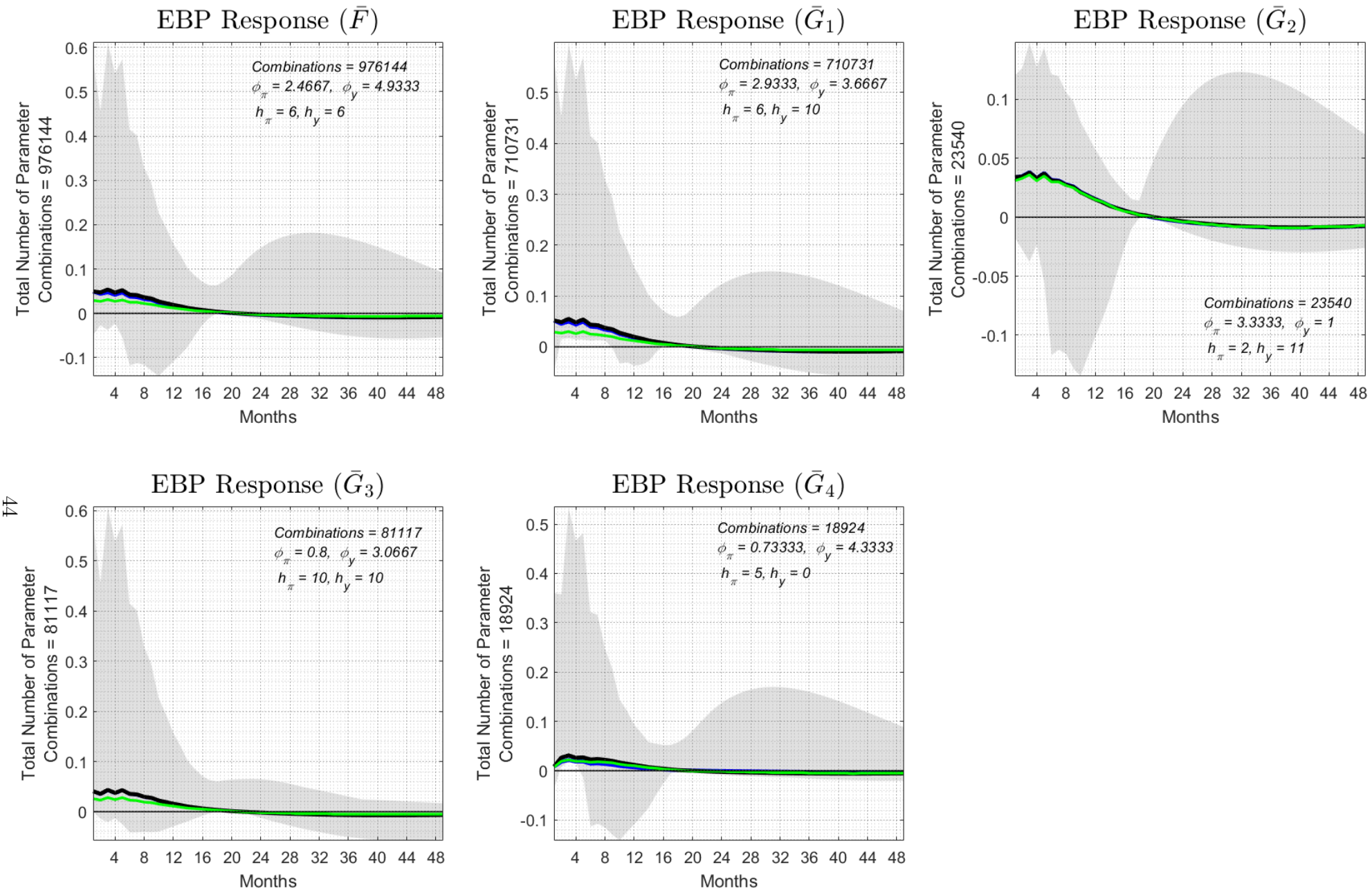


Figure 6: The solid black and blue lines represent the median and mean response across each response set. The green lines connect the modal values across all responses for each horizon.

Figure 7: Responses from Reduced-Form Mapping to Unconstrained Covariance $\bar{G}_0(A)$ Restriction Scheme (Based on 1 million QR rotations. Red lines denote Cholesky Responses)

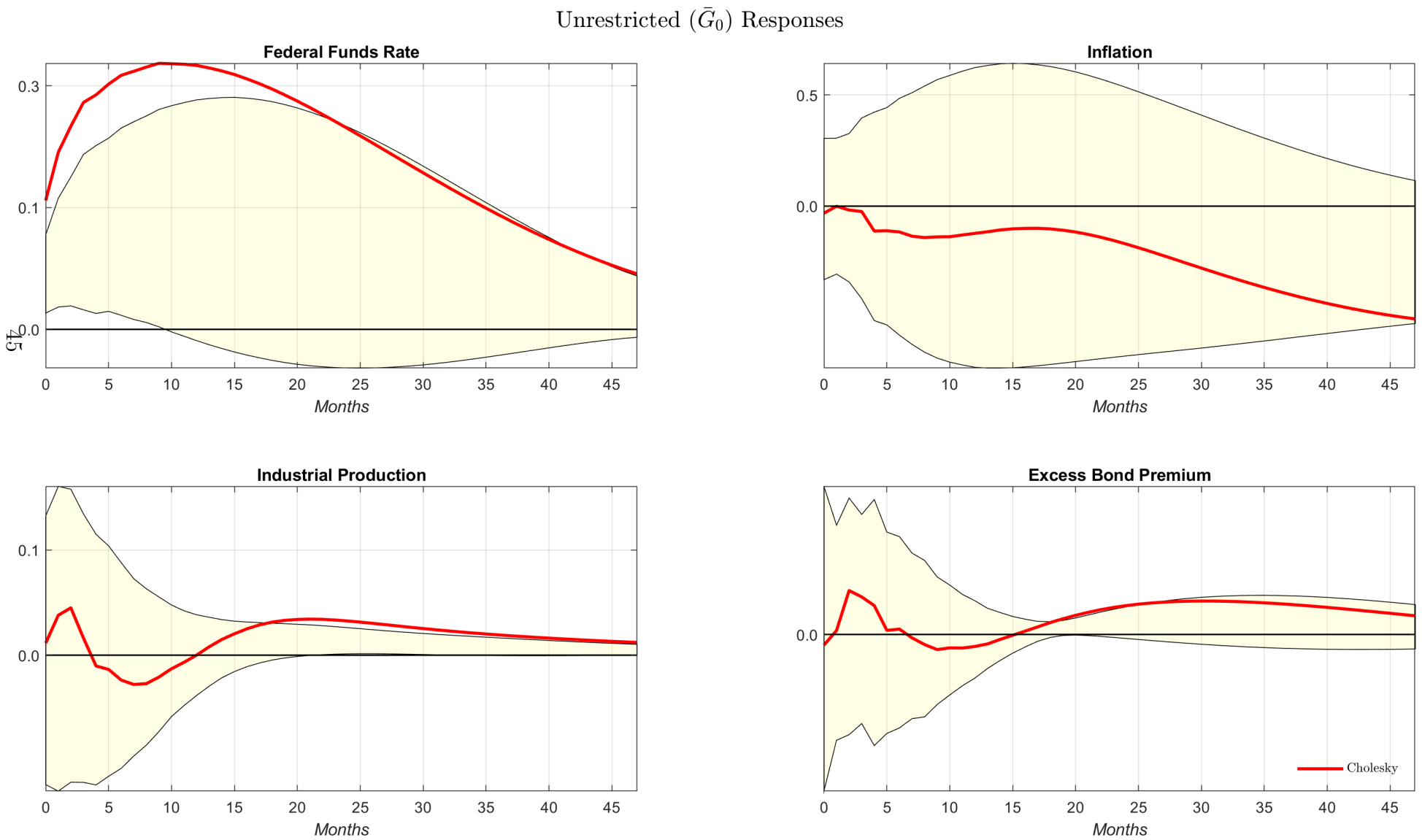


Figure 8: Distribution of Rotated Innovations from the Covariance Restriction $\bar{G}_0(A)$

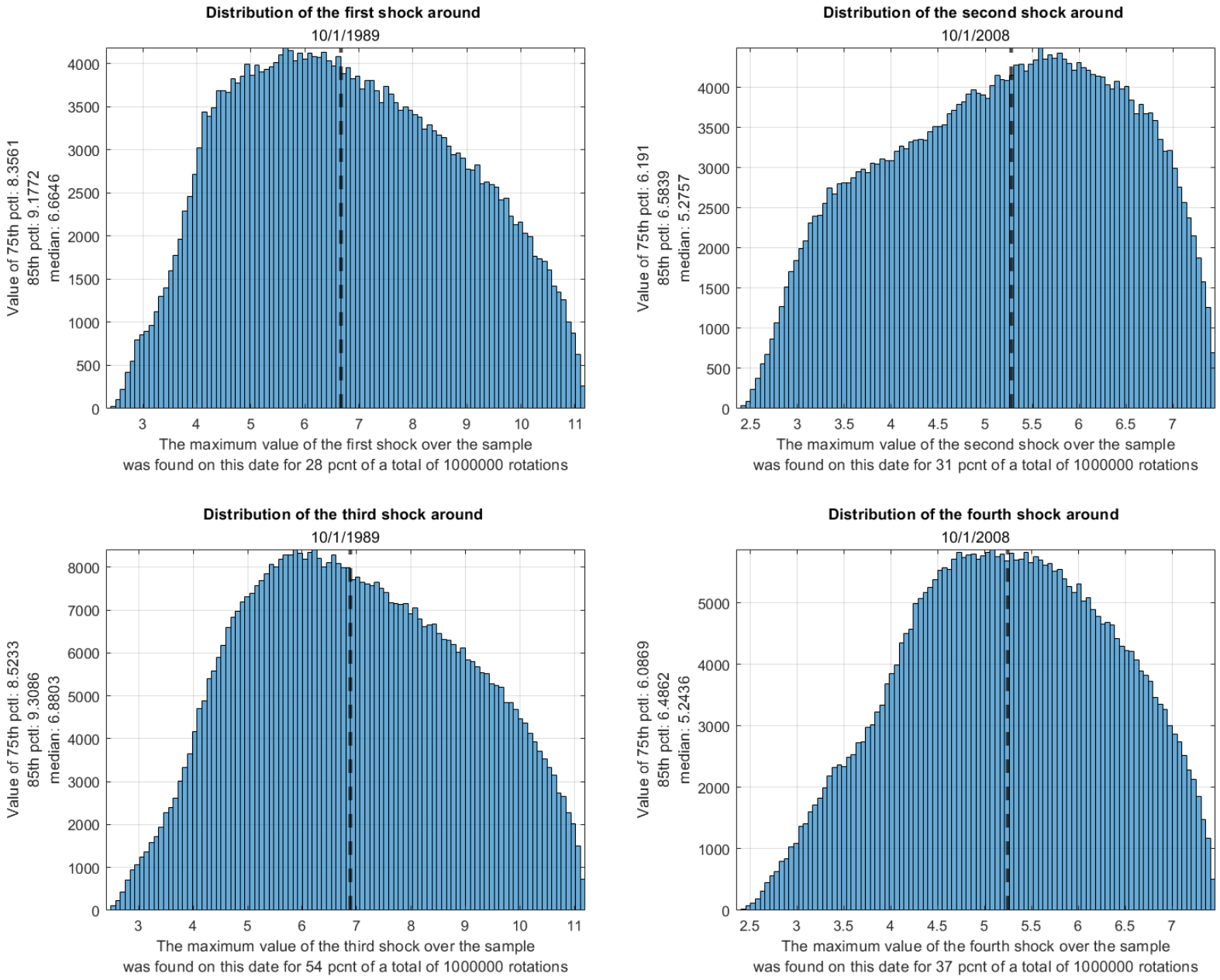
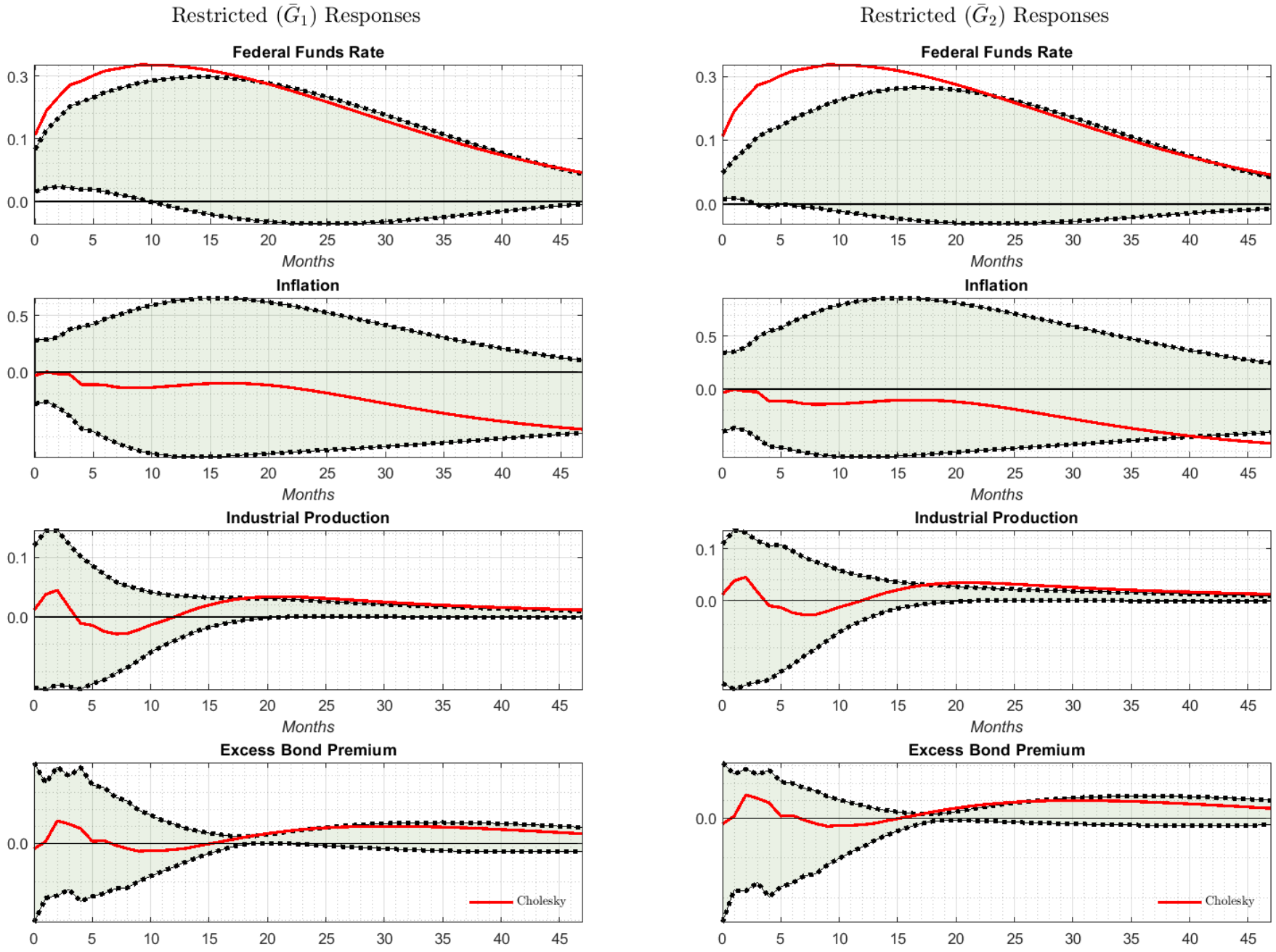


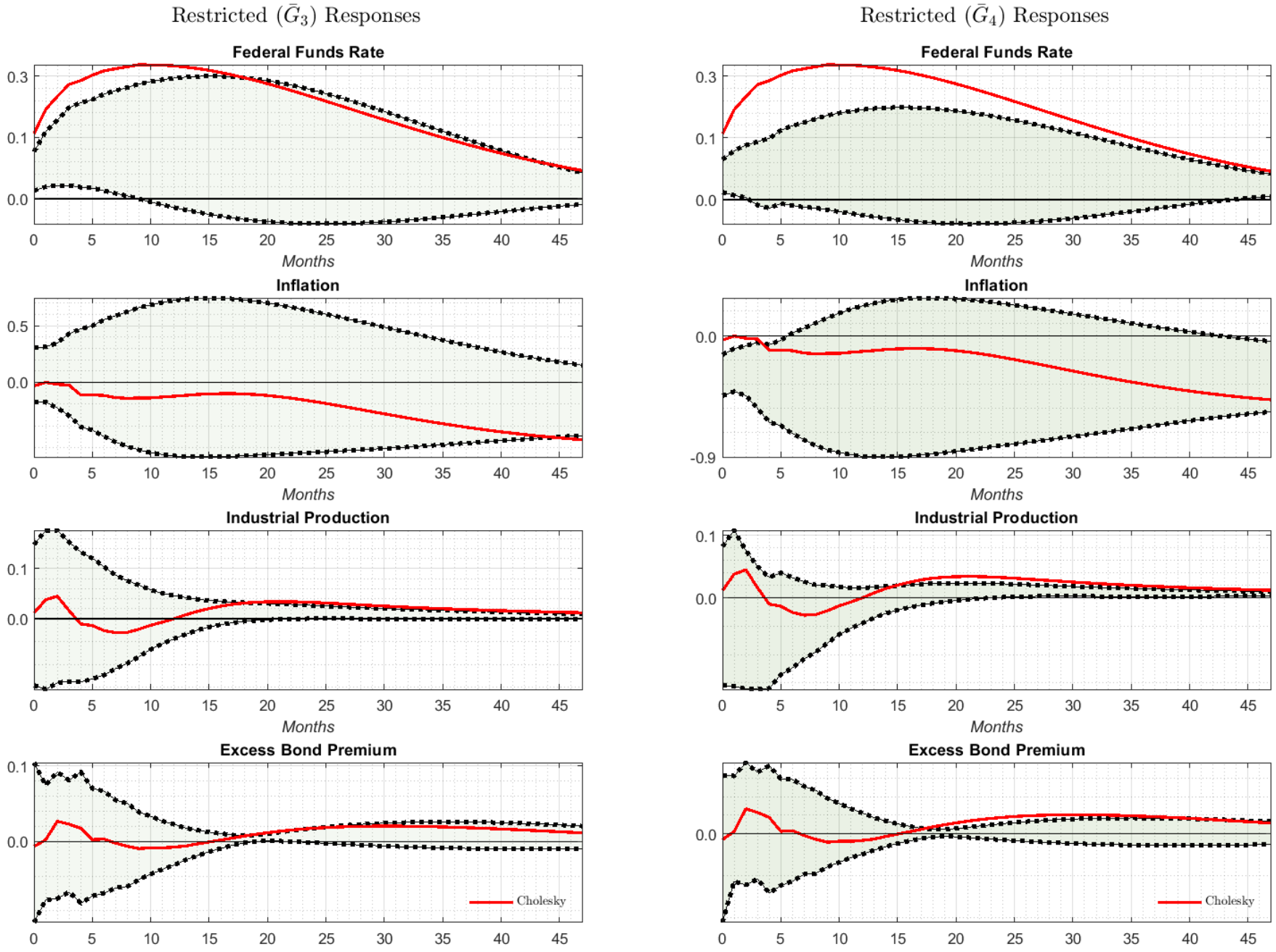
Figure 9: Bootstrapped Responses from Overlaying (*Size/Event*) Shock Restrictions onto $\bar{G}_0(A)$



(a) column: $\bar{\boldsymbol{G}}_1(\boldsymbol{A}) = 200{,}101$ rotations.

(b) column: $\bar{\boldsymbol{G}}_2(\boldsymbol{A}) = 65{,}284$ rotations

Figure 10: Bootstrapped Responses from Overlaying (*Event*/*External*) Shock Restrictions onto $\bar{G}_0(A)$



(a) column: $\bar{\boldsymbol{G}}_3(\boldsymbol{A}) = 215{,}543$ rotations.

(b) column: $\bar{\boldsymbol{G}}_4(\boldsymbol{A}) = 77{,}407$ rotations