Impulse Response Analysis of Structural Nonlinear Time Series Models

Giovanni Ballarin*
University of St. Gallen

December 1, 2024

Abstract: This paper proposes a semiparametric sieve approach to estimate impulse response functions of nonlinear time series within a general class of structural autoregressive models. We prove that a two-step procedure can flexibly accommodate nonlinear specifications while avoiding the need to choose of fixed parametric forms. Sieve impulse responses are proven to be consistent by deriving uniform estimation guarantees, and an iterative algorithm makes it straightforward to compute them in practice. With simulations, we show that the proposed semiparametric approach proves effective against misspecification while suffering only minor efficiency losses. In a US monetary policy application, we find that the pointwise sieve GDP response associated with an interest rate increase is larger than that of a linear model. Finally, in an analysis of interest rate uncertainty shocks, sieve responses imply more substantial contractionary effects both on production and inflation.

Keywords: nonlinear time series, impulse responses, sieve estimation, physical dependence

*E-mail: giovanni.ballarin@unisg.ch. I thank Otilia Boldea, Timo Dimitriadis, Juan Carlos Escanciano, Lyudmila Grigoryeva, Klodiana Istrefi, Marina Khismatullina, So Jin Lee, Yuiching Li, Sarah Mouabbi, Andrey Ramirez, Christoph Rothe, Carsten Trenkler and Mengshan Xu, as well as the participants of the Econonometrics Seminar at the University of Mannheim, the 2023 ENTER Jamboree, the 10th HKMEtrics Workshop, the GSS Weekly Seminar at Tilburg University, the Internal Econometrics Seminar at Vrije Universiteit Amsterdam and the Brown Bag Seminar at the University of St. Gallen for their comments, suggestions and feedback. A significant part of this work was developed at the University of Mannheim thanks to the support of the Center for Doctoral Studies in Economics.

1 Introduction

Linearity is a foundational assumption in structural time series modeling. For example, large classes of macroeconomic models in modern New Keynesian theory can be reduced to linear forms via linearization techniques. This often justifies the use of the linear time series toolbox from a theoretical point of view. The seminal work of Sims (1980) on vector autoregressive (VAR) models brought the study of dynamic economic relationships into focus within the macro-econometric literature, for which the estimation and analysis of impulse response functions (IRFs) is key (Hamilton, 1994b, Lütkepohl, 2005, Kilian and Lütkepohl, 2017). The local projection (LP) approach of Jordà (2005) has also gained popularity as an alternative, thanks to its flexibility and ease of implementation.

Linear models, however, are limited in the kind of effects that they can describe. Asymmetries in monetary policy and non-proportional shock effects are now commonly studied. However, most works construct essentially parametric nonlinear specifications. For example, Tenreyro and Thwaites (2016) studying both sign and size effects of monetary policy (MP) shocks using censoring and cubic transformations, respectively. Caggiano et al. (2017), Pellegrino (2021) and Caggiano et al. (2021) use multiplicative interacted VAR models to estimate effects of uncertainty and MP shocks. From a macro-finance perspective, Forni et al. (2023a,b) study the economic effects of financial shocks following the quadratic VMA specification (Debortoli et al., 2020). Gambetti et al. (2022) study news shocks asymmetries by imposing that news changes enter an autoregressive model through a threshold map. Parametric nonlinear specifications are also common prescriptions in time-varying models (Auerbach and Gorodnichenko, 2012, Caggiano et al., 2015) and state-dependent models (Ramey and Zubairy, 2018).

In this paper, we aim to design a semiparametric, structural nonlinear time series modeling and estimation framework with explicit theoretical properties. Our structural framework is an extension of the block-recursive form from Gonçalves et al. (2021); we combine it with the uniform sieve estimation theory of Chen and Christensen (2015) within a general physical dependence setup (Wu, 2005). Under appropriate regularity

assumptions, we show that a two-step semiparametric series estimation procedure is able to consistently recover the structural model in a uniform sense. Since we ultimately wish to study impulse responses, our theory also encompasses guarantees for estimated nonlinear IRFs: Nonlinear impulse response function estimates are asymptotically consistent and, thanks to an iterative algorithm, straightforward to compute in practice.

To illustrate the validity of our proposed methodology, we first provide simulation evidence. With realistic sample sizes, the efficiency costs of the semiparametric procedure are small compared to correctly-specified parametric estimates. A second set of simulations demonstrates that whenever the nonlinear parametric model is mildly misspecified the large-sample bias is large, while for semiparametric estimates it is negligible. We then evaluate how the IRFs computed with the new method compare with the ones from two previous empirical exercises. In a small, quarterly model of the US macroeconomy, we find that the parametric nonlinear and nonlinear appear to underestimate by intensity the GDP responses by 13% and 16%, respectively, after a large exogenous monetary policy shock. Moreover, sieve responses achieve maximum impact a year before their linear counterparts. Then, we evaluate the effects of interest rate uncertainty on US output, prices, and unemployment following Istrefi and Mouabbi (2018). In this exercise, the impact on industrial production of a one-deviation increase in uncertainty is approximately 54% stronger according to semiparametric IRFs than the comparable linear specification. These findings suggest that structural responses based on linear specifications might be appreciably underestimating shock effects.

A few similar efforts to the one we undertake in this paper have been made thus far. Gourieroux and Lee (2023) provide a framework for nonparametric kernel estimation and inference of IRFs via local projections, although they primarily work in the one-dimensional, single lag case. The seminal work of Jordà (2005) suggested the so-called "flexible local projection" approach based on the Volterra expansion. There are multiple issues with this method: First, the Volterra expension is not formally justified, nor is its truncation, which is key in studying its properties (Sirotko-Sibirskaya et al., 2020, Mova-

hedifar and Dickhaus, 2023). Second, the flexible LP proposal is effectively equivalent to adding mixed polynomial terms to a linear regression, meaning it is a semiparametric method and must be analyzed as such. As we deal with nonlinear impulse responses, we briefly mention here the Generalized IRF (GIRF) approach originated by Koop et al. (1996), Potter (2000) and Gourieroux and Jasiak (2005), of which Teräsvirta et al. (2010) provide a textbook treatment. GIRFs are defined with more sophisticated conditioning sets than standard IRFs. Yet, a core issue with GIRFs is that they do not explicitly address the problem of structural identification (Kilian and Lütkepohl, 2017). In this line of work, Kanazawa (2020) proposed to use radial basis function neural networks to estimate nonlinear reduced-form GIRFs for the US economy.

The remainder of this paper is organized as follows. Section 2 provides the general framework for the structural model. Section 3 describes the two-step semiparametric estimation strategy and Section 4 discusses nonlinear impulse response function computation, validity and consistency. In Section 5 we give a brief overview of simulation results, while Section 6 contains the empirical analyses. Finally, Section 7 concludes. All proofs and additional material can be found in the Supplementary Material. With regard to notation: scalar and vector random variables are denoted in capital or Greek letters, e.g. Y_t or ϵ_t, while realization are shown in lowercase Latin letters, e.g. y_t. For a process $\{Y_t\}_{t\in\mathbb{Z}}$, we write $Y_{t:s} = (Y_t, Y_{t+1}, \ldots, Y_{s-1}, Y_s)$, as well as $Y_{*:t} = (\ldots, Y_{t-2}, Y_{t-1}, Y_t)$ for the left-infinite history and $Y_{t:*} = (Y_t, Y_{t+1}, Y_{t+2}, \ldots)$ for its right-infinite history. The same notation is also used for random variable realizations. For a matrix $A \in \mathbb{R}^{d\times d}$ where $d \geqslant 1$, $\|A\|$ is the spectral norm, $\|A\|_\infty$ is the supremum norm and $\|A\|_r$ for $0 < r < \infty$ is the r-operator norm. For a random vector or matrix, we will use $\|\cdot\|_{L^r}$ to denote the associated L^r norm.

2 Model Framework

In this section, we introduce the general nonlinear time series model. In terms of structural shocks identification, the idea is straightforward: One must choose a scalar series, X_t, to be the *structural variable* identifying shocks, and explicitly model the dynamic effects on the remaining data, vector Y_t. This will enable the derivation of economically meaningful (structural) impulse responses due to an exogenous shock impacting X_t.

2.1 General Model

This paper focuses on the family of nonlinear autoregressive models of the form

$$\begin{aligned} X_t &= \mu_1 + A_{12}(L)Y_{t-1} + A_{11}(L)X_{t-1} + u_{1t}, \\ Y_t &= \mu_2 + G_2(Y_{t-1}, \ldots, Y_{t-p}, X_t, X_{t-1}, \ldots, X_{t-p}) + u_{2t}. \end{aligned} \tag{1}$$

where $X_t \in \mathcal{X} \subseteq \mathbb{R}$ and $Y_t \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$ are scalar and d_Y-dimensional time series, respectively, $u_t = (u_{1t}, u_{2t}')' \in \mathcal{U} \subseteq \mathbb{R}^d$ are innovations, $d = 1 + d_Y$, $G_2 : \mathbb{R}^{1+pd} \mapsto \mathbb{R}$ is a generic nonlinear map, and $A_{12}(L)$ and $A_{11}(L)$ are lag polynomials (Lütkepohl, 2005). We let $Z_t := (X_t, Y_t')' \in \mathbb{R}^d$ be the full data vector. Let us provide some examples for the model classes nested by (1).

Example 2.1 (Linear VAR)**.** In the simplest case, $G_2(Y_{t-1}, \ldots, Y_{t-p}, X_t, X_{t-1}, \ldots, X_{t-p}) = A_{22}(L)Y_{t-1} + A_{21}(L)X_t$, and we recover the class of linear vector autoregressive models.

Example 2.2 (Additively separable model)**.** When $G_2(Y_{t-1}, \ldots, Y_{t-p}, X_t, X_{t-1}, \ldots, X_{t-p}) = \sum_{i=1}^{p} G_{i,22}(Y_{t-i}) + \sum_{j=0}^{p} G_{j,21}(X_{t-j})$, model (1) is additively separable (Fan and Yao, 2003).

Example 2.3 (Nonlinear impact model)**.** An even more parsimonious class than the additively separable one is the one studied in Gonçalves et al. (2021), which may be informally termed the "nonlinear impact model class", where

$$Y_t = \mu_2 + A_{22}(L)Y_{t-1} + \sum_{j=0}^{p} G_{j,21}(X_{t-j}) + u_{2t}.$$

A useful equivalent representation of the above equation for Y_t is

$$Y_t = \mu_2 + A_{22}(L)Y_{t-1} + A_{21}(L)X_{t-1} + \sum_{j=0}^{p} \widehat{G}_{j,21}(X_{t-j}) + u_{2t},$$

where now to identify nonlinear functions $\widehat{G}_{j,21} : \mathbb{R} \mapsto \mathbb{R}^{d_Y}$, $0 \leqslant j \leqslant p$, we require that constant and linear factors be not included at indices $j \geqslant 1$. To make this more compact, write

$$Z_t = \mu + A(L)Z_{t-1} + \widehat{G}(L)X_t + u_t, \quad \text{where} \quad \widehat{G}(L) := \begin{bmatrix} 0 \\ \widehat{G}_{0,21} + \widehat{G}_{1,21}L + \ldots + \widehat{G}_{p,21}L^p \end{bmatrix},$$

with the minor abuse of notation that $\widehat{G}_2(L) := \widehat{G}_{0,21} + \ldots + \widehat{G}_{p,21}L^p$ is now intended as a *functional* lag polynomial, meaning $\widehat{G}_2(L)X_t \equiv \sum_{j=0}^{p} \widehat{G}_{j,21}(X_{t-j})$.[1]

2.2 Structural Framework

Model (1) involves only reduced-form innovations u_{1t} and u_{2t}, meaning additional assumptions are necessary in order to provide any structural interpretation. Many such assumptions have been devised in the macroeconomic literature, but few can be directly applied to nonlinear models (Kilian and Lütkepohl, 2017). Here, we follow the block-recursive identification strategy outlined in Gonçalves et al. (2021) and originally due to Kilian and Vigfusson (2011).

From (1) we derive

$$\begin{aligned} X_t &= \mu_1 + A_{12}(L)Y_{t-1} + A_{11}(L)X_{t-1} + u_{1t}, \\ Y_t &= \mu_2 + A_{22}(L)Y_{t-1} + A_{21}(L)X_{t-1} + \widehat{G}_2(Y_{t-1:t-p}, X_{t:t-p}) + u_{2t}, \end{aligned}$$

where, without loss of generality, we have assumed (as in Example 2.3) that we can separate the linear and non-linear ($\widehat{G}_2$) components from G_2. In general, it can be the case that $\mu_2 = 0$, $A_{22}(L) = 0$ or $A_{21}(L) = 0$ if e.g. G_2 is strictly nonlinear. In vector

[1] The choice to use a functional matrix notation is due to the ease of writing multivariate additive nonlinear models such as (4) in a manner consistent with standard formalisms of linear VAR models, following again e.g. Lütkepohl (2005).

form:

$$Z_t = \mu + A(L)Z_{t-1} + \widehat{G}(Z_{t:t-p}) + u_t, \quad \text{where} \quad \widehat{G}(Z_{t:t-p}) := \begin{bmatrix} 0 \\ \widehat{G}_2(Y_{t-1:t-p}, X_{t:t-p}) \end{bmatrix}. \quad (2)$$

We can now formalize the structural specification of our model.

Assumption 1. There exist (i) a vector $B_0^{21} \in \mathbb{R}^{d_Y}$ and a matrix $B_0^{22} \in \mathbb{R}^{d_Y \times d_Y}$ such that

$$\begin{bmatrix} 1 & 0 \\ B_0^{21} & B_0^{22} \end{bmatrix} =: B_0^{-1}$$

is invertible and has unit diagonal, and (ii) mutually independent innovations sequences $\{\epsilon_{1t}\}_{t\in\mathbb{Z}}$, $\epsilon_{1t} \in \mathcal{E}_1 \subseteq \mathbb{R}$, and $\{\epsilon_{2t}\}_{t\in\mathbb{Z}}$, $\epsilon_{2t} \in \mathcal{E}_2 \subseteq \mathbb{R}^{d_Y}$, such that

$$\begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \right),$$

where $\Sigma_1 > 0$ and Σ_2 is a diagonal positive definite matrix so that

$$\begin{aligned} X_t &= \mu_1 + A_{12}(L)Y_{t-1} + A_{11}(L)X_{t-1} + \epsilon_{1t}, \\ Y_t &= \mu_2 + A_{22}(L)Y_{t-1} + A_{21}(L)X_{t-1} + \widehat{G}_2(Y_{t-1:t-p}, X_{t:t-p}) + B_0^{21}\epsilon_{1t} + B_0^{22}\epsilon_{2t}, \end{aligned} \quad (3)$$

where $u_{1t} \equiv \epsilon_{1t}$, $u_{2t} := B_0^{21}\epsilon_{1t} + B_0^{22}\epsilon_{2t}$ and thus $u_t = B_0^{-1}\epsilon_t$ for $\epsilon_t = (\epsilon_1, \epsilon_2')' \in \mathcal{E} \subseteq \mathbb{R}^d$.

Remark 2.1. Assumption 1 follows Gonçalves et al. (2021) closely. By design, one does not need to identify the model fully, meaning that fewer assumptions on Z_t and ϵ_t are needed to estimate the individual structural effects of ϵ_{1t} on Y_t. This comes at the cost of not being able to simultaneously study structural effects with respect to shocks impacting ϵ_{2t}.

Note that inverting B_0^{-1} gives

$$B_0 = \begin{bmatrix} 1 & 0 \\ -B_{0,12} & B_{0,22} \end{bmatrix},$$

and multiplying both sides we find

$$B_0 Z_t = b + B(L)Z_{t-1} + \widehat{F}(Z_{t:t-p}) + \epsilon_t, \quad (4)$$

where $b = (b_1, b_2')' \in \mathbb{R}^d$ and $\widehat{F}(Z_{t:t-p}) = (0, \widehat{F}_2(Z_{t:t-p}))'$ for $\widehat{F}_2 : \mathbb{R}^{1+pd_Y} \mapsto \mathbb{R}^{d_Y}$, $F_2 =$

$B_{0,22}\widehat{G}_2$. In practice, to estimate the model's coefficients, we will leverage (3). This latter form was termed the *pseudo-reduced form* by Gonçalves et al. (2021).

We observe that $\widehat{G}_2(Y_{t-1:t-p}, X_{t:t-p})$ is correlated with u_{2t} through $B_0^{21}\epsilon_{1t}$. As X_t depends linearly on ϵ_{1t}, if $B_0^{21} \neq 0$ and $\widehat{G}_2(Y_{t-1:t-p}, X_{t:t-p})$ is not independent of X_t, there is endogeneity. Gonçalves et al. (2021) address the issue by proposing a two-step estimation procedure wherein one proxies for ϵ_{1t} by means of regression residuals $\widehat{\epsilon}_t$. As we show in Section 3 below, this approach does also allow consistent semiparametric estimation.

Remark 2.2. (Moving Average Identification). Forni et al. (2023a,b) work with an alternative nonlinear structural identification framework to the block-recursive form. Their approach follows Debortoli et al. (2020), and is based on a vector MA representation. Under appropriate assumptions, the structural model studied by Forni et al. (2023a) is

$$Z_t = \mu + A(L)Z_t + Q_0 F(\epsilon_{1t}) + B_0\epsilon_t, \tag{5}$$

where ϵ_t are independent innovations with zero mean and identity covariance, and ϵ_{1t} identifies the shocks of interest. $Q(L)$ and $B(L)$ are both linear lag polynomials, and $F(x) = x^2$ in their baseline specification. For (5) to overlap with (3), one must impose that (i) X_t is exogenous and independently distributed and (ii) only ϵ_{1t} has nonlinear effects. We emphasize that, if innovation sequence ϵ_{1t} is assumed to be observable, applying our results to the framework of Debortoli et al. (2020) is straightforward.

2.3 Structural Nonlinear Impulse Responses

Starting from pseudo-reduced equations (3), we begin by assuming that the linear autoregressive component is stable.

Assumption 2. The roots of $\det(I_d - A(L)L) = 0$ are outside the complex unit circle.

This is a rather weak assumption which will enable us to write impulse responses in

a manner that can yield simplifications for additively separable models.[2] Then, letting $\Psi(L) = (I_d - A(L)L)^{-1}$, one can write

$$Z_t = \eta + \Theta(L)\epsilon_t + \Gamma(Z_{t:*}), \tag{6}$$

where $\mu := \Psi(1)(\mu_1, \mu_2')'$, $\Theta(L) := \Psi(L)B_0^{-1}$ and $\Gamma(Z_{t:*}) := \Psi(L)(0, \widehat{G}_2(Y_{t-1:t-p}, X_{t:t-p})')'$. We emphasize that the nonlinear term $\Gamma(Z_{t:*})$ generally depends on the entire history of the process Z_t, as $\Psi(L)$ is an infinite-order MA polynomial. To formally define impulse responses, it is useful to partition the polynomial $\Theta(L)$ according to $\Theta(L) := [\Theta_{\cdot 1}(L)\, \Theta_{\cdot 2}(L)]$, where $\Theta_{\cdot 1}(L)$ represents the first column of matrices in $\Theta(L)$, and $\Theta_{\cdot 2}(L)$ the remaining d_Y columns.

Given impulse $\delta \in \mathbb{R}$ at time t, define the shocked innovation process as $\epsilon_{1s}(\delta) = \epsilon_s$ for $s \neq t$ and $\epsilon_{1t}(\delta) = \epsilon_{1t} + \delta$, as well as the shocked structural variable as $Z_s(\delta) = Z_s$ for $s < t$ and $Z_s(\delta) = X_s(\epsilon_{s:t+1}, \epsilon_t + \delta, \epsilon_{t-1:*})$ for $s \geqslant t$. Further, let

$$Z_{t+h} := \eta + \Theta_{\cdot 1}(L)\epsilon_{1t+h} + \Theta_{\cdot 2}(L)\epsilon_{2t+h} + \Gamma(Z_{t:*}),$$

$$Z_{t+h}(\delta) := \eta + \Theta_{\cdot 1}(L)\epsilon_{1t+h}(\delta) + \Theta_{\cdot 2}(L)\epsilon_{2t+h} + \Gamma(Z_{t:*}(\delta)),$$

be the time-t baseline and shocked series, respectively. Then

$$\mathrm{IRF}_h(\delta) = \mathbb{E}\left[Z_{t+h}(\delta) - Z_{t+h}\right] \tag{7}$$

is the unconditional impulse response at horizon h due to shock δ. The difference is directly $Z_{t+h}(\delta) - Z_{t+h} = \Theta_{h,\cdot 1}\delta + \Gamma(Z_{t:*}(\delta)) - \Gamma(Z_{t:*})$, hence

$$\mathrm{IRF}_h(\delta) = \Theta_{h,\cdot 1}\delta + \mathbb{E}\left[\Gamma(Z_{t:*}(\delta)) - \Gamma(Z_{t:*})\right]. \tag{8}$$

Remark 2.3. In additively separable models, it is simple to see that $\Gamma(Z_{t:*})$ is also additively separable over lags of Z_t. Accordingly, the baseline and shock series have an additive form, as terms with time indices $s < t$ remain unaffected by the shock. Therefore,

[2] Stability of the linear VAR component is neither necessary nor sufficient for ensuring stability and stationarity of the entire nonlinear process, c.f. Assumption 9′ in Section 3 below.

(8) reduces to

$$\mathrm{IRF}_h(\delta) = \Theta_{h,\cdot 1}\delta + \mathbb{E}\left[\Gamma_0(Z_{t+h}(\delta)) - \Gamma_0(Z_{t+h})\right] + \ldots + \mathbb{E}\left[\Gamma_h(Z_t(\delta)) - \Gamma_h(Z_t)\right]. \tag{9}$$

Coefficients Γ_j are again functional, and still cannot be collected across $X_{t+j}(\delta)$ and X_{t+j}.

Closed-form computation of nonlinear IRFs is highly non-trivial. Even in the separable case (9), while one can linearly separate expectations in the impulse response formula, terms $\mathbb{E}\left[\Gamma_j Z_{t+j}(\delta) - \Gamma_j Z_{t+j}\right]$ for $0 \leqslant j \leqslant h$ cannot be meaningfully simplified further. Moreover, these expectations involve nonlinear functions of lags of Z_t and are impractical to derive explicitly. To avoid working with $\Theta(L)$ and $\Gamma(L)$, we now present an iterative algorithm which allows one to easily and efficiently compute nonlinear IRFs.

Proposition 2.1. *For any $h = 0, 1, \ldots, H$, with $H \geqslant 1$ fixed, if impulse response $\mathrm{IRF}_h(\delta)$ is finite and well-defined, it can be computed with the following steps:*

(i) *For $j = 0$, let $X_t(\delta) = X_t + \delta$ and $Y_t(\delta) = \mu_2 + G_2(Y_{t-1}, \ldots, Y_{t-p}, X_t(\delta), X_{t-1}, \ldots, X_{t-p}) + B_0^{21}(\epsilon_{1t} + \delta) + \xi_{2t}$.*

(ii) *For $j = 1, \ldots, h$, let*

$$X_{t+j}(\delta) = \mu_1 + A_{12}(L)Y_{t+j-1}(\delta) + A_{11}(L)X_{t+j-1}(\delta) + \epsilon_{1t+j},$$
$$Y_{t+j}(\delta) = \mu_2 + G_2(Y_{t-1}(\delta), \ldots, Y_{t-p}(\delta), X_t(\delta), X_{t-1}(\delta), \ldots, X_{t-p}(\delta)) + B_0^{21}\epsilon_{1t+j} + \xi_{2t+j}.$$

where $X_t(\delta)$ and $Y_t(\delta)$ are the shocked sequences determined by forward iteration after time t, equaling baseline sequences X_t and Y_t at lags before t, respectively.

Setting $Z_{t+j}(\delta) = (X_t(\delta), Y_t(\delta))'$, it holds $\mathrm{IRF}_h(\delta) = \mathbb{E}[Z_{t+j}(\delta) - Z_{t+j}]$.

Proposition 2.1 follows directly from the definition of the unconditional impulse response (7) combined with a direct forward iteration of (3), sidestepping the explicit MA(∞) formulation in (8). This approach dispenses from the need to simulate innovations $\{\epsilon_{t+j}\}_{j=1}^{h-1}$, as the joint distribution of $\{X_{t+h-1}, X_{t+j-1}, \ldots, X_t\}$ already contains all relevant path information. It also improves on the algorithm originally proposed in

Gonçalves et al. (2021): Their computations for step (ii) are recursive, whereas Proposition 2.1 gives iterative forms.

When the model is estimated from data, for residuals $\widehat{\epsilon}_{1t}$ and $\widehat{\xi}_{2t}$ it trivially holds

$$\begin{aligned} X_t &= \widehat{\mu}_1 + \widehat{A}_{12}(L)Y_{t-1} + \widehat{A}_{11}(L)X_{t-1} + \widehat{\epsilon}_{1t}, \\ Y_t &= \widehat{\mu}_2 + \widehat{G}_2(Y_{t-1}, \dots, Y_{t-p}, X_t, X_{t-1}, \dots, X_{t-p}) + \widehat{B}_0^{21}\widehat{\epsilon}_{1t} + \widehat{\xi}_{2t}. \end{aligned}$$

In practice, this means that one can numerically construct the shocked sequence as

$$\begin{aligned} \widehat{X}_{t+j}(\delta) &= \widehat{\mu}_1 + \widehat{A}_{12}(L)\widehat{Y}_{t+j-1}(\delta) + \widehat{A}_{11}(L)\widehat{X}_{t+j-1}(\delta) + \widehat{\epsilon}_{1t+j}, \\ \widehat{Y}_{t+j}(\delta) &= \widehat{\mu}_2 + \widehat{G}_2(\widehat{Y}_{t-1}(\delta), \dots, \widehat{Y}_{t-p}(\delta), \widehat{X}_t(\delta), \widehat{X}_{t-1}(\delta), \dots, \widehat{X}_{t-p}(\delta)) + \widehat{B}_0^{21}\widehat{\epsilon}_{1t+j} + \widehat{\xi}_{2t+j}, \end{aligned}$$

for $j = 1, \dots, h$ where $\widehat{X}_t(\delta) = X_t + \delta$, $\widehat{X}_{t-s} = X_{t-s}$ for all $s \geqslant 1$, and similarly for $\widehat{Y}_t(\delta)$.

3 Estimation

To discuss estimation, we will rewrite the equations in (3) with some minor reordering as

$$\begin{aligned} X_t &= \Pi_1' W_{1t} + \epsilon_{1t}, \\ Y_t &= \Pi_2' W_{2t} + \xi_{2t}, \end{aligned} \tag{10}$$

where $\xi_{2t} = B_0^{22}\epsilon_{2t}$, $\Pi_1 := (\eta_1, A_{1,11}, \cdots, A_{p,11}, A_{1,12}', \cdots, A_{p,12}')'$, $\Pi_1 \in \mathbb{R}^{1+pd}$,

$$\Pi_2 := \left[\begin{array}{c|c|c} & G_{1,2}(\cdot) & \\ \mu_2 & \cdots & B_0^{21} \\ & G_{d_Y,2}(\cdot) & \end{array}\right]',$$

$W_{1t} := (1, X_{t-1}, \dots, X_{t-p}, Y_{t-1}', \dots, Y_{t-p}')' \in \mathbb{R}^{pd}$, and $W_{2t} := (1, X_t, X_{t-1}, \dots, X_{t-p}, Y_{t-1}', \dots, Y_{t-p}', \epsilon_{1t})' \in \mathbb{R}^{2+pd}$. With a slight abuse of notation, similar to the one used in Example 2.3, we have written the functional terms in Π_2 as a "vector product", $G_2 \cdot (X_{t:t-p}', Y_{t-1:t-p}')' \equiv G_2(X_{t:t-p}, Y_{t-1:t-p})$, where G_2 is a vector of functions, one for each component of Y_t.

Whenever $\Pi_1 \neq 0$, W_{2t} is an infeasible vector of regressors due to term ϵ_{1t}. To estimate Π_2, one can use $\widehat{W}_{2t} = (1, X_t, Z_{t-1:t-p}', \widehat{\epsilon}_{1t})'$ instead, which contains generated regressors in the form of residual $\widehat{\epsilon}_{1t}$. A valid two-step estimation procedure (Gonçalves

et al., 2021) is:

1. Regress X_t on W_{1t} to get estimate $\widehat{\Pi}_1$, compute residuals $\widehat{\epsilon}_{1t} = X_t - \widehat{\Pi}_1' W_{1t}$;

2. Semiparametrically regress Y_t on $\widehat{W}_{2t}$ to get estimate $\widehat{\Pi}_2$.

There are many ways to implement Step 2, given that the literature on non- and semiparametric regression is mature. We rely on the sieve framework of Chen and Christensen (2015) as the workhorse to derive the main theoretical results. The sieve framework is known to be rich, encompassing e.g. neural networks (Chen and White, 1999, Shen et al., 2023).

3.1 Semiparametric Series Estimation

The semiparametric regression problem of Step 2 is more readily analyzed by working on each component of Y_t. For $i \in \{1, \ldots, d_Y\}$, consider

$$Y_{t,i} = \mu_{2,i} + G_{2,i}(Y_{t-1}, \ldots, Y_{t-p}, X_t, X_{t-1}, \ldots, X_{t-p}) + B_{0,i}^{21}\epsilon_{1t} + \xi_{2t,i}. \tag{11}$$

Let then $\pi_{2,i} := [\mu_{2,i},\ G_{2,i},\ B_{0,i}^{21}]'$. The regression equation for $\pi_{2,i}$ is thus $Y_i = \pi_{2,i}' W_2 + \xi_{2i}$, where $Y_i = (Y_{1,i}, \ldots, Y_{n,i})'$ and $\xi_{2i} = (\xi_{2t,1}, \ldots, \xi_{2t,n})'$. The estimation target is the conditional expectation $\pi_{2,i}(w) = \mathbb{E}[Y_{t,i} \,|\, W_{2t} = w]$ under the assumption $\mathbb{E}[\xi_{2t,i} \,|\, W_{2t}] = 0$.

Assume that $G_{2,i} \in \Lambda$, where Λ is a sufficiently regular function class to be specified in the following. Given a collection $b_{1\kappa}, \ldots, b_{\kappa\kappa}$ of $\kappa \geqslant 1$ basis functions belonging to sieve $\mathcal{B}_\kappa$, define $b^\kappa(\cdot) := (b_{1\kappa}(\cdot), \ldots, b_{\kappa\kappa}(\cdot))'$ and $B_\kappa := \big(b^\kappa(Y_{0:1-p}, X_{1:1-p}), \ldots, b^\kappa(Y_{n-1:n-p}, X_{n:n-p})\big)'$. For univariate functions, one can directly apply spline, wavelet and Fourier sieves; in the multivariate case, tensor-product sieves are straightforward generalizations (Chen and Christensen, 2015). To construct the final semiparametric sieve for $\pi_{2,i}$, let $b_{\pi,1K}, \ldots, b_{\pi,KK}$ be the sieve basis in $\mathbb{R} \times \mathcal{B}_\kappa \times \mathbb{R}$ for $\kappa \geqslant 1$ and $K = 2 + \kappa$ given by $b_{\pi,1K}(W_{2t}) = 1$, $b_{\pi,\ell K}(W_{2t}) = b_{\ell\kappa}(Y_{t-1:t-p}, X_{t:t-p})$ and $b_{\pi,KK}(W_{2t}) = \epsilon_{1t}$. for $2 \leqslant \ell \leqslant \kappa + 1$. Note that K, the overall size of the sieve, grows linearly in κ, which itself controls the effective dimension of the nonparametric component of the sieve, $b_{\pi,2K}, \ldots, b_{\pi,(\kappa+1)K}$. Introducing

$b_\pi^K(w) := (b_{\pi,1K}(w), \ldots, b_{\pi,KK}(w))'$ and $B_\pi := (b_\pi^K(W_{21}), \ldots, b_\pi^K(W_{2n}))'$, the generally *infeasible* least squares series estimator $\widehat{\pi}_{2,i}^*(w)$ is given by $\widehat{\pi}_{2,i}^*(w) = b_\pi^K(w)'(B_\pi' B_\pi)^{-1} B_K' Y_i$. Similarly, the feasible series regression matrix $\widehat{B}_\pi := (b_\pi^K(\widehat{W}_{21}), \ldots, b_\pi^K(\widehat{W}_{2n}))'$ yields the *feasible* least squares series estimator, $\widehat{\pi}_{2,i}(w) = b_\pi^K(w)'(\widehat{B}_\pi' \widehat{B}_\pi)^{-1} \widehat{B}_K' Y_i$.

To further streamline notation, wherever it does not lead to confusion, we will let π_2 be a generic coefficient vector belonging to $\{\pi_{2,i}\}_{i=1}^p$, as well as define $\widehat{\pi}_2$, Y and u_2 accordingly.

3.2 Distributional and Sieve Assumptions

To derive asymptotic consistency results, we begin by stating conditions on the basic probability structure of the model.

Assumption 3. $\{Z_t\}_{t\in\mathbb{Z}}$ is a strictly stationary and ergodic time series.

Assumption 4. $X_t \in \mathcal{X} \subset \mathbb{R}$, $Y_t \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ and $\epsilon_t \in \mathcal{E} \subset \mathbb{R}^d$ for all $t \in \mathbb{Z}$, where $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{E}$ are compact, convex sets with nonempty interior.

Assumption 3 follows both Gonçalves et al. (2021) and Chen and Christensen (2015). Note that, as W_{2t} depends only on $X_{t:t-p}$, $Y_{t-1:t-p}$ and ϵ_{1t}, the entries of ξ_{2t} in (10) are independent of W_{2t}, so that $\mathbb{E}[u_{2it} \mid W_{2t}] = 0$.

Assumption 4 implies that X_t, Y_t, as well as ϵ_t are bounded random variables. In (semi-)nonparametric estimation, imposing that X_t be bounded almost surely is a standard assumption. Since lags of Y_t and innovations ϵ_t contribute linearly to all components of Z_t, it follows that they too must be bounded. In practice Assumption 4 is not particularly restrictive, as many credibly stationary economic series often have reasonable implicit (e.g. inflation) or explicit bounds (e.g. employment rate).

Remark 3.1. Bounded support assumptions are relatively uncommon in time series econometrics, given the extensive literature available on linear models (Hamilton, 1994a, Lütkepohl, 2005, Kilian and Lütkepohl, 2017, Stock and Watson, 2016). Unbounded regressors are significantly more complex to handle when working in the nonparametric

setting. Chen and Christensen (2015) do work in weighted sup-norms, but their uniform results are stated only under a compact domain assumption. Avoiding Assumption 4 can be achieved with a change in the model's equations – e.g. the lags of Y_t only effect X_t via bounded functions – but this avenue also restricts the model. Establishing a general (uniform) theory of nonparametric regressions with unbounded data domains, on the other hand, is a complex question. For kernel, partitioning and nearest-neighbor methods and i.i.d. data, a handful of papers develop results in L^1 and L^2 norms, see Kohler et al. (2006, 2009) and Kohler and Krzyżak (2013). For wavelet estimators in the i.i.d. regression setting, Zhou (2022) provided the first sup-norm result in Besov spaces, if with suboptimal rates. Construction of a comprehensive nonparametric framework to handle non-independent, unbounded data should thus be considered an important objective of future research.

Without loss of generality, let $\mathcal{Y} = [0,1]^{d_Y}$ and $\mathcal{X} = [0,1]$.

Assumption 5. The unconditional densities of Y_t and X_t are uniformly bounded away from zero and infinity over $\mathcal{Y}$ and $\mathcal{X}$, respectively.

Assumption 6. For all $1 \leqslant i \leqslant d_Y$ the restriction of $G_{2,i}$ to $\mathcal{Y}^p \times \mathcal{X}^{1+p} \equiv [0,1]^{1+pd}$ belongs to the Hölder class $\Lambda^s([0,1]^{1+pd})$ of smoothness $s \geqslant 1$.

Assumptions 5 and 6 are classical in the nonparametric regression literature. Let then $\mathcal{W}_2 \subset \mathbb{R}^d$ be the domain of W_{2t}. By assumption, $\mathcal{W}_2$ is compact and convex and is given by the direct product $\mathcal{W}_2 = \{1\} \times \mathcal{Y}^p \times \mathcal{X}^{1+p} \times \mathcal{E}_1$,where $\mathcal{E}_1$ is the domain of structural innovations ϵ_{1t} i.e. $\mathcal{E} \equiv \mathcal{E}_1 \times \mathcal{E}_2$.

Assumption 7. Define $\zeta_{K,n} := \sup_{w \in \mathcal{W}_2} \|b_\pi^K(w)\|$ and $\lambda_{K,n} := [\lambda_{\min}(\mathbb{E}[\, b_\pi^K(W_{2t}) b_\pi^K(W_{2t})' \,])]^{-1/2}$. It holds: (i) there exist $\omega_1, \omega_2 \geqslant 0$ s.t. $\sup_{w \in \mathcal{W}_2} \|\nabla b_\pi^K(w)\| \lesssim n^{\omega_1} K^{\omega_2}$; (ii) there exist $\overline{\omega}_1 \geqslant 0$, $\overline{\omega}_2 > 0$ s.t. $\zeta_{K,n} \lesssim n^{\overline{\omega}_1} K^{\overline{\omega}_2}$; (iii) $\lambda_{\min}(\mathbb{E}[\, b^K(W_{2t}) b^K(W_{2t})' \,]) > 0$ for all K and n.

Assumption 7 provides mild regularity conditions on the families of sieves that can be used for the series estimator. More generally, letting $\mathcal{W}_2$ be compact and rectangular

makes Assumption 7 hold for commonly used basis functions (Chen and Christensen, 2015). In particular, Assumption 7(i) holds with $\omega_1 = 0$ since the domain is fixed over the sample size. What is also needed is that the nonparametric components of the sieve given by $b_{\pi,1K}, \ldots, b_{\pi,KK}$ are able to approximate $G_{2,i}$ well enough. Throughout this paper, we will consider specific families of sieves, which are known to fulfill the regularity conditions spelled out in Assumption 7. The approximation properties of these sieves are well understood (Chen, 2007).[3]

Assumption 8. Sieve $\mathcal{B}_\kappa$ belongs to $\mathrm{BSpl}(\kappa, \mathcal{W}_2, r)$ or $\mathrm{Wav}(\kappa, \mathcal{W}_2, r)$, the tensor B-spline and tensor wavelet sieve, respectively, of degree r over $\mathcal{W}_2$, with $r \geqslant \max\{s, 1\}$.

We define $\widetilde{b}_\pi^K(w) := \mathbb{E}[\, b_\pi^K(W_{2t}) b_\pi^K(W_{2t})' \,]^{-1/2} b_\pi^K(w)$ and $\widetilde{B}_\pi := \big(\widetilde{b}_\pi^K(W_{21}), \ldots, \widetilde{b}_\pi^K(W_{2n})\big)'$ to be the orthonormalized vector of basis functions and the orthonormalized regression matrix, respectively. To derive uniform converges rates under dependence, we require that the Gram matrix of orthonormalized sieve converges to the identity.

Assumption 9. It holds that $\|(\widetilde{B}_\pi' \widetilde{B}_\pi / n) - I_K\| = o_P(1)$.

Chen and Christensen (2015) introduced Assumption 9 as a key ingredient for their proofs, while also showing that it holds whenever $\{W_{2t}\}_{t\in\mathbb{Z}}$ is either an exponential or algebraic β-mixing process. Unfortunately, mixing conditions are opaque in terms of their connection to the model specification, as they rely on bounding the worst-case "independence gap" between probability events (see Appendix A). We extend their approach to the case of geometrically decaying physical dependence, a metric proposed by Wu (2005), a setting where many estimation and inference results have been derived, see for example Wu et al. (2010), Wu (2011), Chen et al. (2016) and references within.

Assumption 9′. Let $\{Z_t\}_{t\in\mathbb{Z}}$ be such that we can write $Z_{t+h} = \Phi^{(h)}(Z_t, \epsilon_{t+1:t+h})$ for all $h \geqslant 1$, nonlinear maps $\Phi^{(h)}$ and innovations $\{\epsilon_t\}_{t\in\mathbb{Z}}$. Then, for $r \geqslant 2$, there exists constants $a_1 > 0$, $a_2 > 0$ and $\tau \in (0, 1]$ such that it holds

$$\sup_t \left\| Z_{t+h} - \Phi^{(h)}(Z_t', \epsilon_{t+1:t+h}) \right\|_{L^r} \leqslant a_1 \exp(-a_2\, h^\tau).$$

[3] See also Chen (2013), Belloni et al. (2015) for additional discussion and examples of sieve families.

Assumption 9 is subsumed by Assumption 9′. Using physical dependence measure, we argue that it is also possible to swap mixing conditions with more explicit, primitive conditions derived exclusively in terms of model specification (1). In particular, for specific semiparametric model specifications, it is possibly to verify Assumption 9′ directly by leveraging stability/contractivity theory of dynamic systems. We refer the reader to Appendix A for an in-depth discussion of dependence and physical conditions.

3.3 Uniform Convergence and Consistency

We can now state our main result: The two-step estimation procedure for (10) provides consistent estimates.

Theorem 3.1. *Let* $\{Z_t\}_{t\in\mathbb{Z}}$ *be determined by structural model* (4). *Under Assumptions 1, 3, 4, 5, 6, 7, 8 and 9′, let* $\widehat{\Pi}_1$ *and* $\widehat{\Pi}_2$ *be the least squares and two-step semiparametric series estimators for* Π_1 *and* Π_2, *respectively. Then,* $\|\widehat{\Pi}_1 - \Pi_1\|_\infty = O_P(n^{-1/2})$ *and*

$$\|\widehat{\Pi}_2 - \Pi_2\|_\infty \leqslant O_P\left(\zeta_{K,n}\lambda_{K,n}\frac{K}{\sqrt{n}}\right) + \|\widehat{\Pi}_2^* - \Pi_2\|_\infty,$$

where $\widehat{\Pi}_2^*$ *is the infeasible series estimator involving* ϵ_{1t}.

The proof is a moderate extension of Theorem 1 in Chen and Christensen (2015): Sup-norm bounds for $\|\widehat{\Pi}_2^* - \Pi_2\|_\infty$ follow immediately from their Lemma 2.3 and Lemma 2.4. In particular, choosing the optimal nonparametric rate $K \asymp (n/\log(n))^{d/(2s+d)}$ for the infeasible estimator would yield $\|\widehat{\Pi}_2^* - \Pi_2\|_\infty = O_P((n/\log(n))^{-s/(2s+d)})$. The condition for consistency in Theorem 3.1 reduces to $K^{3/2}/\sqrt{n} = o(1)$, since for B-spline and wavelet sieves $\lambda_{K,n} \lesssim 1$ and $\zeta_{K,n} \lesssim \sqrt{K}$. It is simple to show that, if for the feasible estimator $\widehat{\Pi}_2$ the same rate $(n/\log(n))^{d/(2s+d)}$ is chosen for K, consistency is fulfilled assuming e.g. $s \geqslant 1$ and $d = 1$, such as in the setting of the additively separable model in Example 2.3. A number of methods can be used to select K in practice: Cross-validation, generalized cross-validation and Mallow's criterion are commonly employed (Li and Racine, 2009). In the case of piece-wise splines, once size is selected, knots can be chosen to be the K

uniform quantiles of the data. In simulations and applications, for simplicity, we select sieve sizes manually and locate knots approximately following empirical quantiles.

4 Impulse Response Analysis

Once the model's coefficients are estimated, derivation of nonlinear impulse responses must be addressed. To ensure compatibility with bounded support assumptions, we introduce an extension of the classical IRF definition, termed *relaxed impulse response function*. We then show that nonlinear relaxed IRFs can be consistently estimated, and uniformly so with respect to shocks picked within a compact range.

4.1 Relaxed Shocks

Under Assumptions 4 and 5, the standard construction of impulse responses following Section 2.3 is, unfortunately, improper. This is immediately seen by noticing that, at impact, $X_t(\delta) = X_t + \delta$, meaning that $\mathbb{P}(X_t(\delta) \notin \mathcal{X}) > 0$ since there is a translation of size δ in the support of X_t. To address this problem, we introduce an extension to the standard additive shock that is used to define impulse responses.

We begin by defining mean-shift shocks, that is, shocks such that the distribution of time t innovations is shifted to have mean δ, while retaining compact support almost surely.

Definition 4.1. *A mean-shift structural shock $\epsilon_{1t}(\delta)$ is a transformation of ϵ_{1t} such that $\mathbb{P}(\epsilon_{1t}(\delta) \in \mathcal{E}_1) = 1$ and $\mathbb{E}[\epsilon_{1t}(\delta)] = \delta$.*

With a mean-shift shock, at impact it holds $X_t(\delta) = X_t + (\epsilon_{1t}(\delta) - \epsilon_{1t})$. In the standard setting, where $\mathbb{E}[\epsilon_t] = 0$ and $\mathcal{E}_1 \equiv \mathbb{R}$, $\epsilon_{1t}(\delta) = \epsilon_{1t} + \delta$ is clearly valid. More generally, however, imposing $\mathbb{E}[\epsilon_{1t}(\delta)] = \delta$ requires that the distribution of ϵ_{1t} be known. If instead one is willing to assume only that $\mathbb{E}[\epsilon_{1t}(\delta)] \approx \delta$, it is possible to sidestep this need by introducing a *shock relaxation function*.

Definition 4.2. *Assume $\mathcal{E}_1 = [a, b]$. A shock relaxation function is a map $\rho : \mathcal{E}_1 \to [0, 1]$ such that $\rho(e) = 0$ for all $e \in \mathbb{R} \setminus \mathcal{E}_1$, $\rho(e) \geqslant 0$ for all $e \in \mathcal{E}_1$ and there exists $e_0 \in \mathcal{E}_1$ for which $\rho(e_0) = 1$. Moreover, for a given shock $\delta \in \mathbb{R}$,*

(i) If $\delta > 0$, ρ is said to be right-compatible with δ if $e + \rho(e)\delta \leqslant b$ for all $e \in \mathcal{E}_1$.

(ii) If $\delta < 0$, ρ is said to be left-compatible with δ if $e + \rho(e)\delta \geqslant a$ for all $e \in \mathcal{E}_1$.

(iii) ρ is compatible with shock magnitude $|\delta| > 0$ if it is both right- and left-compatible.

By setting $\epsilon_{1t}(\delta) = \epsilon_{1t} + \delta\rho(\epsilon_{1t})$ for a ρ compatible with δ, it follows that $X_t(\delta) = X_t + \delta\rho(\epsilon_{1t})$ and $|\mathbb{E}[\epsilon_{1t}(\delta)]| = |\delta\mathbb{E}[\rho(\epsilon_{1t})]| \leqslant |\delta|$ since $\mathbb{E}[\rho(\epsilon_{1t})] \in [0, 1)$ by definition of ρ. If ρ is a bump function, a relaxed shock is a structural shock that has been mitigated proportionally to the density of innovations at the edges of $\mathcal{E}_1$ and the squareness of ρ. It is important to emphasize that shock relaxation is a generalization of standard shock designs. Indeed, when $\mathcal{X} = \mathbb{R}$ and $\mathcal{E}_1 = \mathbb{R}$, $\rho = \mathbb{I}\{[-\infty, \infty]\}$ is a relaxation function compatible with all $\delta \in \mathbb{R}$. Nonetheless, we may also wonder of how much information on nonlinear term G_2 we can recover at the "boundary" of a finite sample. If X_t is unbounded but well-concentrated, even under strong smoothness conditions and strictly positive density, little can be learned about the *local* structure of regression functions in regions of low density.[4]

Remark 4.1. When studying impulse responses, a researcher should be primarily interested in shock δ itself, not ρ. In this paper, and more specifically in Sections 5 and 6, we choose ρ to be an symmetric exponential bump function, $\rho \in \{x \mapsto \mathbb{I}\{x \leqslant c\} \exp(1 + (|x/c|^\alpha - 1)^{-1}) \mid \alpha > 0\}$. This $\mathcal{C}^\infty$ bump class is widely studied in both functional (Mitrovic and Zubrinic, 1997) and Fourier analysis (Stein and Shakarchi, 2011).[5]

[4] In our regression setting, for example, Theorem 1 in Kohler et al. (2009) on L_2 kernel regression error, assuming $\mathbb{E}[|X_t|^\beta] \leqslant M < \infty$ for some constant $\beta > 2s$, would require the bandwidth to grow over $\mathcal{X}$ faster than $|X_t|$. This question is also linked to issues in kernel density estimation over sets with boundary, see e.g. Karunamuni and Alberts (2005), Malec and Schienle (2014), Berry and Sauer (2017) and references therein.

[5] For generic shock distributions, one can for also consider the class $\{x \mapsto \mathbb{I}\{a \leqslant x \leqslant b\} \exp(1 + (|2(x - b)/(b - a) + 1|^\alpha - 1)^{-1}) \mid \alpha > 0\}$ of exponential bump functions with domain $[a, b] \subset \mathbb{R}$.

We aim to set α to be as large as possible to minimize distortions from a linear shift, while retaining compatibility with $\delta \in \mathcal{D}$, where $\mathcal{D}$ is a set of shocks of empirical interest.

4.2 Relaxed Impulse Response Consistency

We will now study relaxed impulse responses in the setting of additively separable models. The reason is twofold: First, additive separability is a very common assumption in applied work, as we shall impose it in the models applied in both Section 5 and 6. Second, collecting nonlinear terms over lags significantly streamlines notation and analysis. It would be straightforward, if tedious, to extend our derivations below to the more general setting of Theorem 3.1.

Given $\delta \in \mathbb{R}$ and compatible shock relaxation function ρ, let $\widetilde{\delta}_t := \delta\rho(\epsilon_{1t})$. Starting from a path of realization $X_{t+j:t}$ and (9), the relaxed shock path is

$$X_{t+j}(\widetilde{\delta}_t) = X_{t+j} + \Theta_{j,11}\widetilde{\delta}_t + \sum_{k=1}^{j}\Big[\Gamma_{k,11}X_{t+j-k}(\widetilde{\delta}_t) - \Gamma_{k,11}X_{t+j-k}\Big] = \gamma_j(X_{t+j:t};\widetilde{\delta}_t).$$

The relaxed-shock impulse response is thus given by

$$\widetilde{\mathrm{IRF}}_h(\delta) := \mathbb{E}[Z_{t+j}(\widetilde{\delta}_t) - Z_{t+j}] = \Theta_{h,\cdot 1}\delta\,\mathbb{E}\left[\rho(\epsilon_{1t})\right] + \sum_{k=1}^{j}\mathbb{E}\Big[\Gamma_k X_{t+j-k}(\widetilde{\delta}_t) - \Gamma_k X_{t+j-k}\Big].$$

For $1 \leqslant \ell \leqslant d$, we let $V_{j,\ell}(\delta)$ be the sample analog of the horizon j nonlinear effect on the ℓth variable,

$$V_{j,\ell}(\delta) := \frac{1}{n-j}\sum_{t=1}^{n-j}\Big[\Gamma_{j,\ell}\gamma_j(X_{t+j:t};\widetilde{\delta}_t) - \Gamma_{j,\ell}X_{t+j}\Big] = \frac{1}{n-j}\sum_{t=1}^{n-j}v_{j,\ell}(X_{t+j:t};\widetilde{\delta}_t)$$

where $\Gamma_{j,\ell}$ is the ℓth component of functional vector Γ_j. As ϵ_{1t} is not universally observable, we introduce its residual counterpart, $\widehat{\widetilde{\delta}}_t = \delta\rho(\widehat{\epsilon}_{1t})$. The associated plug-in sample estimates are $\widehat{V}_{j,\ell}(\delta) = (n-j)^{-1}\sum_{t=1}^{n-j}\widehat{v}_{j,\ell}\big(X_{t+j:t};\widehat{\widetilde{\delta}}_t\big)$, $\widehat{v}_{j,\ell}(X_{t+j:t};\widehat{\widetilde{\delta}}_t) = \widehat{\Gamma}_{j,\ell}\widehat{\gamma}_j(X_{t+j:t};\widehat{\widetilde{\delta}}_t) - \widehat{\Gamma}_{j,\ell}X_{t+j}$, and

$$\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) = \widehat{\Theta}_{h,\cdot 1}\delta\, n^{-1}\sum_{t=1}^{n}\rho(\widehat{\epsilon}_{1t}) + \sum_{j=0}^{h}\widehat{V}_{j,\ell}(\delta).$$

Our next theorem proves consistency of the relaxed impulse responses estimator based on semiparametric series estimates. We leverage the sup-norm bounds of Theorem 3.1 to

derive a result that is uniform in δ over a compact interval $[-\mathcal{D}, \mathcal{D}]$, $\mathcal{D} > 0$. This allows us to make valid comparisons between IRFs due to shocks of different size.

Theorem 4.1. *Let* $\widehat{\widetilde{\text{IRF}}}_{h,\ell}(\delta)$ *be the semiparametric estimate for the horizon* h *relaxed shock IRF of variable* ℓ *based on relaxation function* ρ *with compatibility range* $[-\mathcal{D}, \mathcal{D}]$. *Under the assumptions in Theorem 3.1 and Assumption 2,*

$$\sup_{\delta \in [-\mathcal{D}, \mathcal{D}]} \left| \widehat{\widetilde{\text{IRF}}}_{h,\ell}(\delta) - \widetilde{\text{IRF}}_{h,\ell}(\delta) \right| = o_P(1)$$

for any fixed integers $0 \leqslant h < \infty$ *and* $1 \leqslant \ell \leqslant d$.

Remark 4.2. By construction of $\widehat{\widetilde{\text{IRF}}}_{h,\ell}(\delta)$, Proposition 2.1 remains valid when computing $\widetilde{\text{IRF}}_h(\delta)$ instead of $\text{IRF}_h(\delta)$. The only adjustment to be made is that in step (i) one must set $X_t(\delta) = X_t + \delta\rho(\epsilon_{1t})$ and iterate forward accordingly. Assumptions 1, 3 and 9′ ensure that the IRFs of interest are well-defined.

Remark 4.3. Our definition of compatible relaxation function is *static*, as it considers only the impact effect of a shock. Nonetheless, $X_t(\delta) \in \mathcal{X}$ for all t must hold to properly define $\widetilde{\text{IRF}}_h(\delta)$. In theory, given δ, one can always either expand $\mathcal{X}$ or strengthen ρ so that compatibility is enforced at all horizons $1 \leqslant h \leqslant H$. In simulations, the choice of domains and relaxation functions can be done transparently. When working with empirical data, unless X_t is exogenous or strictly autoregressive, more care has to be taken to check that there is no dynamic domain violation. In Section 6.2, where X_t is an endogenous series, we discuss such robustness check.

5 Simulations

To analyze the performance of the two-step semiparametric estimation strategy discussed above, we begin by considering the two simulation setups employed by Gonçalves et al. (2021). The setup involves comparing the bias and MSE of the estimated relaxed shocked impulse response functions for different methods. Additionally, we provide simulations under a misspecification design which highlight how in larger samples the nonparametric

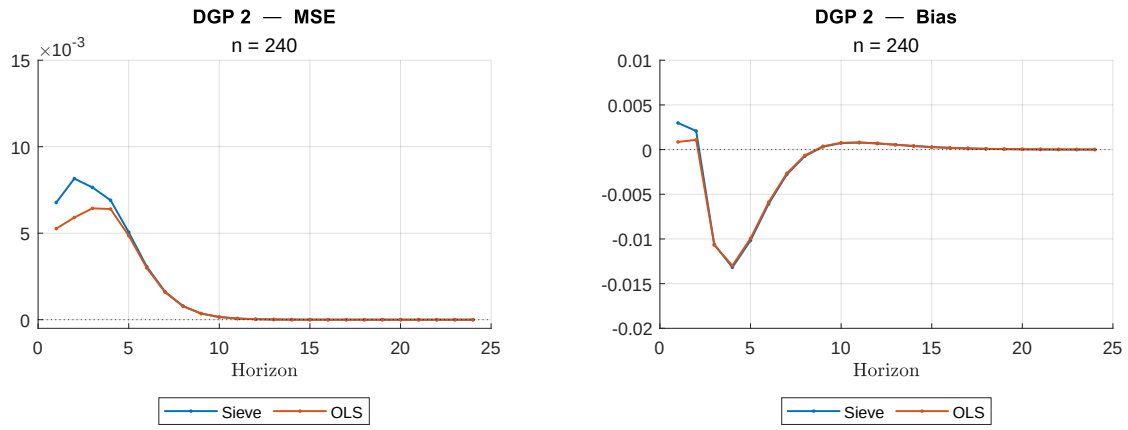


Figure 1: Simulation results for DGP 2 with $\delta = +1$.

sieve estimator consistently recovers impulse responses, while a least-squares estimator constructed with a pre-specified nonlinear transform does not. We compute MSE and bias of both the parametric IRFs obtained via least squares regression on transformed regressors and the semiparametric two-step estimator using 10 000 Monte Carlo replications. Population impulse responses are computed with 10^5 replications. In all setups, we use a cubic B-spline sieve.

Benchmarks. Like in Gonçalves et al. (2021), we consider two simulation setups: A bivariate design with identified shocks (DGPs 1-3), and a three-variable design with partial block-recursive identification (DGPs 4-6). In all cases, we consider a sample of size of $n = 240$, which is realistic for most macroeconomic data settings: this is approximately equivalent to 20 years of monthly data or 60 years of quarterly data (Gonçalves et al., 2021).

Due to space constraints, we discuss here only a bi-variate simulation design with a shock $\delta = +1$. We set either $X_t = \epsilon_{1t}$ (DGP 1), $X_t = 0.5X_{t-1} + \epsilon_{1t}$ (DGP 2) or $X_t = 0.5X_{t-1} + 0.2Y_{t-1} + \epsilon_{1t}$ (DGP 3), and

$$Y_t = 0.5Y_{t-1} + 0.5X_t + 0.3X_{t-1} - 0.4\max(0, X_t) + 0.3\max(0, X_{t-1}) + \epsilon_{2t}.$$

Innovations ϵ_{1t} and ϵ_{2t} are drawn as independent, truncated standard Gaussian variables

over $[-3, 3]$. The shock relaxation function is $\rho(z) = \mathbb{I}\{|z| \leqslant 3\} \exp\left(1 + [|z/3|^4 - 1]^{-1}\right)$, c.f. Remark 4.1. In Figure 1 we show MSE and bias curves for the IRF on Y_t in DGP 2, where X_t is an exogenous AR(1) process. One case see that the sieve IRF leads only to a minor increase in mean squared error at short horizon compared to directly estimating the parameter of the true specification. This marginal increase in MSE is consistent across DGPs 1 through 3.

These simulations show that there is negligible loss of efficiency in terms of either MSE or bias when implementing the fully flexible semiparametric estimates at realistic sample sizes. We confirm these results when studying DGPs 4-6, where estimation of structural matrix B_0 is included in the regression problem. Detailed results can be found in Appendix C.

Misspecified Model. To assess the robustness of the proposed semiparametric approach versus the parametric nonlinear model, we consider a modified version (DGP 7):

$$\begin{aligned} X_t &= 0.8X_{t-1} + \epsilon_{1t}, \\ Y_t &= 0.5Y_{t-1} + 0.9\varphi(X_t) + 0.5\varphi(X_{t-1}) + \epsilon_{2t}. \end{aligned} \tag{12}$$

where $\varphi(x) := (x-1)(0.5+\tanh(x-1)/2)$. In this design, we assume that the researcher's prior is $\varphi(x) = \max(0, x)$, as in the benchmark simulations. To emphasize the difference in estimated IRFs, in this setup we focus on $\delta = \pm 2$ and $n = 2400$; innovations ϵ_{1t} and ϵ_{2t} are drawn from a standard Gaussian distribution truncated over $[-5, 5]$, and $\rho(z) = \exp(1 + [|z/5|^{3.9} - 1]^{-1})$. As Figure 2 shows, positive-shock parametric nonlinear IRF estimates are severely biased, while semiparametric sieve IRFs show comparatively negligible error: This yields an up to 4 times reduction of overall MSE at short horizons. Appendix C provides additional simulation results proving that the same improvements hold when $\delta = -2$. There, we also discuss the setting where $\varphi(x)$ is replaced with map $\widetilde{\varphi}(x) = \varphi(x + 1)$, which agrees closely with $\max(0, x)$. In this last setting, we find that parametric nonlinear regression actually dominates in MSE and bias terms. As one might expect, therefore, parametric modeling is optimal only in cases where a good model prior

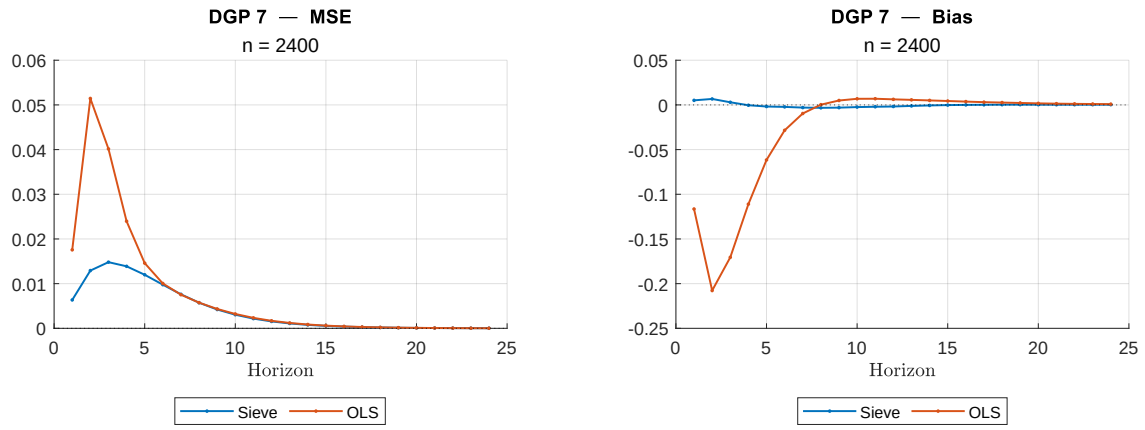


Figure 2: Simulation results for DGP 7 with shock $\delta = +2$.

is available.

6 Empirical Applications

In this section, we showcase the practical utility of the proposed semiparametric sieve estimator by considering two applied exercises.

6.1 Monetary Policy Shocks

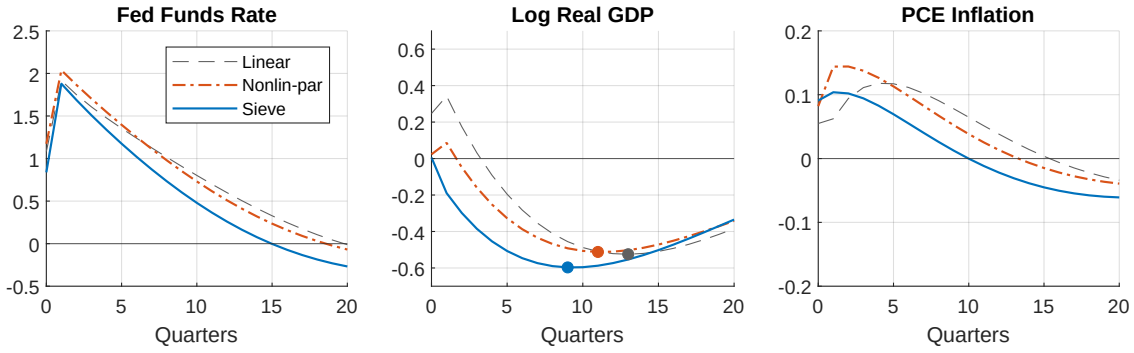
A four-variable model is set up identically to the one analyzed by Gonçalves et al. (2021) based on Tenreyro and Thwaites (2016). Let $Z_t = (X_t, \text{FFR}_t, \text{GDP}_t, \text{PCE}_t)'$, where X_t is the series of narrative U.S. monetary policy shocks, FFR_t is the federal funds rate, GDP_t is log-real GDP and PCE_t is PCE inflation.[6] As a pre-processing step, GDP is transformed to log GDP and then linearly detrended. The data is available quarterly and spans from 1969:Q1 to 2007:Q4. As in Tenreyro and Thwaites (2016), we use a model with one lag, $p = 1$. Narrative shock X_t is considered to be an i.i.d. sequence, i.e. $X_t = \epsilon_{1t}$, therefore we assume no dependence on lagged variables when implementing the pseudo-reduced

[6]In Gonçalves et al. (2021) p. 122, it is mentioned that CPI inflation is included in the model, but both in the replication package made available by one the authors (`https://sites.google.com/site/lkilian2019/research/code`) from which we source the data, and in Tenreyro and Thwaites (2016), PCE inflation is used instead. Moreover, the authors say that both the FFR and PCE enter the model in first differences, yet in their code these variables are kept in levels. We thus consider a model in levels to allow for a proper comparison between estimation methods.

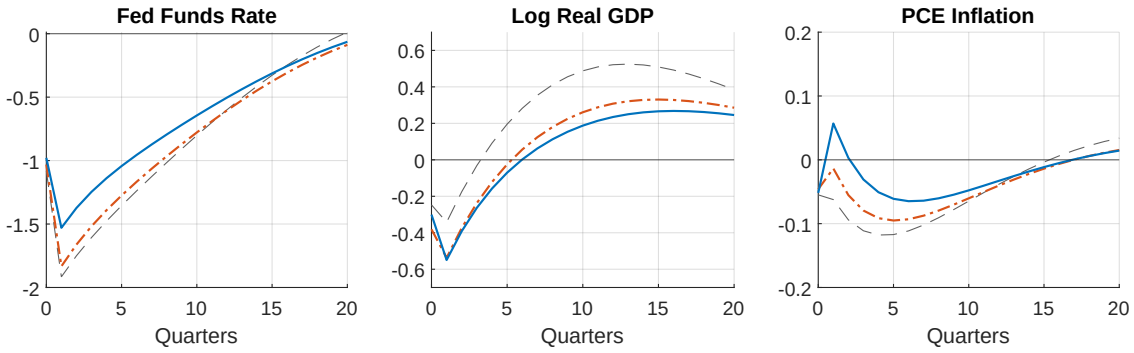
form (3). Like in Gonçalves et al. (2021), we consider positive and negative shocks of size $|\delta| = 1$ and choose $\rho(z) = \mathbb{I}\{|z| \leqslant 4\} \exp(1 + [|z/4|^6 - 1]^{-1})$ to be the shock relaxation function. Figure D.9 in the Online Appendix provides a check for the validity of ρ given the sample distribution of X_t. Knots for sieve estimation are located at $\{-1, 0, 1\}$. The model is block-recursive: U.S. monetary policy shocks are identified without the need to impose additional assumptions on the remaining shocks. Gonçalves et al. (2021), following Tenreyro and Thwaites (2016), use two nonlinear transformations, $F(x) = \max(0, x)$ and $F(x) = x^3$, to try to gauge how negative versus positive and large versus small shocks, respectively, affect the U.S. macroeconomy. They find the two maps yield very similar responses, so we focus on comparing the IRFs estimated via sieve regression with the ones obtained by setting $F(x) = \max(0, x)$, as well as linear IRFs. Figure 3 plots estimated impulse responses to both positive and negative monetary policy shocks. The impact on the federal funds rate is consistent across all three procedures. The semiparametric nonlinear response for GDP, unlike in the case of linear and parametric nonlinear IRFs, is nearly zero at impact and has a monotonic decrease until around 10 quarters ahead. The change in shape is meaningful, as the procedure of Gonçalves et al. (2021) still yields a small short-term upward jump in GDP when a monetary tightening shock hits. Moreover, after the positive shock, the sieve GDP responses reaches its lowest value 4 and 2 quarters before the linear and parametric nonlinear responses, while its size is 13% and 16% larger, respectively.[7] Finally, the sieve PCE response is positive for a shorter interval, but looks to be more persistent once it turns negative also 10 months after impact.

When the shock is expansionary, one sees that the semiparametric FFR response is marginally mitigated compared to the alternative estimates. An important puzzle is due to the clearly negative impact on GDP: Both types of nonlinear responses show a drop in output in the first 5 quarters. Such a quick change seems unrealistic, as one does not expect inflation to suddenly reverse sign, but, as Gonçalves et al. (2021) also remark, the overall impact on inflation of both shocks is small when compared to the change in federal

[7] The strength of this effect changes across different shocks sizes, as Figure D.7 in Appendix D proves. As shocks sizes get smaller, nonlinear IRFs, both parametric and sieve, show decreasing negative effects.



(a) $\delta = +1$



(b) $\delta = -1$

Figure 3: Effect of an unexpected U.S. monetary policy shock on federal funds rate, GDP and inflation. Linear (gray, dashed), parametric nonlinear with $F(x) = \max(0, x)$ (red, point-dashed) and sieve (blue, solid) structural impulse responses. For $\delta = +1$, the lowest point of the GDP response is marked with a dot.

funds rate.

6.2 Uncertainty Shocks

Traditional central bank policymaking is heavily guided by the principle that a central bank can and should influence expectations: Therefore, controlling the (perceived) level of ambiguity in current and future commitments is key. Istrefi and Mouabbi (2018) provide an analysis of the impact of unforeseen changes in the level of subjective interest rate uncertainty on the macroeconomy. They derive a collection of new indices based on short- and long-term profession forecasts. Their empirical study goes in depth into studying the different components that play a role in transmitting uncertainty shocks, but for the sake

of simplicity my evaluation will focus only on their 3-months-ahead uncertainty measure for short-term interest rate maturities (3M3M) and the US economy.

Like in Istrefi and Mouabbi (2018), let $Z_t = (X_t, \text{IP}_t, \text{CPI}_t, \text{PPI}_t, \text{RT}_t, \text{UR}_t)'$ be a vector where X_t is the chosen uncertainty measure, IP_t is the (log) industrial production index, CPI_t is the CPI inflation rate, PPI_t is the producer price inflation rate, RT_t is (log) retail sales and UR_t is the unemployment rate. The nonlinear model specification is given by

$$Z_t = \mu + A_1 Z_{t-1} + A_2 Z_{t-1} + F_1(X_{t-1}) + F_2(X_{t-2}) + DW_t + u_t,$$

where W_t includes a linear time trend and oil price OIL_t.[8] The data has monthly frequency and spans the period between May 1993 and July 2015.[9] Note here that, following the identification strategy of Gonçalves et al. (2021), nonlinear functions F_1 and F_2 are to be understood as not effecting X_t, which is the structural variable. The linear VAR specification of Istrefi and Mouabbi (2018) is recovered by simply assuming $F_1 = F_2 = 0$ prior to estimation. Since they use recursive identification and order the uncertainty measure first, this model too is block-recursive. We consider a positive shock with intensity $\delta = \sigma_{\epsilon,1}$, where $\sigma_{\epsilon,1}$ is the standard deviation of structural innovations. In this empirical exercise, the relaxation function is $\rho(z) = \mathbb{I}\{|z| \leqslant 1/4\} \exp(1 + [|4x|^8 - 1]^{-1})$ and we set $\{0.1, 0.3\}$ to be the cubic spline knots. As 3M3M is a non-negative measure of uncertainty, some care must be taken to make sure that the shocked paths for X_t do not reach negative values. Figure D.10 in Appendix D shows that the relaxation function is compatible, and also that the shocked nonlinear paths of X_t with impulse δ and δ' all do not cross below zero.

Figure 4 presents both the linear and nonlinear structural impulse responses obtained. Importantly, even though Istrefi and Mouabbi (2018) estimate a Bayesian VAR model and here we consider a frequentist vector autoregressive benchmark, the shape of

[8] Inclusion of linear exogenous variables in the semiparametric theoretical framework in Section 3 is straightforward as long as one can assume that they are stationary and weakly dependent. The choice of using $p = 2$ is identical to that of the original authors, based on BIC.

[9] I reuse the original data employed by the authors, who kindly shared it upon request, but rescale retail sales (RT_t) so that the level on January 2000 equals 100.

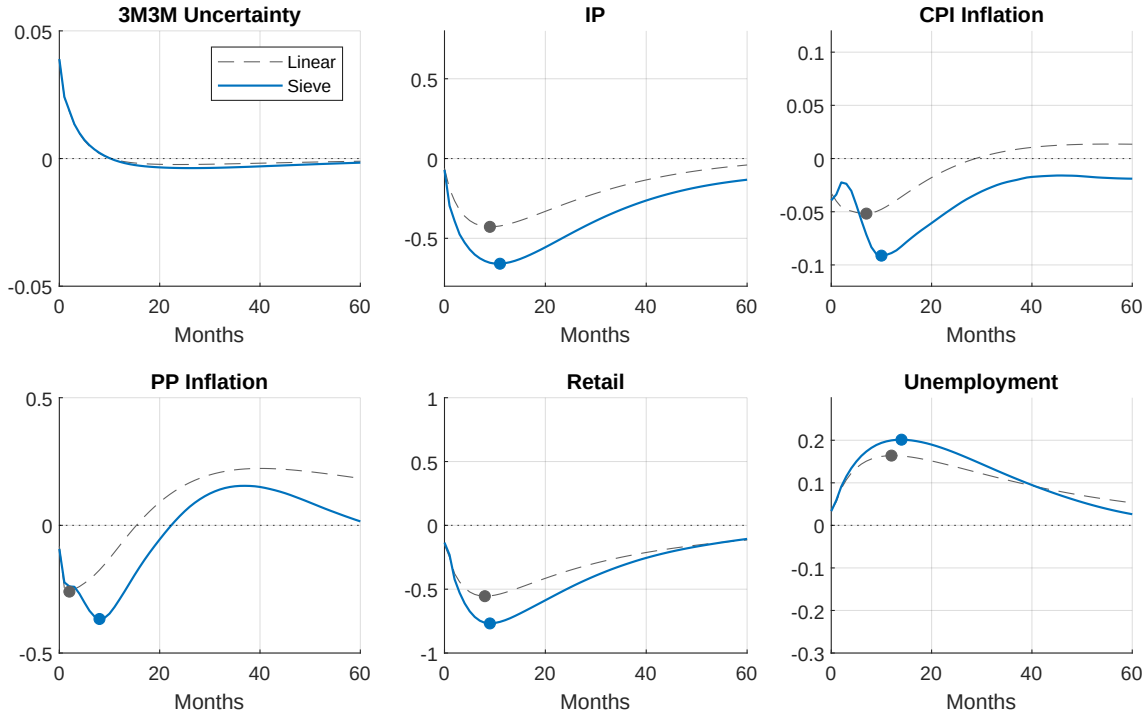


Figure 4: Effect of an unexpected, one-standard-deviation uncertainty shock to US macroeconomic variables. Linear (gray, dashed) and sieve (blue, solid) structural impulse responses. The extreme points of the responses are marked with a dot.

the IRFs is retained, c.f. the median response in the top row of their Figure 4. When uncertainty increases, industrial production drops, and the size and extent of this decrease is intensified in the nonlinear responses. In fact, the sieve IP response reaches a value that is 54% lower than that of the respective linear IRF.[10] A similar behavior holds true for retail sales (38% lower) and unemployment (23% higher), proving that this shock is more profoundly contractionary than suggested by the linear VAR model. Further, CPI and PP inflation both display short-term fluctuations, which strengthen the short- and medium-term impact of the shock. CPI and PP nonlinear inflation responses are 76% and 41% stronger than their linear counterpart, respectively. These differences show that linear IRFs might be both under-estimating the short-term intensity and misrepresenting long-term persistence of inflation reactions. From another perspective, Nowzohour and Stracca (2020) presented evidence that consumer consumption growth, credit growth

[10]Figure D.11 in Appendix D confirms that this difference is consistent over a range of shock sizes, too.

and unemployment do not co-move with the policy uncertainty index (EPU) of Baker et al. (2016), but are negatively correlated with financial volatility. Given the strength of nonlinear IRFs, this discrepancy may also suggest that the 3M3M uncertainty measure partially captures the financial channel, too.

The introduction of nonlinear terms in the structural VAR of Istrefi and Mouabbi (2018) thus provides evidence that fundamental impulse response features might otherwise be missed. Indeed, Figure D.8 in Appendix D - which plots regression functions of endogenous variables with respect to X_t - shows that high and low uncertainty levels may have significantly different effects on endogenous economic variables. In particular, at the second lag, tail effects appear to be milder, while at low levels changes in uncertainty have more pronounced impact.

7 Conclusion

This paper studies the application of semiparametric series estimation to the problem of structural impulse response analysis for time series. After first discussing the partial identification model setup, we have used the conditions of system contractivity and stability to derive physical measures of the dependence for nonlinear systems. In turn, these allow to derive primitive conditions under which series estimation can be employed and structural IRFs are consistently estimated. Simulation results prove that this approach is valid in moderate samples and has the added benefit of being robust to misspecification of the nonlinear model components. Finally, two empirical applications showcase the utility in departing from both linear and parametric nonlinear specifications when estimating structural responses.

A key aspect that we have not touched upon is inference in the form of confidence intervals. This however seems feasible in light of the uniforms inference results obtained by e.g. Belloni et al. (2015) in the i.i.d. setting and Li and Liao (2020) for time series data. Studying other sieve spaces, such as neural networks (Chen and White, 1999, Farrell et al.,

2021) or shape-preserving sieves (Chen, 2007), would also be highly desirable. Finally, in the spirit of Kang (2021), deriving new inference results that are uniform in the selection of series terms is important, as, in practice, the sieve should be tuned in a data-driven way.

References

Auerbach, A. J. and Gorodnichenko, Y. (2012). Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy*, 4(2):1–27.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Berry, T. and Sauer, T. (2017). Density estimation on manifolds with boundary. *Comput. Statist. Data Anal.*, 107:1–17.

Caggiano, G., Castelnuovo, E., Colombo, V., and Nodari, G. (2015). Estimating Fiscal Multipliers: News From A Non-linear World. *The Economic Journal*, 125(584):746–776.

Caggiano, G., Castelnuovo, E., and Figueres, J. M. (2017). Economic policy uncertainty and unemployment in the United States: A nonlinear approach. *Economics Letters*, 151:31–34.

Caggiano, G., Castelnuovo, E., and Pellegrino, G. (2021). Uncertainty shocks and the great recession: Nonlinearities matter. *Economics Letters*, 198:109669.

Chen, X. (2007). Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6, pages 5549–5632. Elsevier.

Chen, X. (2013). Penalized Sieve Estimation and Inference of Seminonparametric Dynamic Models: A Selective Review. In Acemoglu, D., Arellano, M., and Dekel, E., editors, *Advances in Economics and Econometrics*, pages 485–544. Cambridge University Press, 1 edition.

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.

Chen, X., Shao, Q.-M., Wu, W. B., and Xu, L. (2016). Self-normalized Cramér-type moderate deviations under dependence. *The Annals of Statistics*, 44(4):1593–1617.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691.

Debortoli, D., Forni, M., Gambetti, L., and Sala, L. (2020). Asymmetric Effects of Monetary Policy Easing and Tightening. *Working Paper*.

Fan, J. and Yao, Q. (2003). *Nonlinear time series: nonparametric and parametric methods*, volume 20. Springer.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213.

Feng, B. Q. (2003). Equivalence constants for certain matrix norms. *Linear Algebra and its Applications*, 374:247–253.

Forni, M., Gambetti, L., Maffei-Faccioli, N., and Sala, L. (2023a). Nonlinear transmission of financial shocks: Some new evidence. *Journal of Money, Credit and Banking*.

Forni, M., Gambetti, L., and Sala, L. (2023b). Asymmetric effects of news through uncertainty. *Macroeconomic Dynamics*, pages 1–25.

Gambetti, L., Maffei-Faccioli, N., and Zoi, S. (2022). Bad News, Good News: Coverage and Response Asymmetries. *Working Paper*.

Gonçalves, S., Herrera, A. M., Kilian, L., and Pesavento, E. (2021). Impulse response analysis for structural dynamic models with nonlinear regressors. *Journal of Econometrics*, 225(1):107–130.

Gourieroux, C. and Jasiak, J. (2005). Nonlinear Innovations and Impulse Responses with Application to VaR Sensitivity. *Annales d'Économie et de Statistique*, pages 1–31.

Gourieroux, C. and Lee, Q. (2023). Nonlinear impulse response functions and local projections. *Working Paper*.

Hamilton, J. D. (1994a). State-space models. *Handbook of Econometrics*, 4:3039–3080.

Hamilton, J. D. (1994b). *Time Series Analysis*. Princeton University Press.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, second edition.

Istrefi, K. and Mouabbi, S. (2018). Subjective interest rate uncertainty and the macroeconomy: A cross-country analysis. *Journal of International Money and Finance*, 88:296–313.

Jordà, Ò. (2005). Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review*, 95(1):161–182.

Kanazawa, N. (2020). Radial basis functions neural networks for nonlinear time series analysis and time-varying effects of supply shocks. *Journal of Macroeconomics*, 64:103210.

Kang, B. (2021). Inference In Nonparametric Series Estimation with Specification Searches for the Number of Series Terms. *Econometric Theory*, 37(2):311–345.

Karunamuni, R. J. and Alberts, T. (2005). On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191–212.

Kilian, L. and Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Themes in Modern Econometrics. Cambridge University Press, Cambridge.

Kilian, L. and Vigfusson, R. J. (2011). Are the responses of the us economy asymmetric in energy price increases and decreases? *Quantitative Economics*, 2(3):419–453.

Kohler, M. and Krzyżak, A. (2013). Optimal global rates of convergence for interpolation problems with random design. *Statistics & Probability Letters*, 83(8):1871–1879.

Kohler, M., Krzyżak, A., and Walk, H. (2006). Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *Journal of Multivariate Analysis*, 97(2):311–323.

Kohler, M., Krzyżak, A., and Walk, H. (2009). Optimal global rates of convergence for nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 139(4):1286–1296.

Koop, G., Pesaran, M. H., and Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics*, 74(1):119–147.

Li, J. and Liao, Z. (2020). Uniform nonparametric inference for time series. *Journal of Econometrics*, page 14.

Li, Q. and Racine, J. S. (2009). *Nonparametric econometric methods*. Emerald Group Publishing.

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

Malec, P. and Schienle, M. (2014). Nonparametric kernel density estimation near the boundary. *Comput. Statist. Data Anal.*, 72:57–76.

Mitrovic, D. and Zubrinic, D. (1997). *Fundamentals of Applied Functional Analysis*. CRC Press, Boca Raton, FL, USA.

Movahedifar, M. and Dickhaus, T. (2023). On the closed-loop Volterra method for analyzing time series. *Working Paper*.

Nowzohour, L. and Stracca, L. (2020). More than a feeling: Confidence, uncertainty, and macroeconomic fluctuations. *Journal of Economic Surveys*, 34(4):691–726.

Pellegrino, G. (2021). Uncertainty and monetary policy in the US: A journey into nonlinear territory. *Economic Inquiry*, 59(3):1106–1128.

Pötscher, B. M. and Prucha, I. (1997). *Dynamic nonlinear econometric models: Asymptotic theory*. Springer Science & Business Media.

Potter, S. M. (2000). Nonlinear impulse response functions. *Journal of Economic Dynamics and Control*, 24(10):1425–1446.

Ramey, V. A. and Zubairy, S. (2018). Government spending multipliers in good times and in bad: evidence from us historical data. *Journal of political economy*, 126(2):850–901.

Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. (2023). Asymptotic properties of neural network sieve estimators. *Journal of Nonparametric Statistics*, 0(0):1–30.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1):1–48.

Sirotko-Sibirskaya, N., Franz, M. O., and Dickhaus, T. (2020). Volterra bootstrap: Resampling higher-order statistics for strictly stationary univariate time series. *Working Paper*.

Stein, E. M. and Shakarchi, R. (2011). *Fourier analysis: an introduction*, volume 1. Princeton University Press.

Stock, J. H. and Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of Macroeconomics*, volume 2, pages 415–525. Elsevier.

Tenreyro, S. and Thwaites, G. (2016). Pushing on a string: US monetary policy is less powerful in recessions. *American Economic Journal: Macroeconomics*, 8(4):43–74.

Teräsvirta, T., Tjøstheim, D., and Granger, C. W. J. (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press.

Tropp, J. A. (2012). User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and its Interface*, 4(2):207–226.

Wu, W. B., Huang, Y., and Huang, Y. (2010). Kernel estimation for time series: An asymptotic theory. *Stochastic Processes and their Applications*, 120(12):2412–2431.

Zhou, X. (2022). Uniform convergence rates for wavelet curve estimation in sup-norm loss. *Journal of Computational and Applied Mathematics*, 400:113752.

Supplementary Material to "*Impulse Response Analysis of Structural Nonlinear Time Series Models* "

Giovanni Ballarin*
University of St. Gallen

December 1, 2024

A Dependence Conditions

A.1 Mixing

For the sake of completeness, we first recall the definition of β-mixing process: this is originally the dependence frameworks employed by Chen and Christensen (2015). Let $(\Omega, \mathcal{Q}, \mathbb{P})$ be the underlying probability space and define

$$\beta(\mathcal{A}, \mathcal{B}) := \frac{1}{2} \sup \sum_{(i,j)\in I\times J} |\mathbb{P}(A_i \cap B_i) - \mathbb{P}(A_i)\mathbb{P}(B_i)|$$

where $\mathcal{A}, \mathcal{B}$ are two σ-algebras, $\{A_i\}_{i\in I} \subset \mathcal{A}$, $\{B_j\}_{j\in J} \subset \mathcal{B}$ and the supremum is taken over all finite partitions of Ω. The h-th β-mixing coefficient of process $\{W_{2t}\}_{t\in\mathbb{Z}}$ is defined as

$$\beta(h) = \sup_t \beta\big(\sigma(\ldots, W_{2t-1}, W_{2t}), \ldots, \sigma(W_{2t+h}, W_{2t+h+1}, \ldots)\big),$$

and W_{2t} is said to be *geometric* or *exponential* β*-mixing* if $\beta(h) \leqslant \gamma_1 \exp(-\gamma_2 h)$ for some $\gamma_1 > 0$ and $\gamma_2 > 0$. An important consideration to be made regarding mixing assumptions is that they are, in general, hard to study. Especially in nonlinear systems, assuming that $\beta(h)$ decays exponentially over h imposes very high-level assumptions on the model. There are, however, a number of setups (linear and nonlinear) in which it is known that β-mixing holds under primitive assumptions (see Chen (2013) for examples and relevant references).

A.2 Model-Based Physical Dependence

Consider a *non-structural model* of the form

$$Z_t = G(Z_{t-1}, \epsilon_t). \tag{13}$$

This is a generalization of semi-reduced model (2) where linear and nonlinear components are absorbed into one functional term and B_0 is the identity matrix.[1] Indeed, note that

*E-mail: giovanni.ballarin@unisg.ch

[1] In this specific subsection, shock identification does not play a role and, as such, one can safely ignore B_0.

models of the form $Z_t = G(Z_{t-1}, \ldots, Z_{t-p}, \epsilon_t)$ can be rewritten as (13) using a companion formulation. If ϵ_t is stochastic, (13) defines a causal nonlinear stochastic process. More generally, it defines a nonlinear difference equation and an associated dynamical system driven by ϵ_t. Throughout this subsection, we shall assume that $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ as well as $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$.

Relying on the framework of Pötscher and Prucha (1997), we now introduce explicit conditions that allow to control dependence in nonlinear models by using the toolbox of physical dependence measures developed by Wu (2005, 2011). The aim is to use a dynamical system perspective to address the question of imposing meaningful assumptions on nonlinear dynamic models. This makes it possible to give more primitive conditions under which one can actually estimate (10) in a semiparametric way.

A.3 Stability

An important concept for dynamical system theory is that of stability. Stability turns out to play a key role in constructing valid asymptotic theory, as it is well understood in linear models. It is also fundamental in developing the approximation theory of nonlinear stochastic systems.

Example A.1. As a motivating example, first consider the linear system $Z_t = BZ_{t-1} + \epsilon_t$, where we may assume that $\{\epsilon_t\}_{t\in\mathbb{Z}}$, $\epsilon_t \in \mathbb{R}^{d_Z}$, is a sequence of i.i.d. innovations.[2] It is well-known that this system is stable if and only if the largest eigenvalue of B is strictly less than one in absolute value (Lütkepohl, 2005). For a higher order linear system, $Z_t = B(L)Z_{t-1} + \epsilon_t$ where $B(L) = B_1 + B_2L + \ldots + B_pL^{p-1}$, stability holds if and only if $|\lambda_{\max}(\mathbf{B})| < 1$ with $\mathbf{B}$ being the companion matrix associated with $B(L)$.

Extending the notion of stability from linear to nonlinear systems requires some care. Pötscher and Prucha (1997) derived generic conditions allowing to formally extend stability to nonlinear models by first analyzing *contractive* systems.

Definition A.1 (Contractive System)**.** *Let $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$, $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$, where $\{Z_t\}_{t\in\mathbb{Z}}$ is generated according to $Z_t = G(Z_{t-1}, \epsilon_t)$. The system is contractive if for all $(z, z') \in \mathcal{Z} \times \mathcal{Z}$ and $(e, e') \in \mathcal{E} \times \mathcal{E}$ the condition $\|G(z, \epsilon) - G(z', \epsilon')\| \leqslant C_Z\|z - z'\| + C_\epsilon\|e - e'\|$ holds with Lipschitz constants $0 \leqslant C_Z < 1$ and $0 \leqslant C_\epsilon < \infty$.*

Sufficient conditions to establish contractivity are

$$\sup \left\{ \left\| \operatorname{stack}_{i=1}^{d_Z} \left[\frac{\partial G}{\partial Z}(z^i, e^i) \right]_i \right\| \,\middle|\, z^i \in \mathcal{Z}, e^i \in \mathcal{E} \right\} < 1 \tag{14}$$

[2] One could alternatively think of the case of a deterministic input, setting $\epsilon_t \sim P_t(a_t)$, where $P_t(a_t)$ is a Dirac density on the deterministic sequence $\{a_t\}_{t\in\mathbb{Z}}$.

and

$$\left\| \frac{\partial G}{\partial \epsilon} \right\| < \infty, \tag{15}$$

where the stacking operator $\text{stack}_{i=1}^{d_Z}[\,\cdot\,]_i$ progressively stacks the rows, indexed by i, of its argument (which can be changing with i) into a matrix. Values $(z^i, e^i) \in \mathcal{Z} \times \mathcal{E}$ change with index i as the above condition is derived using the mean value theorem. Therefore it is necessary to consider a different set of values for each component of Z_t.

It is easy to see, as Pötscher and Prucha (1997) point out, that contractivity is often a too strong condition to be imposed. Indeed, even in the simple case of a scalar AR(2) model $Z_t = b_1 Z_{t-1} + b_2 Z_{t-2} + \epsilon_t$, regardless of the values of $b_1, b_2 \in \mathbb{R}$ contractivity is violated. This is due to the fact that in a linear AR(2) model studying contractivity reduces to checking $\|\mathbf{B}\| < 1$ instead of $|\lambda_{\max}(\mathbf{B})| < 1$, and the former is a stronger condition than the latter.[3] One can weaken contractivity – which must hold for G as a map from Z_{t-1} to Z_t – to the idea of *eventual contractivity*. That is, intuitively, one can impose conditions on the dependence of Z_{t+h} on Z_t for $h > 1$ sufficiently large. To do this formally, we first introduce the definition of system map iterates.

Definition A.2 (System Map Iterates)**.** *Let $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$, $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$, where $\{Z_t\}_{t\in\mathbb{Z}}$ is generated from a sequence $\{\epsilon_t\}_{t\in\mathbb{Z}}$ according to $Z_t = G(Z_{t-1}, \epsilon_t)$. The h-order system map iterate is defined to be*

$$\begin{aligned} G^{(h)}(Z_t, \epsilon_{t+1}, \epsilon_{t+2}, \ldots, \epsilon_{t+h}) &:= G(G(\cdots G(Z_t, \epsilon_{t+1}) \cdots, \epsilon_{t+h-1}), \epsilon_{t+h}) \\ &= G(\cdot, \epsilon_{t+h}) \circ G(\cdot, \epsilon_{t+h-1}) \circ \cdots \circ G(Z_t, \epsilon_{t+1}), \end{aligned}$$

where $\circ$ signifies function composition and $G^{(0)}(Z_t) = Z_t$.

To shorten notation, in place of $G^{(h)}(Z_t, \epsilon_{t+1}, \epsilon_{t+2}, \ldots, \epsilon_{t+h})$ we shall use $G^{(h)}(Z_t, \epsilon_{t+1:t+h})$. Additionally, for $1 \leqslant j \leqslant h$, the partial derivative $\partial G^{(h^*)}/\partial \epsilon_j$ for some fixed h^* is to be intended with respect to ϵ_{t+j}, the j-th entry of the input sequence. This derivative does not depend on the time index since by assumption G is time-invariant and so is $G^{(h)}$.

Taking again the linear autoregressive model as an example,

$$Z_{t+h} = G^{(h)}(Z_t, \epsilon_{t+1:t+h}) = B_1^h Z_t + \sum_{i=0}^{h-1} B_1^i \epsilon_{t+h-i}$$

since $G(z, \epsilon) = B_1 z + \epsilon$. If B_1 determines a stable system, then $\|B_1^h\| \to 0$ as $h \to \infty$ since G^h converges to zero, and therefore $\|B_1^h\| \leqslant C_Z < 1$ for h sufficiently large. It is thus possible to use system map iterates to define stability for higher-order nonlinear systems.

[3] See Pötscher and Prucha (1997), pp.68-69.

Definition A.3 (Stable System). *Let $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$, $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$, where $\{Z_t\}_{t\in\mathbb{Z}}$ is generated according to the system $Z_t = G(Z_{t-1}, \epsilon_t)$. The system is stable if there exists $h^* \geqslant 1$ such that for all $(z, z') \in \mathcal{Z} \times \mathcal{Z}$ and $(e_1, e_2, \dots e_{h^*}, e'_1, e'_2, \dots, e'_{h^*}) \in \bigtimes_{i=1}^{2h^*} \mathcal{E}$*

$$\|G^{(h^*)}(z, e_{1:h^*}) - G^{(h^*)}(z', e'_{1:h^*})\| \leqslant C_Z\|z - z'\| + C_\epsilon\|e_{1:h^*} - e'_{1:h^*}\|$$

holds with Lipschitz constants $0 \leqslant C_Z < 1$ and $0 \leqslant C_\epsilon < \infty$.

It is important to remember that this definition encompasses systems with an arbitrary finite autoregressive structure, i.e., $Z_t = G(Z_{t-p+1}, \dots, Z_{t-1}, \epsilon_t)$ for $p \geqslant 1$, thanks to the companion formulation of the process. An explicit stability condition, similar to that discussed above for contractivity, can be derived by means of the mean value theorem. Indeed, for a system to be stable it is sufficient that, at iterate h^*,

$$\sup \left\{ \left\| \operatorname{stack}_{i=1}^{d_Z} \left[\frac{\partial G^{(h^*)}}{\partial Z}(z^i, e^i_{1:h^*}) \right]_i \right\| \,\middle|\, z^i \in \mathcal{Z}, e^i_{1:h^*} \in \bigtimes_{i=1}^{h^*} \mathcal{E} \right\} < 1 \tag{16}$$

and

$$\sup \left\{ \left\| \frac{\partial G^{(h^*)}}{\partial \epsilon_j}(z, e_{1:h^*}) \right\| \,\middle|\, z \in \mathcal{Z}, e_{1:h^*} \in \bigtimes_{i=1}^{h^*} \mathcal{E} \right\} < \infty, \qquad j = 1, \dots, h^*. \tag{17}$$

Remark A.1. Pötscher and Prucha (1997) have used conditions (14)-(15) and (16)-(17) as basis for uniform laws of large numbers and central limit theorems for L^r-approximable and near epoch dependent processes.

A.4 Physical Dependence

Wu (2005) first proposed alternatives to mixing concepts by proposing dependence measures rooted in a dynamical system view of a stochastic process. Much work has been done to use such measures to derive approximation results and estimator properties, see for example Wu et al. (2010), Wu (2011), Chen et al. (2016), and references within.

Definition A.4. *Let $\{Z_t\}_{t\in\mathbb{Z}}$ be a process that can be written as $Z_{t+h} = G^{(h)}(Z_t, \epsilon_{t+1:t+h})$ for all $h \geqslant 1$, (nonlinear) maps $G^{(h)}$ and innovations $\{\epsilon_t\}_{t\in\mathbb{Z}}$. If for all $t \in \mathbb{Z}$ and chosen $r \geqslant 1$, Z_t has finite rth moment, the functional physical dependence measure Δ_r is*

$$\Delta_r(h) := \sup_t \left\| Z_{t+h} - G^{(h)}(Z'_t, \epsilon_{t+1:t+h}) \right\|_{L^r}$$

where $\{Z'_t\}_{t\in\mathbb{Z}}$ is an independent copy of $\{Z_t\}_{t\in\mathbb{Z}}$ based on innovation process $\{\epsilon'_t\}_{t\in\mathbb{Z}}$, itself an independent copy of $\{\epsilon_t\}_{t\in\mathbb{Z}}$.

Chen et al. (2016), among others, show how one may replace the geometric β-mixing

assumption with a physical dependence assumption.[4] We will consider the setting where models that have dependence – as measured by $\Delta_r(h)$ – which decays exponentially with h.

Definition A.5 (Geometric Moment Contracting Process)**.** *$\{Z_t\}_{t\in\mathbb{Z}}$ is geometric moment contracting (GMC) in L^r norm if there exists $a_1 > 0$, $a_2 > 0$ and $\tau \in (0, 1]$ such that*

$$\Delta_r(h) \leqslant a_1 \exp(-a_2\, h^\tau).$$

GMC conditions can be considered more general than β-mixing, as they encompass well-known counterexamples, e.g., the known counterexample provided by $Z_t = (Z_{t-1} + \epsilon_t)/2$ for ϵ_t i.i.d. Bernoulli r.v.s (Chen et al., 2016).

In the following proposition we prove that if contractivity or stability conditions as defined by Pötscher and Prucha (1997) hold for G and $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence, then the process $\{Z_t\}_{t\in\mathbb{Z}}$ is GMC – according to Definition A.5 – under weak moment assumptions.

Proposition A.1. *Assume that $\{\epsilon_t\}_{t\in\mathbb{Z}}$, $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$ are i.i.d. and $\{Z_t\}_{t\in\mathbb{Z}}$ is generated according to $Z_t = G(Z_{t-1}, \epsilon_t)$, where $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ and G is a measurable function.*

(a) If contractivity conditions (14)-(15) hold, $\sup_{t\in\mathbb{Z}}\|\epsilon_t\|_{L^r} < \infty$ for $r \geqslant 2$ and $\|G(\overline{z}, \overline{\epsilon})\| < \infty$ for some $(\overline{z}, \overline{\epsilon}) \in \mathcal{Z} \times \mathcal{E}$, then $\{Z_t\}_{t\in\mathbb{Z}}$ is GMC with $\Delta_r(k) \leqslant a \exp(-\gamma h)$ where $\gamma = -\log(C_Z)$ and $a = 2\|Z_t\|_{L^r} < \infty$.

(b) If stability conditions (16)-(17) hold, $\sup_{t\in\mathbb{Z}}\|\epsilon_t\|_{L^r} < \infty$ for $r \geqslant 2$ and $\|\partial G/\partial Z\| \leqslant M_Z < \infty$, then $\{Z_t\}_{t\in\mathbb{Z}}$ is GMC with $\Delta_r(k) \leqslant \bar{a} \exp(-\gamma_{h^*} h)$ where $\gamma_{h^*} = -\log(C_Z)/h^*$ and $\bar{a} = 2\|Z_t\|_{L^r} \max\{M_Z^{h-1}, 1\}/C_Z < \infty$.

Proposition A.1 is important in that it links the GMC property to transparent conditions on the structure of the nonlinear model. It also allows to handle multivariate systems, while previous work has focused on scalar systems (c.f. Wu (2011) and Chen et al. (2016)).

Lastly, the following lemma shows that if $\{W_{2t}\}_{t\in\mathbb{Z}}$ is geometric moment contracting, Assumption 9 is fulfilled.[5]

Lemma A.1. *If Assumption 7(iii) holds and $\{W_{2t}\}_{t\in\mathbb{Z}}$ is strictly stationary and GMC then one may choose an integer sequence $q = q(n) \leqslant n/2$ with $(n/q)^{r+1} q K^\rho \Delta_r(q) = o(1)$ for $\rho = 5/2 - (r/2 + 2/r) + \omega_2$ and $r > 2$ such that*

$$\|(\widetilde{B}_\pi' \widetilde{B}_\pi/n) - I_K\| = O_P\left(\zeta_{K,n}\lambda_{K,n}\sqrt{\frac{q \log K}{n}}\right) = o_P(1)$$

[4] We adapt here the definitions of Chen et al. (2016) to work with a system of the form $Z_t = G(Z_{t-1}, \epsilon_t)$.

[5] Compare also with Lemma 2.2 in Chen and Christensen (2015).

provided $\zeta_{K,n}\lambda_{K,n}\sqrt{(q\log K)/n} = o(1)$.

It can be seen that Lemma A.1 holds by setting $\sqrt{K(\log(n))^2/n} = o(1)$ and choosing $q(n) = \gamma^{-1}\log(K^{\rho}n^{r+1})$, where γ is a GMC factor, see Proposition A.1 in the Online Appendix for details. Therefore, the rate is the same as the one derived by Chen and Christensen (2015) for exponentially β-mixing regressors. It is straightforward to prove that, if one assume $\{Z_t\}_{t\in\mathbb{Z}}$ fulfills GMC conditions, then $\{W_{2t}\}_{t\in\mathbb{Z}}$ is also a geometric moment contracting process, see Remark B.1 below. Accordingly, Lemma A.1 applies and Assumption 9 is automatically verified.

B Proofs

Matrix Norms Let

$$\|A\|_r := \max\left\{\|Ax\|_r \,\middle|\, \|x\|_r \leqslant 1\right\}$$

be the r-operator norm of matrix $A \in \mathbb{C}^{d_1\times d_2}$. The following Theorem establishes the equivalence between different operator norms as well as the compatibility constants.

Theorem B.1 (Feng (2003))**.** *Let* $1 \leqslant p, q \leqslant \infty$. *Then for all* $A \in \mathbb{C}^{d_1\times d_2}$,

$$\|A\|_p \leqslant \lambda_{p,q}(d_1)\lambda_{q,p}(d_2)\|A\|_q \quad \text{where} \quad \lambda_{a,b}(d) := \begin{cases} 1 & \text{if } a \geqslant b, \\ d^{1/a-1/b} & \text{if } a < b. \end{cases}$$

This norm inequality is sharp.

In particular, if $p > q$ then it holds $(d_2)^{-(1/q-1/p)}\|A\|_p \leqslant \|A\|_q \leqslant (d_1)^{1/q-1/p}\|A\|_p$.

B.1 GMC Conditions and Proposition A.1

Lemma B.1. *Assume that* $\{\epsilon_t\}_{t\in\mathbb{Z}}$, $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$ *are i.i.d., and* $\{Z_t\}_{t\in\mathbb{Z}}$ *is generated according to* $Z_t = G(Z_{t-1}, \epsilon_t)$, *where* $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ *and* G *is a measurable function. If either*

(a) Contractivity conditions (14)-(15) hold, $\sup_{t\in\mathbb{Z}}\|\epsilon_t\|_{L^r} < \infty$ *and* $\|G(\bar{z}, \bar{\epsilon})\| < \infty$ *for some* $(\bar{z}, \bar{\epsilon}) \in \mathcal{Z}\times\mathcal{E}$;

(b) Stability conditions (16)-(17) hold, $\sup_{t\in\mathbb{Z}}\|\epsilon_t\|_{L^r} < \infty$ *and* $\|\partial G/\partial Z\| \leqslant M_Z < \infty$;

then $\sup_t\|Z_t\|_{L^r} < \infty$ *w.p.*1.

Proof.

(a) In a first step, we show that, given event $\omega \in \Omega$, realization $Z_t(\omega)$ is unique with probability one. To do this, introduce initial condition $z_\circ$ for $\ell > 1$ such that $z_\circ \in \mathcal{Z}$ and $\|z_\circ\| < \infty$. Define $Z_t^{(-\ell)}(\omega) = G^{(\ell)}(y_\circ, \epsilon_{t-\ell+1:t}(\omega))$. Further, let $Z_t'^{(-\ell)}$ be the realization with initial condition $z_\circ' \neq z_\circ$ and innovation realizations $\epsilon_{t-\ell+1:t}(\omega)$. Note that $\|Z_t^{(-\ell)}(\omega) - Z_t'^{(-\ell)}(\omega)\| \leqslant C_Z^\ell \|z_\circ - z_\circ'\|$, which goes to zero as $\ell \to \infty$. Therefore, if we set $Z_t(\omega) := \lim_{\ell\to\infty} Z_t^{(-\ell)}(\omega)$, $Z_t(\omega)$ is unique with respect to the choice of $z_\circ$ w.p.1. A similar recursion shows that

$$\left\|Z_t^{(-\ell)}(\omega)\right\| \leqslant C_Z^\ell \|z_\circ\| + \sum_{k=0}^{\ell-1} C_Z^k C_\epsilon \|\epsilon_{t-k}(\omega)\| .$$

By norm equivalence, this implies

$$\left\|Z_t^{(-\ell)}\right\|_{L^r} \leqslant C_Z^\ell \|z_\circ\|_r + \sum_{k=0}^{\ell-1} C_Z^k C_\epsilon \|\epsilon_{t-k}\|_{L^r} \leqslant C_Z^\ell \|z_\circ\|_r + \frac{C_\epsilon}{1-C_Z} \sup_{t\in\mathbb{Z}} \|\epsilon_t\|_{L^r} < \infty,$$

and taking the limit $\ell \to \infty$ proves the claim.

(b) Consider again distinct initial conditions $z_\circ' \neq z_\circ$ and innovation realizations $\epsilon_{t-\ell+1:t}(\omega)$, yielding $Z_t'^{(-\ell)}(\omega)$ and $Z_t^{(-\ell)}(\omega)$, respectively. We may use the contraction bound derived in the proof of Proposition A.1 (b) below, that is, $\| Z_t^{(-\ell)}(\omega) - Z_t'^{(-\ell)}(\omega) \|_r \leqslant C_Z^\ell C_2 \|z_\circ - z_\circ'\|_r$, where $C_2 > 0$ is a constant. With trivial adjustments, the uniqueness and limit arguments used for (a) above apply here too.

□

Proof of Proposition A.1.

(a) By assumption for all $(z, z') \in \mathcal{Z} \times \mathcal{Z}$ and $(e, e') \in \mathcal{E} \times \mathcal{E}$ it holds that $\|G(z, \epsilon) - G(z', \epsilon')\| \leqslant C_Z\|z - z'\| + C_\epsilon\|e - e'\|$, where $0 \leqslant C_Z < 1$ and $0 \leqslant C_\epsilon < \infty$. The equivalence of norms directly generalizes this inequality to any r-norm for $r > 2$. We study $\|Z_{t+h} - Z'_{t+h}\|_r$ where Z'_{t+h} is constructed with a time-t perturbation of the history of Z_{t+h}. Therefore, for any given t and $h \leqslant 1$ it holds that

$$\begin{aligned}\left\| Z_{t+h} - G^{(h)}(Z'_t, \epsilon_{t+1:t+h}) \right\|_r &\leqslant C_Z \|G^{(h-1)}(Z_t, \epsilon_{t+1:t+h-1}) - G^{(h-1)}(Z'_t, \epsilon_{t+1:t+h-1})\|_r \\ &\leqslant C_Z^h \|Z_t - Z'_t\|_r,\end{aligned}$$

since sequence $\epsilon_{t+1:t+h}$ is common between Z_{t+h} and Z'_{t+h}. Clearly then

$$\left\| Z_{t+h} - G^{(h)}(Z'_t, \epsilon_{t+1:t+h}) \right\|_r \leqslant 2\|Z_t\|_r \exp(-\gamma h)$$

for $\gamma = -\log(C_Z)$. Letting $a = 2\|Z_t\|_r$ and shifting time index t backward by h, since $\sup_t \|Z_t\|_{L^r} < \infty$ w.p.1 from Lemma B.1 the result for L^r follows with $\tau = 1$.

(b) Proceed similar to (a), but notice that now we must handle cases of steps $1 \leqslant h < h^*$. Consider iterate $h^* + 1$, for which

$$\begin{aligned}
&\left\| Z_{t+h+1} - G^{(h+1)}(Z'_t, \epsilon_{t+1:t+h+1}) \right\|_r \\
&\qquad \leqslant C_Z \|G^{(h)}(G(Z_t, \epsilon_{t+1}), \epsilon_{t+2:t+h}) - G^{(h)}(G(Z'_t, \epsilon_{t+1}), \epsilon_{t+2:t+h})\|_r \\
&\qquad \leqslant C_Z^h \|G(Z_t, \epsilon_{t+1}) - G(Z'_t, \epsilon_{t+1})\|_r \\
&\qquad \leqslant C_Z^h M_Z \|Z_t - Z'_t\|_r
\end{aligned}$$

by the mean value theorem. Here we may assume that $M_Z \geqslant 1$ otherwise we would fall under case (a), so that $M_Z \leqslant M_Z^2 \leqslant \ldots \leqslant M_Z^{h^*-1}$. More generally,

$$\left\| Z_{t+h+1} - G^{(h+1)}(Z'_t, \epsilon_{t+1:t+h+1}) \right\|_r \leqslant C_Z^{j(h)} \max\{M_Z^{h^*-1}, 1\} \|Z_t - Z'_t\|_r$$

for $j(h) := \lfloor h/h^* \rfloor$. Result (b) then follows by noting that $j(h) \geqslant h/h^* - 1$ and then proceeding as in (a) to derive GMC coefficients.

□

Remark B.1. The assumption of GMC for a process translates naturally to vectors that are composed of stacked lags of realizations. This, for example, is important in the discussion of Section 3, since one needs that regressors $\{W_{2t}\}_{t\in\mathbb{Z}}$ be geometric moment contracting.

Recall that $W_{2t} = (X_t, X_{t-1}, \ldots, X_{t-p}, Y_{t-1}, \ldots, Y_{t-p}, \epsilon_{1t})$. Here we shall reorder this vector slightly to actually be $W_{2t} = (X_t, X_{t-1}, Y_{t-1}, \ldots, X_{t-p}, Y_{t-p}, \epsilon_{1t})$. For $h > 0$ and $1 \leqslant l \leqslant h$, let $Z'_{t+j} := \Phi^{(l)}(Z'_t, \ldots, Z'_{t-p}; \epsilon_{t+1:t+j})$ be the a perturbed version of Z_t, where $Z'_t, \ldots, Z'_{t-p}$ are taken from an independent copy of $\{Z_t\}_{t\in\mathbb{Z}}$. Define $W'_{2t} = (X'_t, X'_{t-1}, Y'_{t-1}, \ldots, X'_{t-p}, Y'_{t-p}, \epsilon_{1t})$. Using Minkowski's inequality

$$\begin{aligned}
\|W_{2t+h} - W'_{2t+h}\|_{L^r} &\leqslant \|X_{t+h} - X'_{t+h}\|_{L^r} + \sum_{j=1}^{p} \|Z_{t+h-j} - Z'_{t+h-j}\|_{L^r} \\
&\leqslant \sum_{j=0}^{p} \|Z_{t+h-j} - Z'_{t+h-j}\|_{L^r},
\end{aligned}$$

thus, since $p > 0$ is fixed finite,

$$\sup_t \|W_{2t+h} - W'_{2t+h}\|_{L^r} \leqslant \sum_{j=0}^{p} \Delta_r(h-j) \leqslant (p+1)\, a_{1Z} \exp(-a_{2Z} h).$$

Above, a_{1Z} and a_{2Z} are the GMC coefficients of $\{Z_t\}_{t\in\mathbb{Z}}$.

B.2 Lemma A.1 and Matrix Inequalities under Dependence

In order to prove Lemma A.1, we modify the approach of Chen and Christensen (2015), which relies on Berbee's Lemma and an interlaced coupling, to handle variables with physical dependence. This is somewhat similar to the proof strategies used in Chen et al. (2016).

First of all, we recall below a Bernstein-type inequality for independent random matrices of Tropp (2012).

Theorem B.2. *Let $\{\Xi_i\}_{i=1}^n$ be a finite sequence of independent random matrices with dimensions $d_1 \times d_2$. Assume $\mathbb{E}[\Xi_i] = 0$ for each i and $\max_{1\leqslant i\leqslant n}\|\Xi_i\| \leqslant R_n$ and define*

$$\varsigma_n^2 := \max\left\{\left\|\sum_{i=1}^n \mathbb{E}\left[\Xi_{i,n}\Xi_{j,n}'\right]\right\|, \left\|\sum_{i=1}^n \mathbb{E}\left[\Xi_{i,n}'\Xi_{j,n}\right]\right\|\right\}.$$

Then for all $z \geqslant 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_i\right\| \geqslant z\right) \leqslant (d_1+d_2)\exp\left(\frac{-z^2/2}{nq\varsigma_n^2 + qR_n z/3}\right).$$

The main exponential matrix inequality due to Chen and Christensen (2015), Theorem 4.2 is as follows.

Theorem B.3. *Let $\{X_i\}_{i\in\mathbb{Z}}$ where $X_i \in \mathcal{X}$ be a β-mixing sequence and let $\Xi_{i,n} = \Xi_n(X_i)$ for each i where $\Xi_n : \mathcal{X} \mapsto \mathbb{R}^{d_1\times d_2}$ be a sequence of measurable $d_1 \times d_2$ matrix-valued functions. Assume that $\mathbb{E}[\Xi_{i,n}] = 0$ and $\|\Xi_{i,n}\| \leqslant R_n$ for each i and define*

$$S_n^2 := \max\left\{\mathbb{E}\left[\|\Xi_{i,n}\Xi_{j,n}'\|\right], \mathbb{E}\left[\|\Xi_{i,n}'\Xi_{j,n}\|\right]\right\}.$$

Let $1 \leqslant q \leqslant n/2$ be an integer and let $I_\bullet = q\lfloor n/q\rfloor, \ldots, n$ when $q\lfloor n/q\rfloor < n$ and $I_\bullet = \varnothing$ otherwise. Then, for all $z \geqslant 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_{i,n}\right\| \geqslant 6z\right) \leqslant \frac{n}{q}\beta(q) + \mathbb{P}\left(\left\|\sum_{i\in I_\bullet} \Xi_{i,n}\right\| \geqslant z\right) + 2(d_1+d_2)\exp\left(\frac{-z^2/2}{nqS_n^2 + qR_n z/3}\right),$$

where $\|\sum_{i\in I_\bullet}\Xi_{i,n}\| := 0$ whenever $I_\bullet = \varnothing$.

To fully extend Theorem B.3 to physical dependence, we will proceed in steps. First, we derive a similar matrix inequality by directly assuming that random matrices $\Xi_{i,n}$ have physical dependence coefficient $\Delta_r^{\Xi}(h)$. In the derivations we will use that

$$\frac{1}{(d_2)^{1/2-1/r}}\|A\|_r \leqslant \|A\|_2 \leqslant (d_1)^{1/2-1/r}\|A\|_r.$$

for $r \geqslant 2$.

Theorem B.4. *Let $\{\epsilon_j\}_{j\in\mathbb{Z}}$ be a sequence of i.i.d. variables and let $\{\Xi_{i,n}\}_{i=1}^n$,*

$$\Xi_{i,n} = G_n^{\Xi}(\ldots, \epsilon_{i-1}, \epsilon_i)$$

for each i, where $\Xi_n : \mathcal{X} \mapsto \mathbb{R}^{d_1\times d_2}$, be a sequence of measurable $d_1 \times d_2$ matrix-valued functions. Assume that $\mathbb{E}[\Xi_{i,n}] = 0$ and $\|\Xi_{i,n}\| \leqslant R_n$ for each i and define

$$S_n^2 := \max\left\{\mathbb{E}\left[\|\Xi_{i,n}\Xi_{j,n}'\|\right], \mathbb{E}\left[\|\Xi_{i,n}'\Xi_{j,n}\|\right]\right\}.$$

Additionally assume that $\|\Xi_{i,n}\|_{L^r} < \infty$ for $r > 2$ and define the matrix physical dependence measure $\Delta_r^{\Xi}(h)$ as

$$\Delta_r^{\Xi}(h) := \max_{1\leqslant i\leqslant n} \left\| \Xi_{i,n} - \Xi_{i,n}^{h*} \right\|_{L^r},$$

where $\Xi_{i,n}^{h*} := G_n^{\Xi}(\ldots, \epsilon_{i-h-1}^*, \epsilon_{i-h}^*, \epsilon_{i-h+1}, \ldots, \epsilon_{i-1}, \epsilon_i)$ for independent copy $\{\epsilon_j^*\}_{j\in\mathbb{Z}}$. Let $1 \leqslant q \leqslant n/2$ be an integer and let $I_\bullet = q\lfloor n/q\rfloor, \ldots, n$ when $q\lfloor n/q\rfloor < n$ and $I_\bullet = \varnothing$ otherwise. Then, for all $z \geqslant 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_{i,n}\right\| \geqslant 6z\right) \leqslant \frac{n^{r+1}}{q^r (d_2)^{r/2-1} z^r}\Delta_r^{\Xi}(q) + \mathbb{P}\left(\left\|\sum_{i\in I_\bullet}\Xi_{i,n}\right\| \geqslant z\right) + 2(d_1+d_2)\exp\left(\frac{-z^2/2}{nqS_n^2 + qR_n z/3}\right),$$

where $\|\sum_{i\in I_\bullet}\Xi_{i,n}\| := 0$ whenever $I_\bullet = \varnothing$.

Proof. To control dependence, we can adapt the interlacing block approach outlined by Chen et al. (2016). To interlace the sum, split it into

$$\sum_{i=1}^n \Xi_{i,n} = \sum_{j\in K_e} J_k + \sum_{j\in J_o} W_k + \sum_{i\in I_\bullet}\Xi_{i,n},$$

where $W_j := \sum_{i=q(j-1)+1}^{qj}\Xi_{i,n}$ for $j = 1, \ldots, \lfloor n/q\rfloor$ are the blocks, $I_\bullet := \{q\lfloor n/q\rfloor + 1, \ldots, n\}$ if $q\lfloor n/q\rfloor < n$ and J_e and J_o are the subsets of even and odd numbers of $\{1, \ldots, \lfloor n/q\rfloor\}$, respectively. For simplicity define $J = J_e \cup J_o$ as the set of block indices and let

$$W_j^\dagger := \mathbb{E}\big[\, W_j \,|\, \epsilon_\ell,\ q(j-2)+1 \leqslant \ell \leqslant qj \,\big].$$

Note that by construction $\{W_j^\dagger\}_{j\in J_e}$ are independent and also $\{W_j^\dagger\}_{j\in J_o}$ are independent. Using the triangle inequality we find

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_{i,n}\right\| \geqslant 6z\right) \leqslant \mathbb{P}\left(\left\|\sum_{j\in J}(W_j - W_j^\dagger)\right\| + \left\|\sum_{j\in J} W_j^\dagger\right\| + \left\|\sum_{i\in I_\bullet}\Xi_{i,n}\right\| \geqslant 6z\right)$$

$$
\begin{aligned}
&\leqslant \mathbb{P}\left(\left\|\sum_{j\in J}(W_j - W_j^\dagger)\right\| \geqslant z\right) + \mathbb{P}\left(\left\|\sum_{j\in J_e} W_j^\dagger\right\| \geqslant z\right) \\
&\quad + \mathbb{P}\left(\left\|\sum_{j\in J_o} W_j^\dagger\right\| \geqslant z\right) + \mathbb{P}\left(\left\|\sum_{i\in I_\bullet} \Xi_{i,n}\right\| \geqslant z\right) \\
&= I + II + III + IV.
\end{aligned}
$$

We keep term IV as is. As in the proof of Chen and Christensen (2015), terms II and III consist of sums of independent matrices, where each $W_j^\dagger$ satisfies $\|W_j^\dagger\| \leqslant qR_n$ and

$$
\max\left\{\mathbb{E}\left[\|W_j^\dagger W_j^{\dagger\prime}\|\right], \mathbb{E}\left[\|W_j^{\dagger\prime} W_j^\dagger\|\right]\right\} \leqslant qS_n^2.
$$

Then, using the exponential matrix inequality of Tropp (2012),

$$
\mathbb{P}\left(\left\|\sum_{j\in J_e} W_k^\dagger\right\| \geqslant z\right) \leqslant (d_1 + d_2)\exp\left(\frac{-z^2/2}{nqS_n^2 + qR_n z/3}\right).
$$

The same holds for the sum over J_o. Finally, we use the physical dependence measure Δ_r^Ξ to bound I. Start with the union bound to find

$$
\begin{aligned}
\mathbb{P}\left(\left\|\sum_{j\in J}(W_j - W_j^\dagger)\right\| \geqslant z\right) &\leqslant \mathbb{P}\left(\sum_{j\in J}\left\|W_j - W_j^\dagger\right\| \geqslant z\right) \\
&\leqslant \frac{n}{q}\,\mathbb{P}\left(\left\|W_j - W_j^\dagger\right\| \geqslant \frac{q}{n}\,z\right),
\end{aligned}
$$

where we have used that $\lfloor n/q \rfloor \leqslant n/q$. Since W_j and $W_j^\dagger$ differ only over a σ-algebra that is q steps in the past, by assumption

$$
\left\|W_j - W_j^\dagger\right\|_{L^r} \leqslant q\,\Delta_r^\Xi(q),
$$

which implies, by means of the rth moment inequality,

$$
\mathbb{P}\left(\left\|W_j - W_j^\dagger\right\| \geqslant \frac{q}{n}\,z\right) \leqslant \mathbb{P}\left((d_2)^{1/r-1/2}\left\|W_j - W_j^\dagger\right\|_r \geqslant \frac{q}{n}\,z\right) \leqslant \frac{n^r}{q^{r-1}(d_2)^{r/2-1}z^r}\,\Delta_r^\Xi(q).
$$

where $(d_2)^{1/r-1/2}$ is the operator norm equivalence constant such that $\|\cdot\| \geqslant (d_2)^{1/r-1/2}\|\cdot\|_r$ (Feng, 2003). Therefore,

$$
\mathbb{P}\left(\left\|\sum_{j\in J}(W_j - W_j^\dagger)\right\| \geqslant z\right) \leqslant \frac{n^{r+1}}{q^r(d_2)^{r/2-1}z^r}\,\Delta_r^\Xi(q)
$$

as claimed. □

Notice that the first term in the bound is weaker than that derived by Chen and

Christensen (2015). The β-mixing assumption and Berbee's Lemma give strong control over the probability $\mathbb{P}(\|\sum_{j\in J}(W_j - W_j^\dagger)\| \geqslant z)$. In contrast, assuming physical dependence means we have to explicitly handle a moment condition. One might think of sharpening Theorem B.4 by sidestepping the rth moment inequality (c.f. avoiding Chebyshev's inequality in concentration results), but we do not explore this approach here.

The second step is to map the physical dependence of a generic vector time series $\{X_i\}_{i\in\mathbb{Z}}$ to matrix functions.

Proposition B.1. *Let $\{X_i\}_{i\in\mathbb{Z}}$ where $X_i = G(\ldots, \epsilon_{i-1}, \epsilon_i) \in \mathcal{X}$ for $\{\epsilon_j\}_{j\in\mathbb{Z}}$ i.i.d. be a sequence with finite rth moment, where $r > 0$, and functional physical dependence coefficients*

$$\Delta_r(h) = \sup_i \left\| X_{i+h} - G^{(h)}(X_i^*, \epsilon_{i+1:i+h}) \right\|_{L^r}$$

for $h \geqslant 1$. Let $\Xi_{i,n} = \Xi_n(X_i)$ for each i where $\Xi_n : \mathcal{X} \mapsto \mathbb{R}^{d_1\times d_2}$ be a sequence of measurable $d_1 \times d_2$ matrix-valued functions such that $\Xi_n = (v_1, \ldots, v_{d_2})$ for $v_\ell \in \mathbb{R}^{d_1}$. If $\|\Xi_{i,n}\|_{L^r} < \infty$ and

$$C_{\Xi,\ell} := \sup_{x\in\mathcal{X}} \|\nabla v_\ell(x)\| \leqslant C_\Xi < \infty,$$

then matrices $\Xi_{i,n}$ have physical dependence coefficients

$$\Delta_r^\Xi(h) = \sup_i \left\| \Xi_{i,n} - \Xi_{i,n}^{h*} \right\|_{L^r} \leqslant \sqrt{d_1} \left(\frac{d_2}{d_1}\right)^{1/r} C_\Xi \, \Delta_r(h),$$

where $\Xi_{i,n}^{h*} = \Xi_n(G^{(h)}(X_i', \epsilon_{i+1:i+h}))$.

Proof. To derive the bound, we use $\Xi_n(X_i)$ and $\Xi_n(X_i^{h*})$ in place of $\Xi_{i,n}$ and $\Xi_{i,n}^{h*}$, respectively, where $X_i^{h*} = G^{(h)}(X_i^*, \epsilon_{i+1:i+h})$. First we move from studying the operator r-norm (recall, $r > 2$) to the Frobenius norm,

$$\left\|\Xi_n(X_i) - \Xi_n(X_i^{h*})\right\|_r \leqslant (d_2)^{1/2-1/r} \left\|\Xi_n(X_i) - \Xi_n(X_i^{h*})\right\|_F.$$

where as intermediate step we use the 2-norm. Let $\Xi_n = (v_1, \ldots, v_{d_2})$ for $v_\ell \in \mathbb{R}^{d_1}$ and $\ell \in 1, \ldots, d_2$, so that

$$\|\Xi_n\|_F = \sqrt{\sum_{\ell=1}^{d_2} \|v_\ell\|^2}$$

where $v_\ell = (v_{\ell 1}, \ldots, v_{\ell d_1})'$. Since $v_\ell : \mathcal{X} \mapsto \mathbb{R}^{d_1}$ are vector functions, the mean value theorem gives that

$$\left\|\Xi_n(X_i) - \Xi_n(X_i^{h*})\right\|_F \leqslant \sqrt{\sum_{\ell=1}^{d_2} C_{\Xi,\ell}^2 \, \|X_i - X_i^{h*}\|^2} \leqslant \sqrt{d_2}\, C_\Xi \, \|X_i - X_i^{h*}\|.$$

Combining results and moving from the vector r-norm to the 2-norm yields

$$\left\|\Xi_n(X_i) - \Xi_n(X_i^{h*})\right\|_r \leqslant (d_2)^{1-1/r}(d_1)^{1/2-1/r}\, C_\Xi \, \|X_i - X_i^{h*}\|_r.$$

The claim involving the L^r norm follows immediately. □

The following Corollary, which specifically handles matrix functions defined as outer products of vector functions, is immediate and covers the setups of series estimation.

Corollary B.1. *Under the conditions of Proposition B.1, if*

$$\Xi_n(X_i) = \xi_n(X_i)\xi_n(X_i)' + Q_n$$

where $\xi_n : \mathcal{X} \mapsto \mathbb{R}^d$ is a vector function and $Q_n \in \mathbb{R}^{d\times d}$ is nonrandom matrix, then

$$\Delta_r^{\Xi}(h) \leqslant d^{3/2-2/r}\, C_\xi \, \Delta_r(h),$$

where $C_\xi := \sup_{x\in\mathcal{X}}\|\nabla\xi_n(x)\| < \infty$.

Proof. Matrix Q_n cancels out since it is nonrandom and appears in both $\Xi_n(X_i)$ and $\Xi_n(X_i^{h*})$. Since $\Xi_n(X_i)$ is square, the ratio of row to column dimensions simplifies. □

The following Corollaries to Theorem B.4 can now be derived in a straightforward manner.

Corollary B.2. *Under the conditions of Theorem B.4 and Proposition B.1, for all $z \geqslant 0$*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_{i,n}\right\| \geqslant 6z\right) \leqslant \frac{n^{r+1}}{q^r z^r}(d_2)^{2-(r/2+1/r)}(d_1)^{1/2-1/r} C_\Xi \, \Delta_r(q) + \mathbb{P}\left(\left\|\sum_{i\in I_\bullet} \Xi_{i,n}\right\| \geqslant z\right) + 2(d_1+d_2)\exp\left(\frac{-z^2/2}{nqS_n^2 + qR_n z/3}\right).$$

where $\Delta_r(\cdot)$ if the functional physical dependence coefficient of X_i.

Corollary B.3. *Under the conditions of Theorem B.4 and Proposition B.1, if $q = q(n)$ is chosen such that*

$$\frac{n^{r+1}}{q^r}(d_2)^{2-(r/2+1/r)}(d_1)^{1/2-1/r} C_\Xi \, \Delta_r(q) = o(1)$$

and $R_n\sqrt{q\log(d_1+d_2)} = o(S_n\sqrt{n})$, then

$$\left\|\sum_{i=1}^n \Xi_{i,n}\right\| = O_P\left(S_n\sqrt{nq\log(d_1+d_2)}\right).$$

This result is almost identical to Corollary 4.2 in Chen and Christensen (2015), with the only adaptation of using Theorem B.4 as a starting point. Condition $R_n\sqrt{q\log(d_1+d_2)} = o(S_n\sqrt{n})$ is simple to verify by assuming, e.g., $q = o(n/\log(n))$ since $\log(d_1+d_2) \lesssim \log(K)$ and $K = o(n)$.

Note that when $d_1 = d_2 \equiv K$, which is the case of interest in the series regression setup, the first condition in Corollary B.3 reduces to

$$K^{5/2-(r/2+2/r)}\, C_\Xi\, \Delta_r(q) = o(1),$$

which also agrees with the rate of Corollary B.1. Assumption 7(i) and a compact domain further allow to explicitly bound factor C_Ξ by

$$C_\Xi \lesssim K^{\omega_2},$$

so that the required rate becomes

$$K^\rho\, \Delta_r(q) = o(1), \quad \text{where} \quad \rho := \frac{3}{2} - \frac{r}{2} + \omega_2.$$

Proof of Lemma A.1. The proof follows from Corollary B.3 by the same steps of the proof of Lemma 2.2 in Chen and Christensen (2015). Simply take

$$\Xi_{i,n} = n^{-1}(\widetilde{(}b)_\pi^K(X_i)\widetilde{(}b)_\pi^K(X_i)' - I_K)$$

and note that $R_n \leqslant n^{-1}(1+\zeta_{K,n}^2\lambda_{K,n}^2)$ and $S_n \leqslant n^{-2}(1+\zeta_{K,n}^2\lambda_{K,n}^2)$. □

For Lemma A.1 to hold under GMC assumptions a valid choice for $q(n)$ is

$$q(n) = \gamma^{-1}\log(K^\rho n^{r+1})$$

where γ as in Proposition A.1. This is due to

$$\begin{aligned}\left(\frac{n}{q}\right)^{r+1} qK^\rho\Delta_r(q) &\lesssim \frac{n^{r+1}}{q^r}K^\rho\exp(-\gamma q)\\ &\lesssim \frac{n^{r+1}K^\rho}{\log(K^\rho n^{r+1})^r}(K^\rho n^{r+1})^{-1}\\ &= \frac{1}{\log(K^\rho n^{r+1})^r} = o(1).\end{aligned}$$

Note then that, if $\lambda_{K,n} \lesssim 1$ and $\zeta_{K,n} \lesssim \sqrt{K}$, since

$$\zeta_{K,n}\lambda_{K,n}\sqrt{\frac{q\log K}{n}} \lesssim \sqrt{\frac{K\log(K^\rho n^{r+1})\log(K)}{n}} \lesssim \sqrt{\frac{K\log(n^{\rho+r+2})\log(n)}{n}} \lesssim \sqrt{\frac{K\log(n)^2}{n}},$$

to satisfy Assumption 9 we may assume $\sqrt{K\log(n)^2/n} = o(1)$ as in Remark 2.3 of Chen and Christensen (2015) for the case of exponential β-mixing regressors.

B.3 Theorem 3.1

Before delving into the proof of Theorem 3.1, note that we can decompose $\widehat{\Pi}_2 - \Pi_2$ as

$$\widehat{\Pi}_2 - \Pi_2 = (\widehat{\Pi}_2 - \widehat{\Pi}_2^*) + (\widehat{\Pi}_2^* - \widetilde{\Pi}_2) + (\widetilde{\Pi}_2 - \Pi_2),$$

where $\widetilde{\Pi}_2$ is the projection of Π_2 onto the linear space spanned by the sieve. The last two terms can be handled directly with the theory developed by Chen and Christensen (2015). Specifically, their Lemma 2.3 controls the second term (variance term), while Lemma 2.4 handles the third term (bias term). This means here we can focus on the first term, which is due to using generated regressors $\widehat{\epsilon}_{1t}$ in the second step.

Since $\widehat{\Pi}_2$ can be decomposed in d_Y rows of semiparametric coefficients, we further reduce to the scalar case. Let π_2 be any row of Π_2 and, with a slight abuse of notation, Y the vector of observations of the component of Y_t of the same row, so that one may write

$$\begin{aligned}\widehat{\pi}_2(x) - \widehat{\pi}_2^*(x) &= \widetilde{b}_\pi^K(x)\big(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi\big)^-\big(\widehat{\widetilde{B}}_\pi - \widetilde{B}_\pi\big)'Y + \widetilde{b}_\pi^K(x)\left[\big(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi\big)^- - \big(\widetilde{B}_\pi'\widetilde{B}_\pi\big)^-\right]\widetilde{B}_\pi'Y \\ &= I + II\end{aligned}$$

where $\widetilde{b}_\pi^K(x) = \Gamma_{B,2}^{-1/2} b_\pi^K(x)$ is the orthonormalized sieve according to $\Gamma_{B,2} := \mathbb{E}[b_\pi^K(W_{2t})\, b_\pi^K(W_{2t})']$, $\widetilde{B}_\pi$ is the *infeasible* orthonormalized design matrix (involving ϵ_{1t}) and $\widehat{\widetilde{B}}_\pi$ is *feasible* orthonormalized design matrix (involving $\widehat{\epsilon}_{1t}$). In particular, note that

$$\widehat{B}_\pi = B_\pi + R_n, \quad \text{where} \quad R_n := \begin{bmatrix} 0 & & 0 & \widehat{\epsilon}_{11} - \epsilon_{11} \\ \vdots & \dots & \vdots & \vdots \\ 0 & & 0 & \widehat{\epsilon}_{1n} - \epsilon_{1n} \end{bmatrix} \in \mathbb{R}^{n\times K},$$

which implies $\widehat{\widetilde{B}}_\pi - \widetilde{B}_\pi = R_n\,\Gamma_{B,2}^{-1/2} =: \widetilde{R}_n$.

The next Lemma provides a bound for the difference $(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n) - (\widetilde{B}_\pi'\widetilde{B}_\pi/n)$ that will be useful in the proof of Theorem 3.1 below.

Lemma B.2. *Under the setup of Theorem 1 in Chen and Christensen (2015), it holds*

$$\left\|(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n) - (\widetilde{B}_\pi'\widetilde{B}_\pi/n)\right\| = O_P(\sqrt{K/n}).$$

Proof. Using the expansion $\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi = \widetilde{B}_\pi'\widetilde{B}_\pi + (\widetilde{B}_\pi'\widetilde{R}_n + \widetilde{R}_n'\widetilde{B}_\pi) + \widetilde{R}_n'\widetilde{R}_n$, one immediately finds that $\left\|(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n) - (\widetilde{B}_\pi'\widetilde{B}_\pi/n)\right\| \leqslant 2\left\|\widetilde{B}_\pi'\widetilde{R}_n/n\right\| + \left\|\widetilde{R}_n'\widetilde{R}_n/n\right\|$. The second right-hand

side factor satisfies $\left\|\widetilde{R}_n'\widetilde{R}_n/n\right\| \leqslant \lambda_{K,n}^2\left\|R_n'R_n/n\right\|$. Moreover,

$$\begin{aligned}\left\|R_n'R_n/n\right\| &= \left\|\frac{1}{n}\sum_{t=1}^{n}(\widehat{\epsilon}_{1t}-\epsilon_{1t})^2\right\| = \left\|\frac{1}{n}\sum_{t=1}^{n}(\Pi_1-\widehat{\Pi}_1)'W_{1t}W_{1t}'(\Pi_1-\widehat{\Pi}_1)\right\| \\ &\leqslant \left\|\Pi_1-\widehat{\Pi}_1\right\|^2\left\|W_1'W_1/n\right\| = O_P(n^{-1}),\end{aligned}$$

since $\|W_1'W_1/n\| = O_P(1)$. Under Assumption 8, $\lambda_{K,n}^2/n = o_P(\sqrt{K/n})$ since B-splines and wavelets satisfy $\lambda_{K,n} \lesssim 1$. Consequently, $\left\|\widetilde{R}_n'\widetilde{R}_n/n\right\| = o_P(\sqrt{K/n})$. Factor $\|\widetilde{B}_\pi'R_n/n\|$ is also straightforward, but depends on sieve dimension K,

$$\begin{aligned}\left\|\widetilde{B}_\pi'R_n/n\right\| &\leqslant \left\|\frac{1}{n}\sum_{t=1}^{n}\widetilde{b}_\pi^K(W_{2t})(\widehat{\epsilon}_{1t}-\epsilon_{1t})\right\| = \left\|\frac{1}{n}\sum_{t=1}^{n}\widetilde{b}_\pi^K(W_{2t})W_{1t}'(\Pi_1-\widehat{\Pi}_1)\right\| \\ &\leqslant \left\|\Pi_1-\widehat{\Pi}_1\right\|\left\|\widetilde{B}_\pi'W_1/n\right\| = O_P(\sqrt{K/n}),\end{aligned}$$

since $\|\widetilde{B}_\pi'W_1/n\| = O_P(\sqrt{K})$ as the column dimension of W_1 is fixed. The claim then follows by noting $O_P(\sqrt{K/n})$ is the dominating order of convergence. □

Proof of Theorem 3.1. Since $\widehat{\Pi}_1$ is a least squares estimator of a linear equation, the rate of convergence is the parametric rate $n^{-1/2}$. The first result is therefore immediate. For the second step, we use $\left\|\widehat{\Pi}_2-\Pi_2\right\|_\infty \leqslant \left\|\widehat{\Pi}_2-\widehat{\Pi}_2^*\right\|_\infty + \left\|\widehat{\Pi}_2^*-\Pi_2\right\|_\infty$, and bound explicitly the first right-hand side term. For a given component of the regression function,

$$|\widehat{\pi}_2(x)-\widehat{\pi}_2^*(x)| \leqslant |I|+|II|.$$

We now control each term on the right side.

(1) It holds

$$\begin{aligned}|I| &\leqslant \|\widetilde{b}_\pi^K(x)\|\left\|\big(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n\big)^-\right\|\left\|\big(\widehat{\widetilde{B}}_\pi-\widetilde{B}_\pi\big)'Y/n\right\| \\ &\leqslant \sup_{x\in\mathcal{W}_2}\|\widetilde{b}_\pi^K(x)\|\left\|\big(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n\big)^-\right\|\left\|\big(\widehat{\widetilde{B}}_\pi-\widetilde{B}_\pi\big)'Y/n\right\| \\ &\leqslant \zeta_{K,n}\lambda_{K,n}\left\|\big(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n\big)^-\right\|\left\|\big(\widehat{\widetilde{B}}_\pi-\widetilde{B}_\pi\big)'Y/n\right\|.\end{aligned}$$

Let $\mathcal{A}_n$ denote the event on which $\left\|\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n - I_K\right\| \leqslant 1/2$, so that $\left\|\big(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n\big)^-\right\| \leqslant 2$ on $\mathcal{A}_n$. Notice that since $\|(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n) - (\widetilde{B}_\pi'\widetilde{B}_\pi/n)\| = o_P(1)$ (Lemma B.2) and, by assumption, $\|\widetilde{B}_\pi'\widetilde{B}_\pi/n - I_K\| = o_P(1)$, then $\mathbb{P}(\mathcal{A}_n^c) = o(1)$. On $\mathcal{A}_n$, then

$$|I| \lesssim \zeta_{K,n}\lambda_{K,n}^2\left\|\big(\widehat{B}_\pi-B_\pi\big)'Y/n\right\| = \zeta_{K,n}\lambda_{K,n}^2\left\|R_n'Y/n\right\|.$$

From $R_n'Y = \sum_{t=1}^n b_\pi^K(W_{2t})(\widehat{\epsilon}_{1t}-\epsilon_{1t})Y_t = (\Pi_1-\widehat{\Pi}_1)'W_1'Y$ it follows that $\left\|R_n'Y/n\right\| \leqslant$

$\left\|\Pi_1 - \widehat{\Pi}_1\right\| \left\|W_1'Y/n\right\|$ on $\mathcal{A}_n$, meaning $|I| = O_P\left(\zeta_{K,n}\lambda_{K,n}^2/\sqrt{n}\right)$ as $\|W_1'Y/n\| = O_P(1)$ and $\mathbb{P}(\mathcal{A}_n^{\mathrm{c}}) = o(1)$.

(2) Again we proceed by uniformly bounding II according to

$$|II| \leqslant \zeta_{K,n}\lambda_{K,n} \left\|\left(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n\right)^- - \left(\widetilde{B}_\pi'\widetilde{B}_\pi/n\right)^-\right\| \left\|\widetilde{B}_\pi'Y/n\right\|.$$

The last factor has order $\|\widetilde{B}_\pi'Y/n\| = O_P(\sqrt{K})$ since $\widetilde{B}_\pi$ is growing in row dimension with K. For the middle term, introduce $\Delta_B := \widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n - \widetilde{B}_\pi'\widetilde{B}_\pi/n$ and event $\mathcal{B}_n := \left\{\left\|\left(\widetilde{B}_\pi'\widetilde{B}_\pi/n\right)^- \Delta_B\right\| \leqslant 1/2\right\} \cap \left\{\left\|\widetilde{B}_\pi'\widetilde{B}_\pi/n - I_K\right\| \leqslant 1/2\right\}$. On $\mathcal{B}_n$, we can apply the bound (Horn and Johnson, 2012)

$$\left\|\left(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n\right)^- - \left(\widetilde{B}_\pi'\widetilde{B}_\pi/n\right)^-\right\| \leqslant \frac{\|(\widetilde{B}_\pi'\widetilde{B}_\pi/n)^-\|^2 \|\Delta_B\|}{1 - \|(\widetilde{B}_\pi'\widetilde{B}_\pi/n)^- \Delta_B\|} \lesssim \left\|\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n - \widetilde{B}_\pi'\widetilde{B}_\pi/n\right\|.$$

Since $\left\|\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n - \widetilde{B}_\pi'\widetilde{B}_\pi/n\right\| = O_P(\sqrt{K/n})$ by Lemma B.2, we get

$$|II| = O_P\left(\zeta_{K,n}\lambda_{K,n}\frac{K}{\sqrt{n}}\right)$$

on $\mathcal{B}_n$. Finally, using $\mathbb{P}((A \cap B)^{\mathrm{c}}) \leqslant \mathbb{P}(A^{\mathrm{c}}) + \mathbb{P}(B^{\mathrm{c}})$ we note that $\mathbb{P}(\mathcal{B}_n^{\mathrm{c}}) = o(1)$ so that the bound asymptotically holds irrespective of event $\mathcal{B}_n$.

Thus, we have shown that

$$|\widehat{\pi}_2(x) - \widehat{\pi}_2^*(x)| \leqslant O_P\left(\zeta_{K,n}\lambda_{K,n}^2\frac{1}{\sqrt{n}}\right) + O_P\left(\zeta_{K,n}\lambda_{K,n}\frac{K}{\sqrt{n}}\right) = O_P\left(\zeta_{K,n}\lambda_{K,n}\frac{K}{\sqrt{n}}\right)$$

as clearly $\sqrt{n}^{-1} = o(K/\sqrt{n})$ and, as discussed in the proof of Lemma B.2, $\lambda_{K,n}^2/n = o_P(\sqrt{K/n})$. This bound is uniform in x and holds for each of the (finite number of) components of $\widehat{\Pi}_2$, therefore the proof is complete. □

B.4 Theorem 4.1

Before proving impulse response consistency, we show that the functional moving average coefficient matrices Γ_j can be consistently estimated with $\widehat{\Pi}_1$ and $\widehat{\Pi}_2$.

Lemma B.3. *Under the assumptions of Theorem 3.1 and for any fixed integer $j \geqslant 0$, it holds*

$$\|\widehat{\Gamma}_j - \Gamma_j\|_\infty = o_P(1).$$

Proof. By definition, recall that $\Gamma(L) = \Psi(L)G(L)$ where $\Psi = (I_d - A(L)L)^{-1}$. Since $\Psi(L)$ is an MA(∞) lag polynomial, we have that $\Gamma(L) = (\sum_{k=0}^{\infty}\Psi_k L^k)(G_0 + G_1 L + \ldots + G_p L^p)$, where $\Psi_0 = I_d$, $\{\Psi_k\}_{k=1}^{\infty}$ are purely real matrices and G_0 is a functional vector

that may also contain linear components (i.e. allow linear functions of X_t). This means that Γ_j is a convolution of real and functional matrices, $\Gamma_j = \sum_{k=1}^{\min\{j,p\}} \Psi_{j-k} G_k$. The linear coefficients of $A(L)$ can be consistently estimated by $\widehat{\Pi}_1$ and $\widehat{\Pi}_2$, and, thus, the plug-in estimate $\widehat{\Psi}_j$ is consistent for Ψ_j (Lütkepohl, 2005). Therefore,

$$\begin{aligned}\|\widehat{\Gamma}_j - \Gamma_j\|_\infty &\leqslant \sum_{k=1}^{\min\{j,p\}} \left\|\Psi_{j-k} - \widehat{\Psi}_{j-k}\right\|_\infty \|G_k\|_\infty + \left\|\widehat{\Psi}_{j-k}\right\|_\infty \left\|G_k - \widehat{G}_k\right\|_\infty \\ &\leqslant \sum_{k=1}^{\min\{j,p\}} o_p(1) C_{G,k} + O_P(1) o_p(1) = o_p(1),\end{aligned}$$

where $C_{G,k}$ is a constant and $\|G_k - \widehat{G}_k\|_\infty = o_p(1)$ as a direct consequence of Theorem 3.1. □

Recall now that the sample estimate for the relaxed-shock impulse response is

$$\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) = \widehat{\Theta}_{h,\cdot 1} \delta\, n^{-1} \sum_{t=1}^{n} \rho(\widehat{\epsilon}_{1t}) + \sum_{j=0}^{h} \widehat{V}_{j,\ell}(\delta)$$

where

$$\widehat{V}_{j,\ell}(\delta) = \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widehat{\widetilde{\delta}}_t\big) = \frac{1}{n-j} \sum_{t=1}^{n-j} \left[\widehat{\Gamma}_j \widehat{\gamma}_j(X_{t+j:t}; \widehat{\widetilde{\delta}}_t) - \widehat{\Gamma}_j X_{t+j}\right].$$

Therefore, the estimated horizon h impulse response of the ℓth variable is

$$\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) := \widehat{\Theta}_{h,\ell 1} \delta\, n^{-1} \sum_{t=1}^{n} \rho(\widehat{\epsilon}_{1t}) + \sum_{j=0}^{h} \left[\frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widehat{\widetilde{\delta}}_t\big)\right].$$

Lemma B.4. *Under the assumptions of Theorem 4.1 , let $x_{j:0} = (x_j, \ldots, x_0) \in \mathcal{X}^j$ and $\varepsilon \in \mathcal{E}_1$ be nonrandom quantities. Let $\widetilde{\delta}$ be the relaxed shock determined by δ, ρ and ε. Then,*

(i) $\sup_{x_{j:0},\varepsilon} |\widehat{\gamma}_j(x_{j:0}; \widetilde{\delta}) - \gamma_j(x_{j:0}; \widetilde{\delta})| = o_P(1)$,

(ii) $\sup_{x_{j:0},\varepsilon} |\widehat{v}_{j,\ell}\big(x_{j:0}; \widetilde{\delta}\big) - v_{j,\ell}\big(x_{j:0}; \widetilde{\delta}\big)| = o_P(1)$,

for any fixed integers $j \geqslant 0$ and $\ell \in \{1, \ldots, d\}$.

Proof.

(i) We have that

$$|\widehat{\gamma}_j(x_{j:0}; \delta) - \gamma_j(x_{j:0}; \delta)| = \left|\sum_{k=1}^{j} \left[(\widehat{\Gamma}_{k,11} x_{j-k}(\widetilde{\delta}) - \widehat{\Gamma}_{k,11} x_{j-k}) - (\Gamma_{k,11} x_{j-k}(\widetilde{\delta}) - \Gamma_{k,11} x_{j-k})\right]\right|$$

$$\leqslant \sum_{k=1}^{j} \left|\widehat{\Gamma}_{k,11} x_{j-k}(\widetilde{\delta}) - \Gamma_{k,11} x_{j-k}(\widetilde{\delta})\right| + \sum_{k=1}^{j} \left|\widehat{\Gamma}_{k,11} x_{j-k} - \Gamma_{k,11} x_{j-k}\right|.$$

This yields $\sup_{x_{j:0},\varepsilon} |\widehat{\gamma}_j(x_{j:0};\widetilde{\delta}) - \gamma_j(x_{j:0};\widetilde{\delta})| \leqslant 2j \sup_{x\in\mathcal{X}} \left|\widehat{\Gamma}_{k,11} x - \Gamma_{k,11} x\right|$. Since j is finite and fixed and the uniform consistency bound of Lemma B.3 holds, a fortiori $\sup_{x\in\mathcal{X}} \left|\widehat{\Gamma}_{k,11} x - \Gamma_{k,11} x\right| = o_P(1)$.

(ii) Similarly to above,

$$\begin{aligned}
|\widehat{v}_{j,\ell}(x_{j:0};\widetilde{\delta}) - v_{j,\ell}(x_{j:0};\widetilde{\delta})| &= \left|\left(\widehat{\Gamma}_{j,\ell}\widehat{\gamma}_j(x_{j:0};\widetilde{\delta}) - \Gamma_{j,\ell}\gamma_j(x_{j:0};\widetilde{\delta})\right) - \left(\widehat{\Gamma}_{j,\ell} x_j - \Gamma_{j,\ell} x_j\right)\right| \\
&\leqslant \|\widehat{\Gamma}_{j,\ell} - \Gamma_{j,\ell}\|_\infty + \|\Gamma_{j,\ell}\|_\infty |\widehat{\gamma}_j(x_{j:0};\delta) - \gamma_j(x_{j:0};\delta)| \\
&\quad + |\widehat{\Gamma}_{j,\ell} x_j - \Gamma_{j,\ell} x_j| \\
&\leqslant 2\|\widehat{\Gamma}_{j,\ell} - \Gamma_{j,\ell}\|_\infty + C_{\Gamma,j,l}\, |\widehat{\gamma}_j(x_{j:0};\delta) - \gamma_j(x_{j:0};\delta)|,
\end{aligned}$$

where we have used that $\gamma_j(x_{j:0};\widetilde{\delta}) \in \mathcal{X}$ to derive the first term in the second line. In the last line, $C_{\Gamma,j,l}$ is a constant such that $\|\Gamma_{j,\ell}\|_\infty \leqslant \sum_{k=1}^{\min\{j,p\}} \|\Psi_{j-k}\|_\infty \|G_k\|_\infty \leqslant C_{\Gamma,j,l}$. The claim then follows thanks to Lemma B.3 and (i).

□

In what follows, define $\widehat{v}_{j,\ell}(X_{t+j:t};\widetilde{\delta}_t)$ to be a version of $v_{j,\ell}$ that is constructed using coefficient estimates from $\{\widehat{\Pi}_1, \widehat{\Pi}_2\}$ but evaluated on the true innovations ϵ_t.

Proof of Theorem 4.1. If we introduce

$$\widetilde{\mathrm{IRF}}_{h,\ell}(\delta)^* := \widehat{\Theta}_{h,\ell 1}\delta\, n^{-1} \sum_{t=1}^{n} \rho(\epsilon_{1t}) + \sum_{j=0}^{h} \left[\frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}(X_{t+j:t};\widetilde{\delta}_t)\right],$$

then clearly

$$\begin{aligned}
\left|\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) - \widetilde{\mathrm{IRF}}_{h,\ell}(\delta)\right| &\leqslant \left|\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) - \widetilde{\mathrm{IRF}}^*_{h,\ell}(\delta)\right| + \left|\widetilde{\mathrm{IRF}}^*_{h,\ell}(\delta) - \widetilde{\mathrm{IRF}}_{h,\ell}(\delta)\right| \\
&= I + II.
\end{aligned}$$

To control II, we can observe

$$\begin{aligned}
II \leqslant &\left|\widehat{\Theta}_{h,\ell 1}\delta\, n^{-1} \sum_{t=1}^{n} \rho(\epsilon_{1t}) - \Theta_{h,\ell 1}\delta\mathbb{E}[\rho(\epsilon_{1t})]\right| \\
&+ \sum_{j=0}^{h} \left|\frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}(X_{t+j:t};\widetilde{\delta}_t) - \mathbb{E}[v_{j,\ell}(X_{t+j:t};\widetilde{\delta})]\right|
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \delta \left|\widehat{\Theta}_{h,\ell 1} - \Theta_{h,\ell 1}\right| \left|n^{-1}\sum_{t=1}^{n} \rho(\epsilon_{1t})\right| + \delta \left|\widehat{\Theta}_{h,\ell 1}\right| \left|n^{-1}\sum_{t=1}^{n} \rho(\epsilon_{1t}) - \mathbb{E}[\rho(\epsilon_{1t})]\right| \\
&\quad + \sum_{j=0}^{h} \left| \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) - \mathbb{E}[v_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}\big)] \right| \\
&\leqslant \delta \left|\widehat{\Theta}_{h,\ell 1} - \Theta_{h,\ell 1}\right| \left|n^{-1}\sum_{t=1}^{n} \rho(\epsilon_{1t})\right| + \delta \left|\widehat{\Theta}_{h,\ell 1}\right| \left|n^{-1}\sum_{t=1}^{n} \rho(\epsilon_{1t}) - \mathbb{E}[\rho(\epsilon_{1t})]\right| \\
&\quad + \sum_{j=0}^{h} \left| \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) - v_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) \right| \\
&\quad + \sum_{j=0}^{h} \left| \frac{1}{n-j} \sum_{t=1}^{n-j} v_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) - \mathbb{E}[v_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}\big)] \right|.
\end{aligned}
$$

The first two terms in the last bound are $o_P(1)$ since $\left|\widehat{\Theta}_{h,\ell 1} - \Theta_{h,\ell 1}\right| = o_P(1)$, as discussed in Lemma B.3, and $n^{-1}\sum_{t=1}^{n} \rho(\epsilon_{1t}) \xrightarrow{p} \mathbb{E}[\rho(\epsilon_{1t})]$ by a WLLN. For the other terms in the last sum above, we similarly note that

$$
\left| \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) - v_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) \right| = o_P(1)
$$

from Lemma B.4, while, thanks again to a WLLN, it holds

$$
\left| \frac{1}{n-j} \sum_{t=1}^{n-j} v_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) - \mathbb{E}[v_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}\big)] \right| = o_P(1).
$$

Since h is fixed finite, this implies that $II = o_P(1)$.

Considering now I, we can write

$$
\begin{aligned}
I &\leqslant \delta \left|\widehat{\Theta}_{h,\ell 1}\right| \left|n^{-1}\sum_{t=1}^{n} \rho(\widehat{\epsilon}_{1t}) - \rho(\epsilon_{1t})\right| + \sum_{j=0}^{h} \left| \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widehat{\widetilde{\delta}}_t\big) - \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) \right| \\
&= I' + I''.
\end{aligned}
$$

Since by assumption ρ is a bump function, thus continuously differentiable over the range of ϵ_t, by the mean value theorem

$$
\left|n^{-1}\sum_{t=1}^{n} \rho(\widehat{\epsilon}_{1t}) - \rho(\epsilon_{1t})\right| \leqslant n^{-1}\sum_{t=1}^{n} |\rho'_t| \left|\widehat{\epsilon}_{1t} - \epsilon_{1t}\right|
$$

for a sequence $\{\rho'_t\}_{t=1}^{n}$ of evaluations of first-order derivative ρ' at values $\overline{\epsilon_t}$ in the interval with endpoint ϵ_t and $\widehat{\epsilon}_t$. One can use $|\rho'_t| \leqslant C_{\rho'}$ with a finite positive constant $C_{\rho'}$, and

by recalling that $\widehat{\epsilon}_{1t} - \epsilon_{1t} = (\Pi_1 - \widehat{\Pi}_1)' W_{1t}$ one, thus, gets

$$\left| n^{-1} \sum_{t=1}^{n} \rho(\widehat{\epsilon}_{1t}) - \rho(\epsilon_{1t}) \right| \leqslant C_{\rho'} \frac{1}{n} \sum_{t=1}^{n} \left| (\Pi_1 - \widehat{\Pi}_1)' W_{1t} \right| \leqslant C_{\rho'} \|\Pi_1 - \widehat{\Pi}_1\|_2 \frac{1}{n} \sum_{t=1}^{n} \|W_{1t}\|_2 = o_P(1).$$

This proves that term I' is itself $o_P(1)$. Finally, to control I'', we use that by construction estimator $\widehat{\Pi}_2$ is composed of sufficiently regular functional elements i.e. B-spline estimates of order 1 or greater. Thanks again to the mean value theorem

$$\begin{aligned} \left| \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widehat{\widetilde{\delta}}\big) - \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) \right| &\leqslant \frac{1}{n-j} \sum_{t=1}^{n-j} \left| \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widehat{\widetilde{\delta}}\big) - \widehat{v}_{j,\ell}\big(X_{t+j:t}; \widetilde{\delta}_t\big) \right| \\ &\leqslant C_{\widehat{v}',j,\ell} \frac{1}{n-j} \sum_{t=1}^{n-j} \left| \widehat{\epsilon}_{1t} - \epsilon_{1t} \right| \end{aligned}$$

for any fixed j and some $C_{\widehat{v}',j,\ell} > 0$. This holds since $\widehat{v}_{j,\ell}$ is uniformly continuous by construction. Note that we have assumed that the nonlinear part of Π_2 belongs to a Hölder class with smoothness $s > 1$ (for simplicity, assume here that s is integer, otherwise a similar argument can be made). Then, even though $C_{\widehat{v}',j,\ell}$ depends on the sample, it is bounded above in probability for n sufficiently large. Following the discussion of term I', we deduce that the last line in the display above is $o_p(1)$. As h is finite and independent of n, it follows that also I'' is of order $o_P(1)$.

Finally, to obtain uniformity with respect to $\delta \in [-\mathcal{D}, \mathcal{D}]$, simply note that bounds on I and II are explicit in δ, therefore

$$\sup_{\delta \in [-\mathcal{D}, \mathcal{D}]} \left| \widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) - \widetilde{\mathrm{IRF}}_{h,\ell}(\delta) \right| \leqslant \mathcal{D} \times o_P(1),$$

concluding the proof. □

C Simulation Details

C.1 Benchmark Bivariate Design

The first simulation setup involves a bivariate DGP where the structural shock does not directly affect other observables. This is a simple environment to check that indeed the two-step estimator recover the nonlinear component of the model and impulse responses are consistently estimated, and that the MSE does not worsen excessively.

I consider three bivariate data generation processes. DGP 1 sets X_t to be a fully

exogenous innovation process,

$$\begin{aligned} X_t &= \epsilon_{1t}, \\ Y_t &= 0.5Y_{t-1} + 0.5X_t + 0.3X_{t-1} - 0.4\max(0, X_t) + 0.3\max(0, X_{t-1}) + \epsilon_{2t}. \end{aligned} \tag{18}$$

DGP 2 adds an autoregressive component to X_t, but maintains exogeneity,

$$\begin{aligned} X_t &= 0.5X_{t-1} + \epsilon_{1t}, \\ Y_t &= 0.5Y_{t-1} + 0.5X_t + 0.3X_{t-1} - 0.4\max(0, X_t) + 0.3\max(0, X_{t-1}) + \epsilon_{2t}. \end{aligned} \tag{19}$$

Finally, DGP 3 add an endogenous effect of Y_{t-1} on the structural variable by setting

$$\begin{aligned} X_t &= 0.5X_{t-1} + 0.2Y_{t-1} + \epsilon_{1t}, \\ Y_t &= 0.5Y_{t-1} + 0.5X_t + 0.3X_{t-1} - 0.4\max(0, X_t) + 0.3\max(0, X_{t-1}) + \epsilon_{2t}. \end{aligned} \tag{20}$$

Following Assumption 1, innovations are mutually independent. To accommodate Assumption 4, both ϵ_{1t} and ϵ_{2t} are drawn from a truncated standard Gaussian distribution over $[-3, 3]$.[6] All DGPs are centered to have zero intercept in population.

We evaluate bias and MSE plots using 10 000 Monte Carlo simulation. For a chosen horizon H, the impact of a relaxed shock on ϵ_{1t} is evaluated on Y_{t+h} for $h = 1, \ldots, H$. To compute the population IRF, we employ a direct simulation strategy that replicates the shock's propagation through the model and we use 10^5 replications. To evaluate the estimated IRF, the two-step procedure is implemented: a sample of length n is drawn, the linear least squares and the semiparametric series estimators of the model are used to estimate the model and the relaxed IRF is computed following Proposition 2.1. For the sake of brevity, we discuss the case of $\delta = 1$ and we set the shock relaxation function to be

$$\rho(z) = \mathbb{I}\{x \leqslant 3\} \exp\left(1 + \left[\left|\frac{z}{3}\right|^4 - 1\right]^{-1}\right)$$

It can be easily checked that this choice of ρ is compatible with shocks of size $0 \leqslant |\delta| \leqslant 1$. Choices of $\delta = -1$ and $\delta = \pm 0.5$ yield similar results in simulations, so we do not discuss them here.

Figure D.1 contains the results for sample size $n = 240$. This choice is motivated by considering the average sample sizes found in most macroeconometric settings: it is equivalent to 20 years of monthly data or 60 yearly of quarterly data (Gonçalves et al., 2021). The benchmark method is an OLS regression that relies on a priori knowledge of the underlying DGP specification. Given the moderate sample size, to construct the

[6] Let $e_{it} \sim \mathcal{N}(0, 1)$ for $i = 1, 2$, then the truncated Gaussian innovations used in simulation are set to be $\epsilon_{it} = \min(\max(-3, e_{it}), 3)$. The resulting r.v.s have a non-continuous density with two mass points at -3 and 3. However, in practice, since these masses are negligible, for the moderate sample sizes used this choice does not create issues.

cubic spline sieve estimator of the nonlinear component of the model we use a single knot, located at 0. The simulations in Figure D.1 show that while the MSE is slighlty higher for the sieve model, the bias is comparable across methods. Note that for DGP 3, due to the dependence of the structural variable on non-structural series lags, the MSE and bias increase significantly, and there is no meaningful difference in performance between the two estimation approaches.

C.2 Structural Partial Identification Design

To showcase the validity of the proposed sieve estimator under the type of partial structural identification discussed in the paper, we again rely on the simulation design proposed by Gonçalves et al. (2021). All specifications are block-recursive, and require estimating the contemporaneous effects of a structural shock on non-structural variables, unlike in the previous section.

The form of the DGPs is

$$B_0 Z_t = B_1 Z_{t-1} + C_0 f(X_t) + C_1 f(X_{t-1}) + \epsilon_t,$$

where in all variations of the model

$$B_0 = \begin{bmatrix} 1 & 0 & 0 \\ -0.45 & 1 & -0.3 \\ -0.05 & 0.1 & 1 \end{bmatrix}, \quad C_0 = \begin{bmatrix} 0 \\ -0.2 \\ 0.08 \end{bmatrix}, \quad \text{and} \quad C_1 = \begin{bmatrix} 0 \\ -0.1 \\ 0.2 \end{bmatrix}.$$

I focus on the case $f(x) = \max(0, x)$, since this type of nonlinearity is simpler to study. DGP 4 treats X_t as an exogenous shock by setting

$$B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.15 & 0.17 & -0.18 \\ -0.08 & 0.03 & 0.6 \end{bmatrix};$$

DGP 5 add serial correlation to X_t,

$$B_1 = \begin{bmatrix} -0.13 & 0 & 0 \\ 0.15 & 0.17 & -0.18 \\ -0.08 & 0.03 & 0.6 \end{bmatrix};$$

and DGP 6 includes dependence on Y_{t-1},

$$B_1 = \begin{bmatrix} -0.13 & 0.05 & -0.01 \\ 0.15 & 0.17 & -0.18 \\ -0.08 & 0.03 & 0.6 \end{bmatrix}.$$

For these data generating processes, we employ the same setup of simulations with DGPs 1-3, including the number of replications as well as the type of relaxed shock. as well as the sieve grid. Here too we evaluate MSE and bias of both the sieve and the correct

specification OLS estimators with as sample size of $n = 240$ observations. The results in Figure D.2 show again that there is little difference in terms of performance between the semiparametric sieve approach and a correctly-specified OLS regression.

C.3 Model Misspecification

The results from benchmark simulations support the use of the sieve IRF estimator in a sample of moderate size, since it performs comparably to a regression performed with a priori knowledge of the underlying DGP. We now show that the semiparametric approach is also robust to model misspecification compared to simpler specifications involving fixed choices for nonlinear transformations.

To this end, we modify DGP 2 to use a smooth nonlinear transformation to define the effect of structural variable X_t on Y_t. That is, there is no compounding of linear and nonlinear effects. The autoregressive coefficient in the equation for X_t is also increased to make the shock more persistent. The new data generating process, DGP 7, is, thus, given by

$$\begin{aligned} X_t &= 0.8X_{t-1} + \epsilon_{1t}, \\ Y_t &= 0.5Y_{t-1} + 0.9\varphi(X_t) + 0.5\varphi(X_{t-1}) + \epsilon_{2t}. \end{aligned} \tag{21}$$

where $\varphi(x) := (x-1)(0.5 + \tanh(x-1)/2)$.

To emphasize the difference in estimated IRFs, in this setup we focus on $\delta = \pm 2$, which requires adapting the choice of innovations and shock relaxation function. In simulations of DGP 7, ϵ_{1t} and ϵ_{2t} are both drawn from a truncated standard Gaussian distribution over $[-5, 5]$. The shock relaxation function of this setup is given by

$$\rho(z) = \mathbb{I}\{x \leqslant 5\} \exp\left(1 + \left[\left|\frac{z}{5}\right|^{3.9} - 1\right]^{-1}\right).$$

This form of ρ is adapted to choices of δ such that $0 < |\delta| \leqslant 2$. The sieve grid now consists of 4 equidistant knots within $(-5, 5)$. We use the same numbers of replications as in the previous simulations. Finally, the regression design is identical to that used for DGP 2 under correct specification.

The results obtained with sample size $n = 2400$ are collected in Figure D.3. We choose this larger sample size to clearly showcase the inconsistency of impulse responses under misspecification: as it can be observed, the simple OLS estimator involving the negative-censoring transform produces IRF estimates with consistently worse MSE and bias than those of the sieve estimator at almost all horizons. Similar results are also obtained for more moderate shocks $\delta = \pm 1$, but the differences are less pronounced. These simulations

suggest that the semiparametric sieve estimator can produce substantially better IRF estimates in large samples than methods involving nonlinear transformations selected a priori.

In this setup, it is also important to highlight the fact that the poor performance of OLS IRF estimates does not come from $\varphi(x)$ being "complex", and, thus, hard to approximate by combinations of simple functions. In fact, if in DGP 7 function φ is replaced by $\widetilde{\varphi}(x) := \varphi(x+1)$, the differences between sieve and OLS impulse response estimates become minimal in simulations, with the bias of the latter decreasing by approximately an order of magnitude, see Figure D.4. This is simply due to the fact that $\widetilde{\varphi}(x)$ is well approximated by $\max(0, x)$ directly. However, one then requires either prior knowledge or sheer luck when constructing the nonlinear transforms of X_t for an OLS regression. The proposed series estimator, instead, just requires an appropriate choice of sieve. Many data-driven procedures to select sieves in applications have been proposed, see for example the discussion in Kang (2021).

D Additional Plots

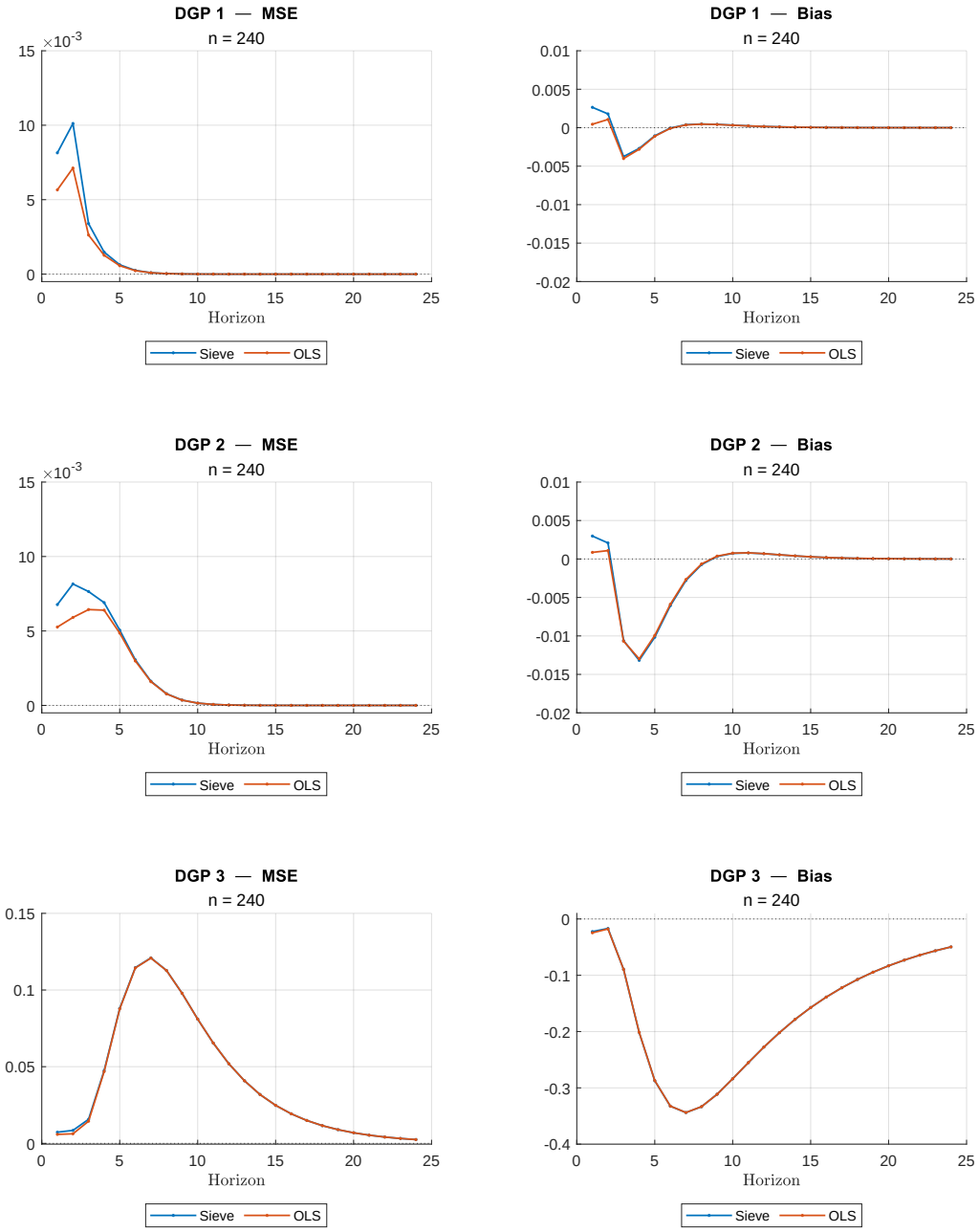


Figure D.1: Simulations results for DGPs 1-3.

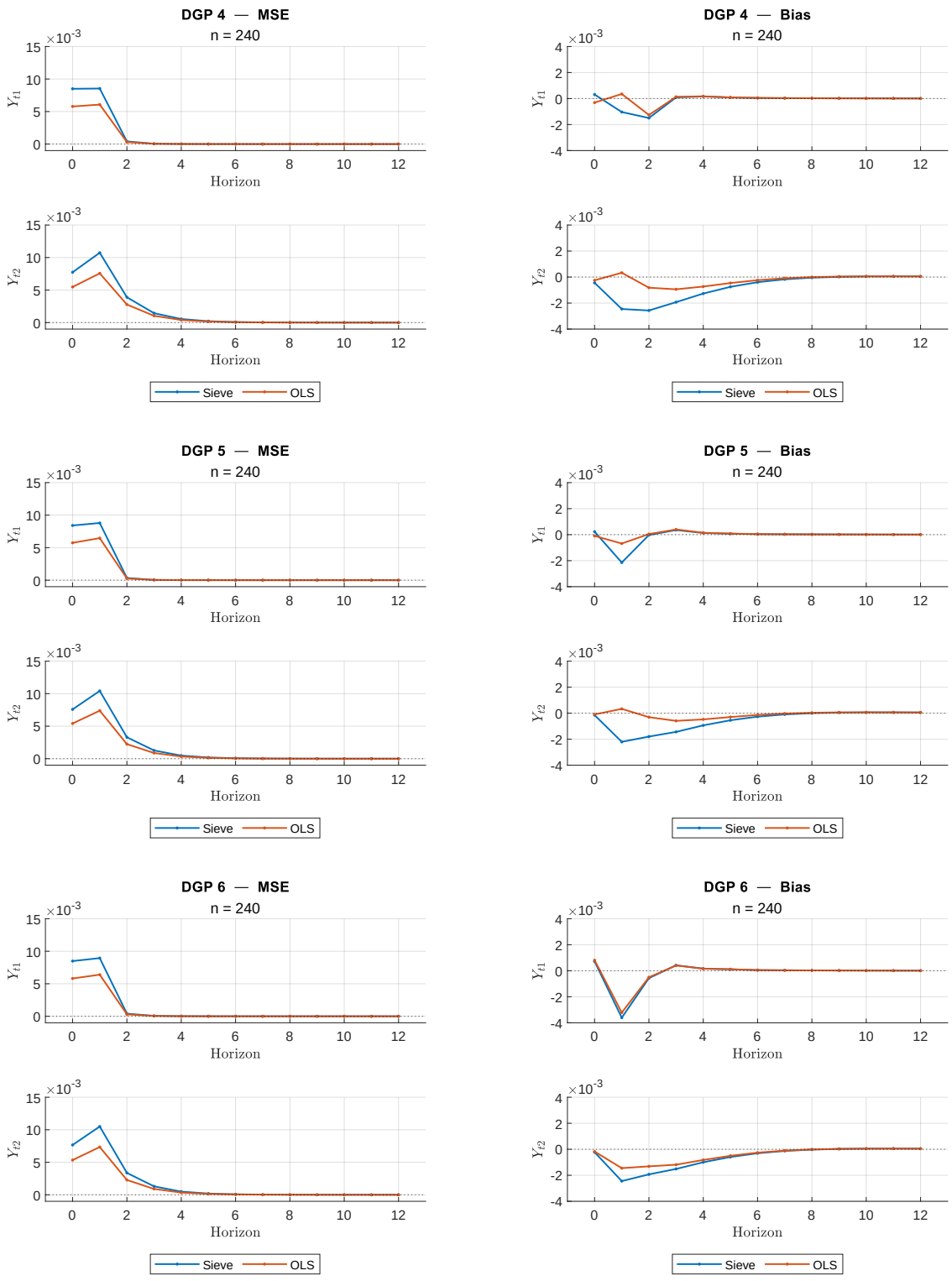
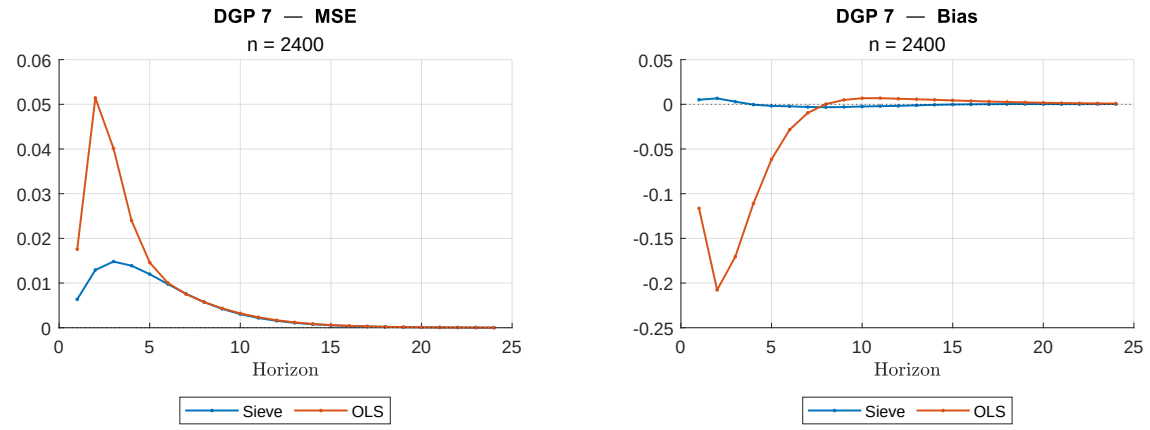
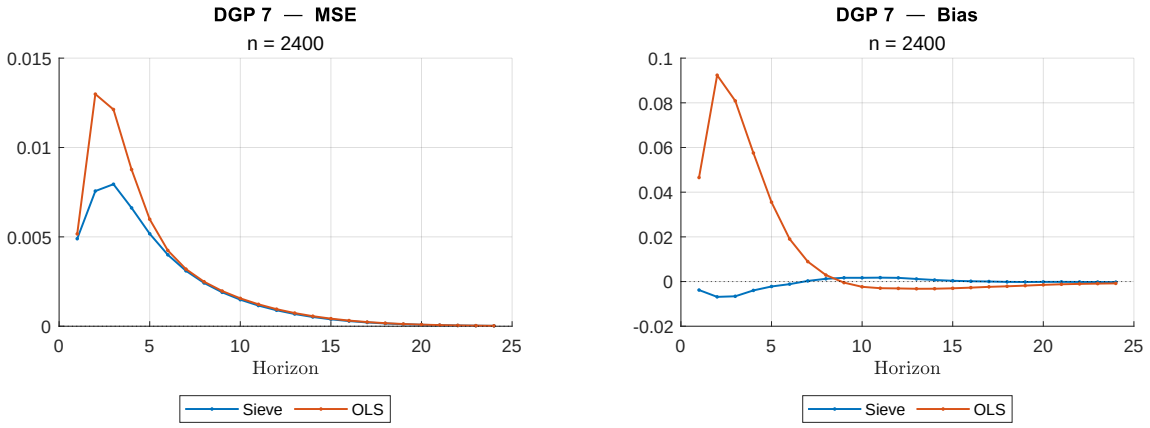


Figure D.2: Simulations results for DGPs 4-6.

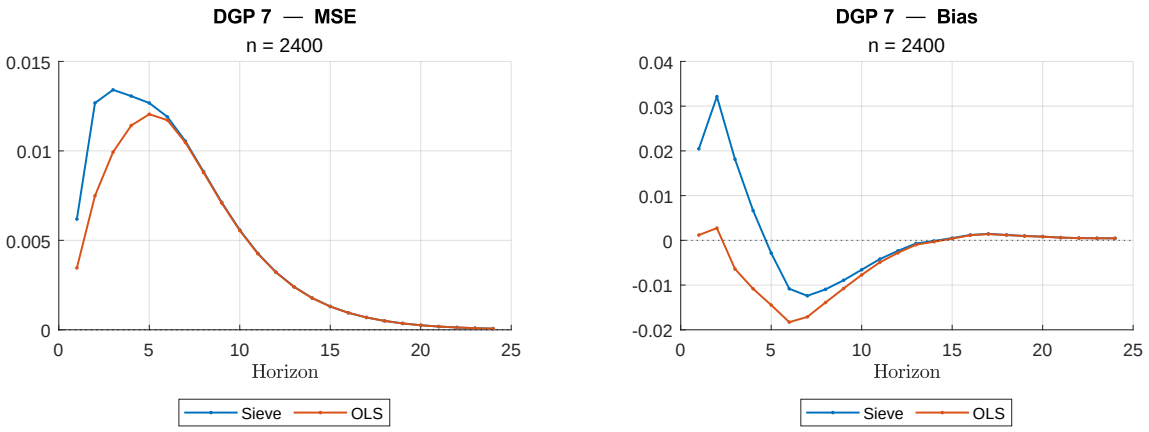


(a) $\delta = +2$

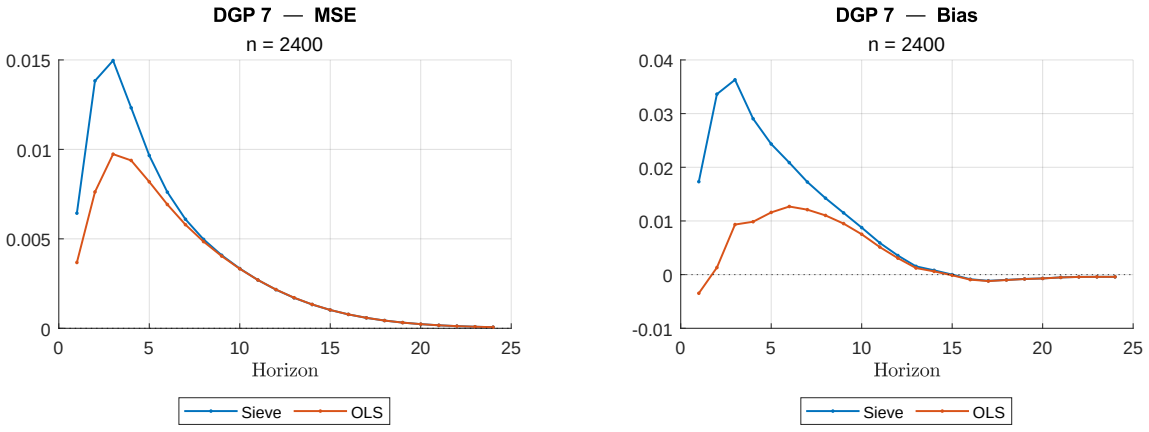


(b) $\delta = -2$

Figure D.3: Simulations results for DGP 7.



(a) $\delta = +2$



(b) $\delta = -2$

Figure D.4: Simulation results for DGP 7 when considering $\widetilde{\varphi}$ in place of φ.

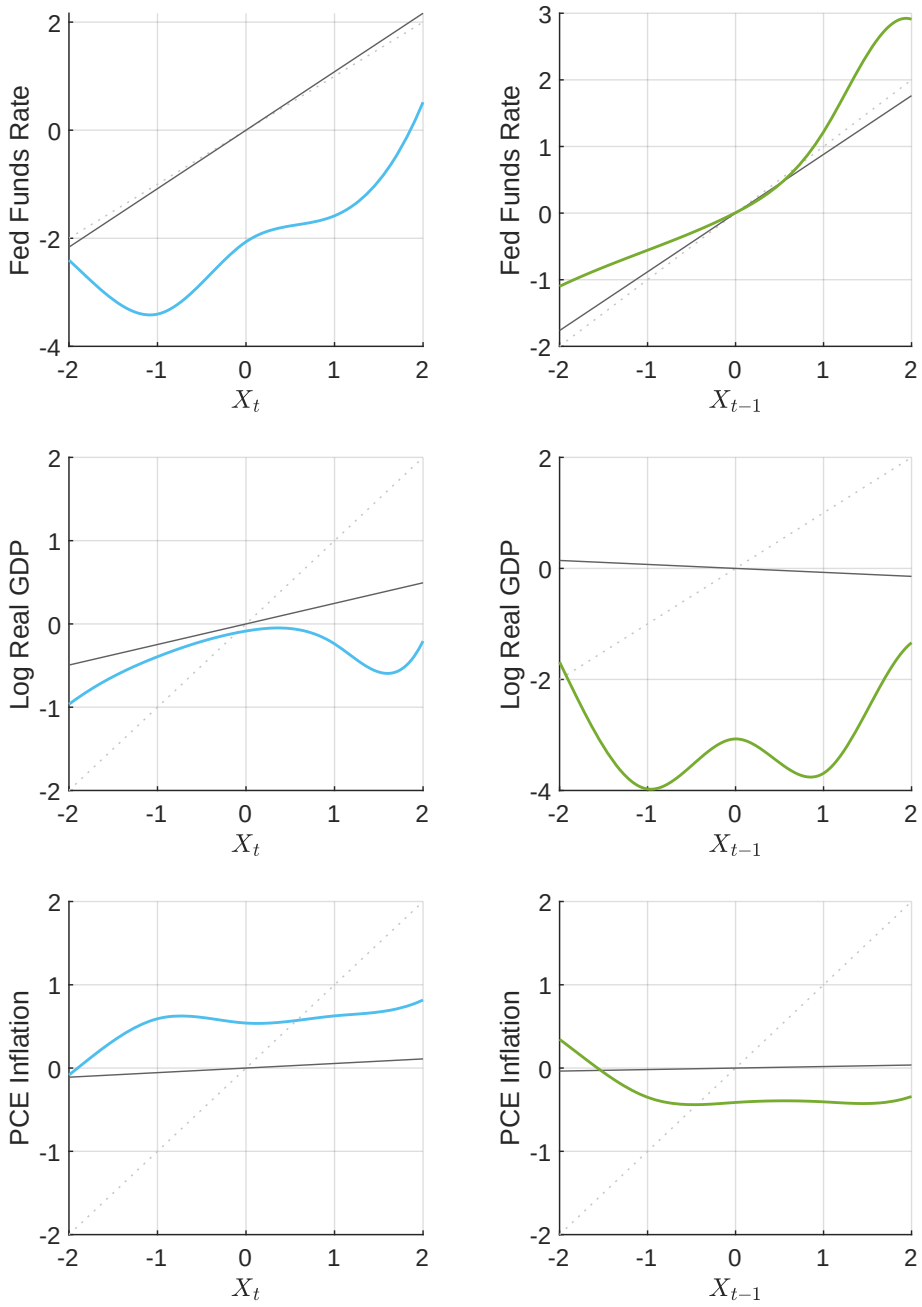
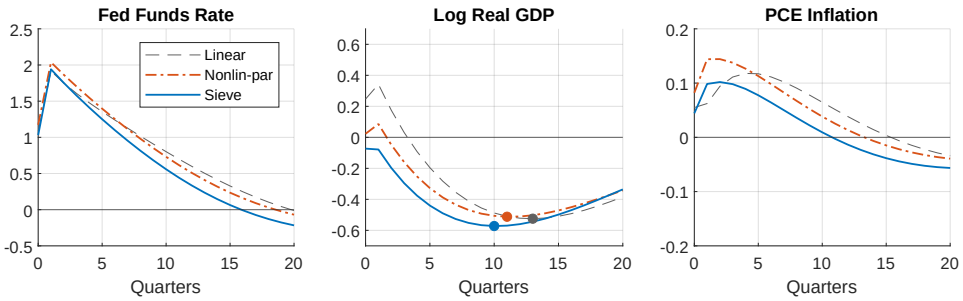
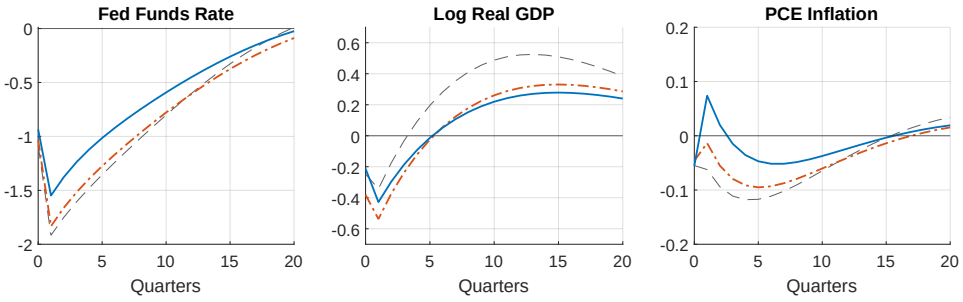


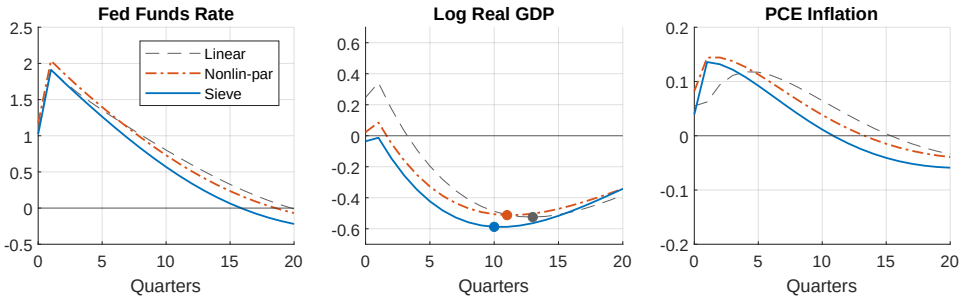
Figure D.5: Estimated nonlinear regression functions for the narrative U.S. monetary policy variable. Contemporaneous (left side) and one-period lag (right side) effects are shown, linear and nonlinear functions. For comparison, linear VAR coefficients (dark gray) and the identity map (light gray, dashed) are shown as lines.



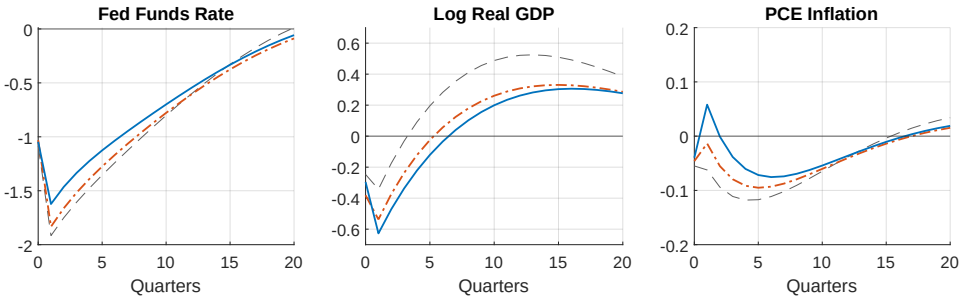
(a) $\delta = +1$, knots at $\{-1, 1\}$



(b) $\delta = -1$, knots at $\{-1, 1\}$



(c) $\delta = +1$, knot at $\{0\}$



(d) $\delta = -1$, knot at $\{0\}$

Figure D.6: Robustness plots for U.S. monetary policy shock when changing knots compared to those used in Figure 3. Note that linear and parametric nonlinear responses do not change.

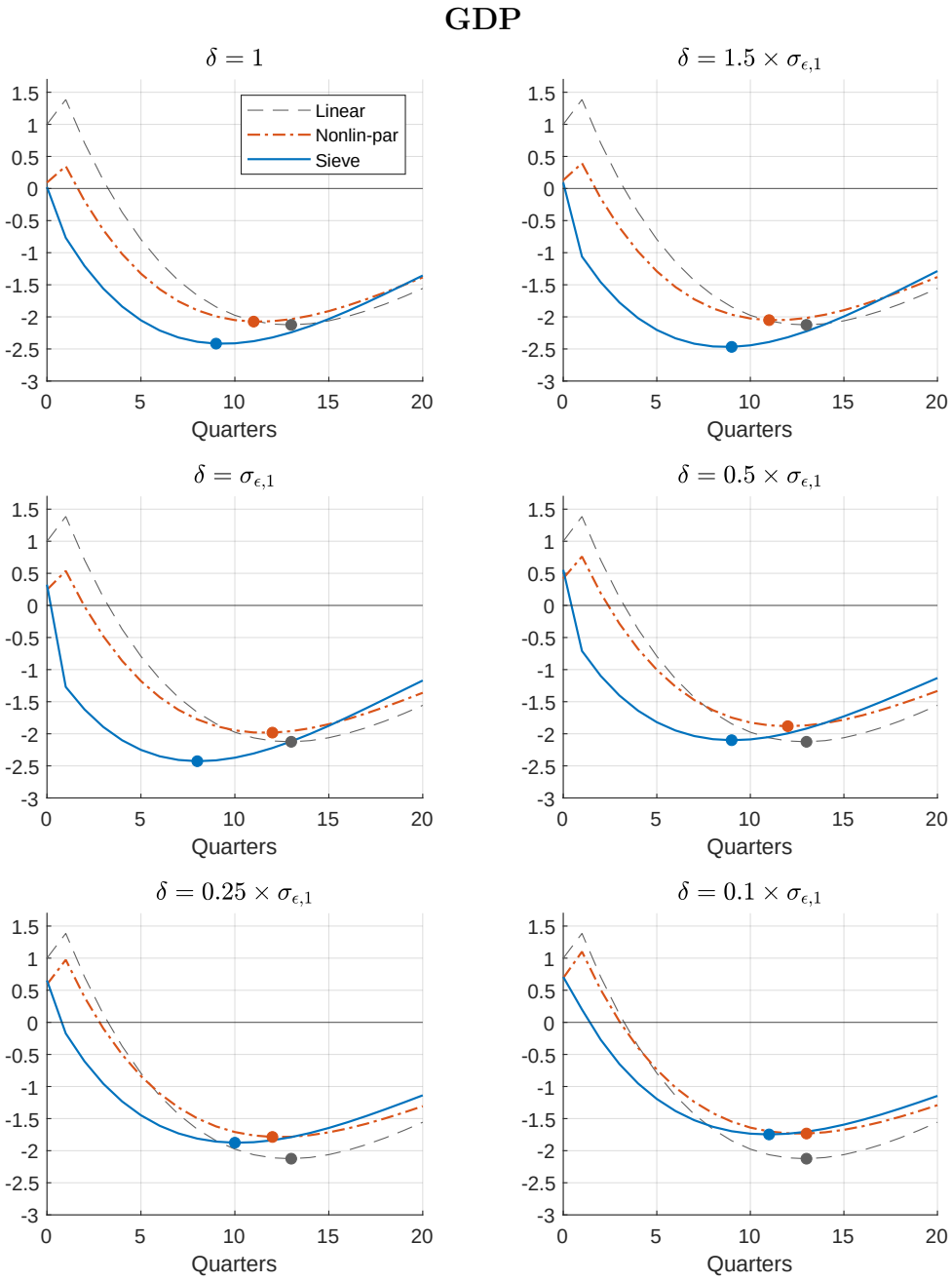


Figure D.7: Relative changes in the GDP impulse responses function when the size of the shock is reduced from that used in Figure 3. The standard deviation of $X_t \equiv \epsilon_{1t}$ is $\sigma_{\epsilon,1} \approx 0.5972$. Linear IRFs are re-scaled such that for all values of δ the linear response at $h = 0$ is one in absolute value. Nonlinear IRFs are re-scaled by δ times the linear response scaling factor.

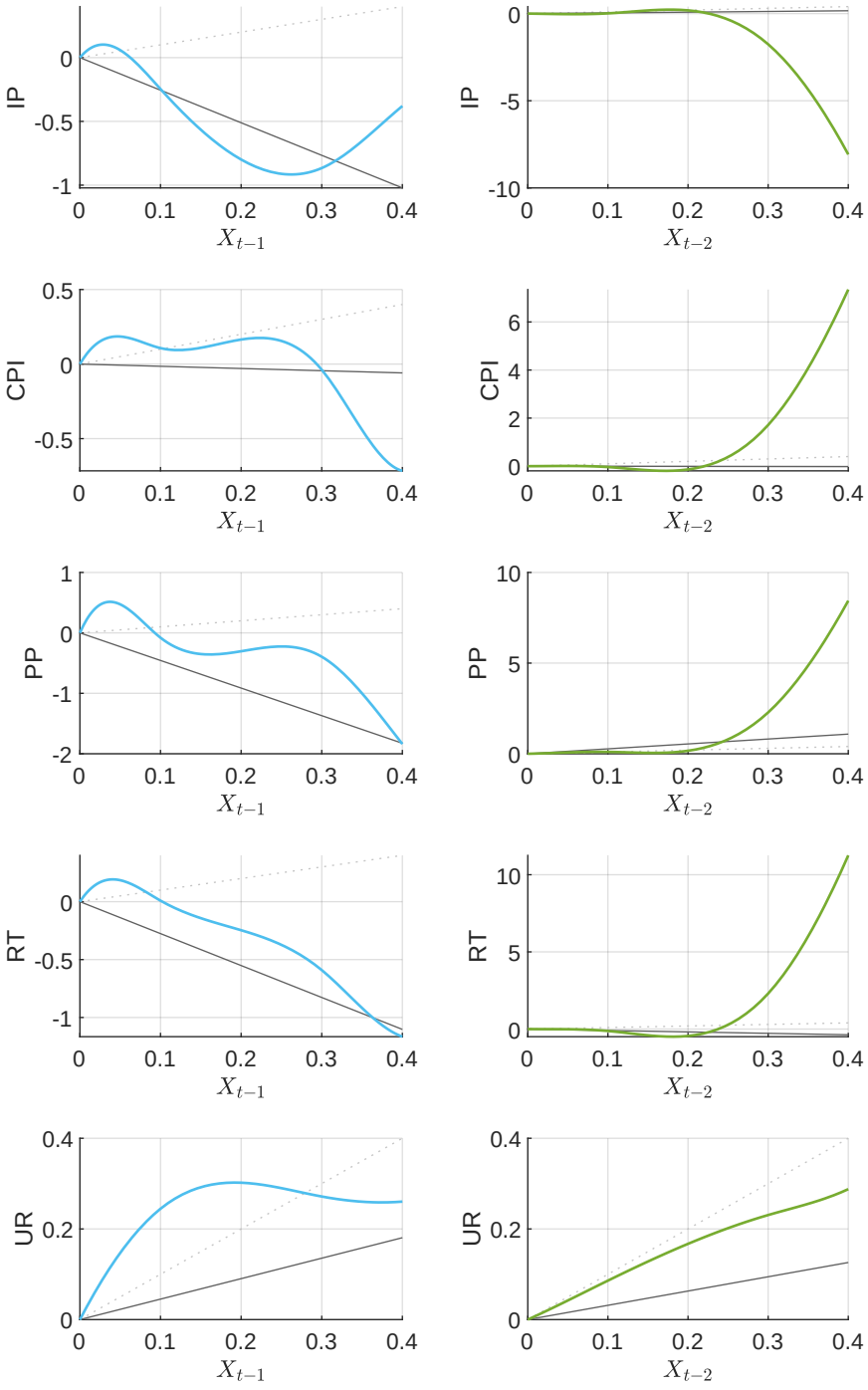
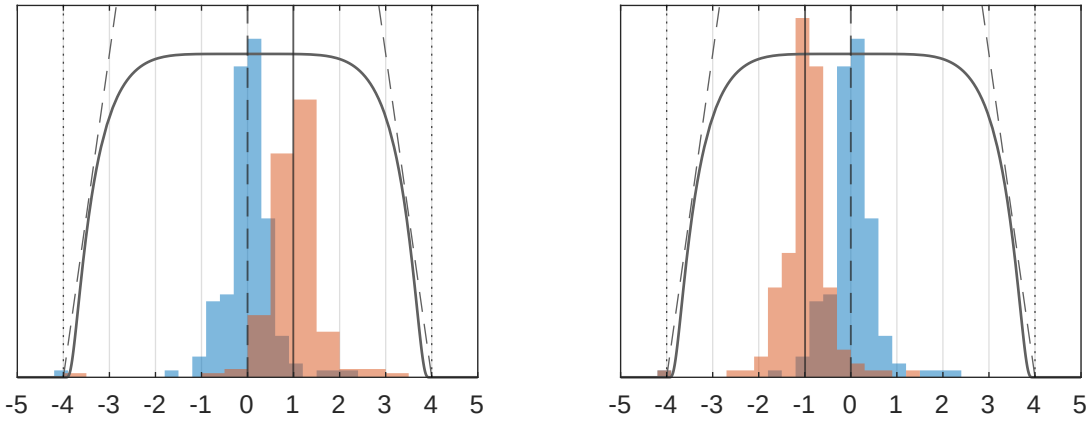
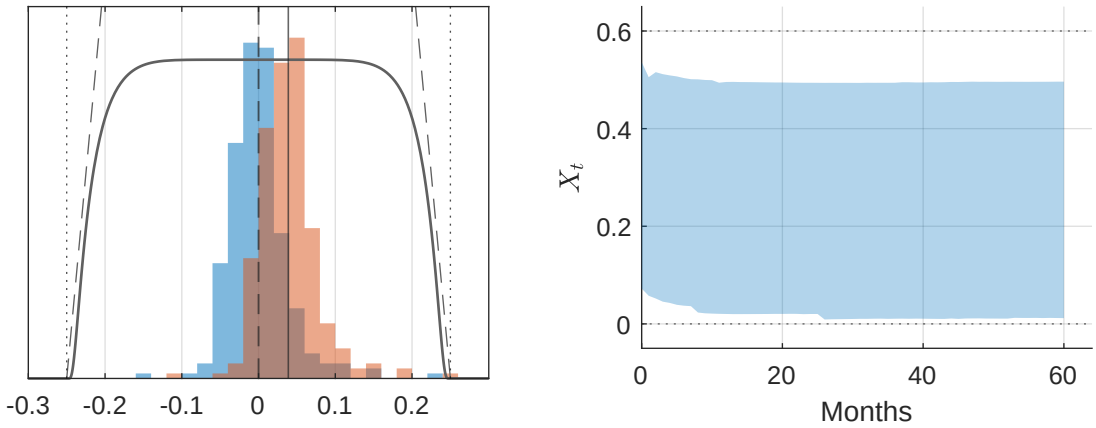


Figure D.8: Estimated nonlinear regression functions for the 3M3M subjective interest rate uncertainty measure. One-period (left side) and two-period lag (right side) effects are shown, combining linear and nonlinear functions. For comparison, linear VAR coefficients (dark gray) and the identity map (light gray, dashed) are shown as lines.



(a) $\delta = +1$ (b) $\delta = -1$

Figure D.9: Comparison of histograms and shock relaxation function for a positive (left) and negative (right) shock in monetary policy. Original (blue) versus shocked (orange) distribution of the sample realization of ϵ_{1t}. The dashed vertical line is the mean of the original distribution, while the solid vertical line is the mean after the shock.



(a) $\delta = \sigma_\epsilon$ (b) Envelope

Figure D.10: Left: Histograms and shock relaxation function for a one-standard-deviation shock in interest rate uncertainty. Original (blue) versus shocked (orange) distribution of the sample realization of ϵ_{1t}. The dashed vertical line is the mean of the original distribution, while the solid vertical line is the mean after the shock. Right: Envelope (min-max) of shocked paths for one-standard-deviation impulse response.

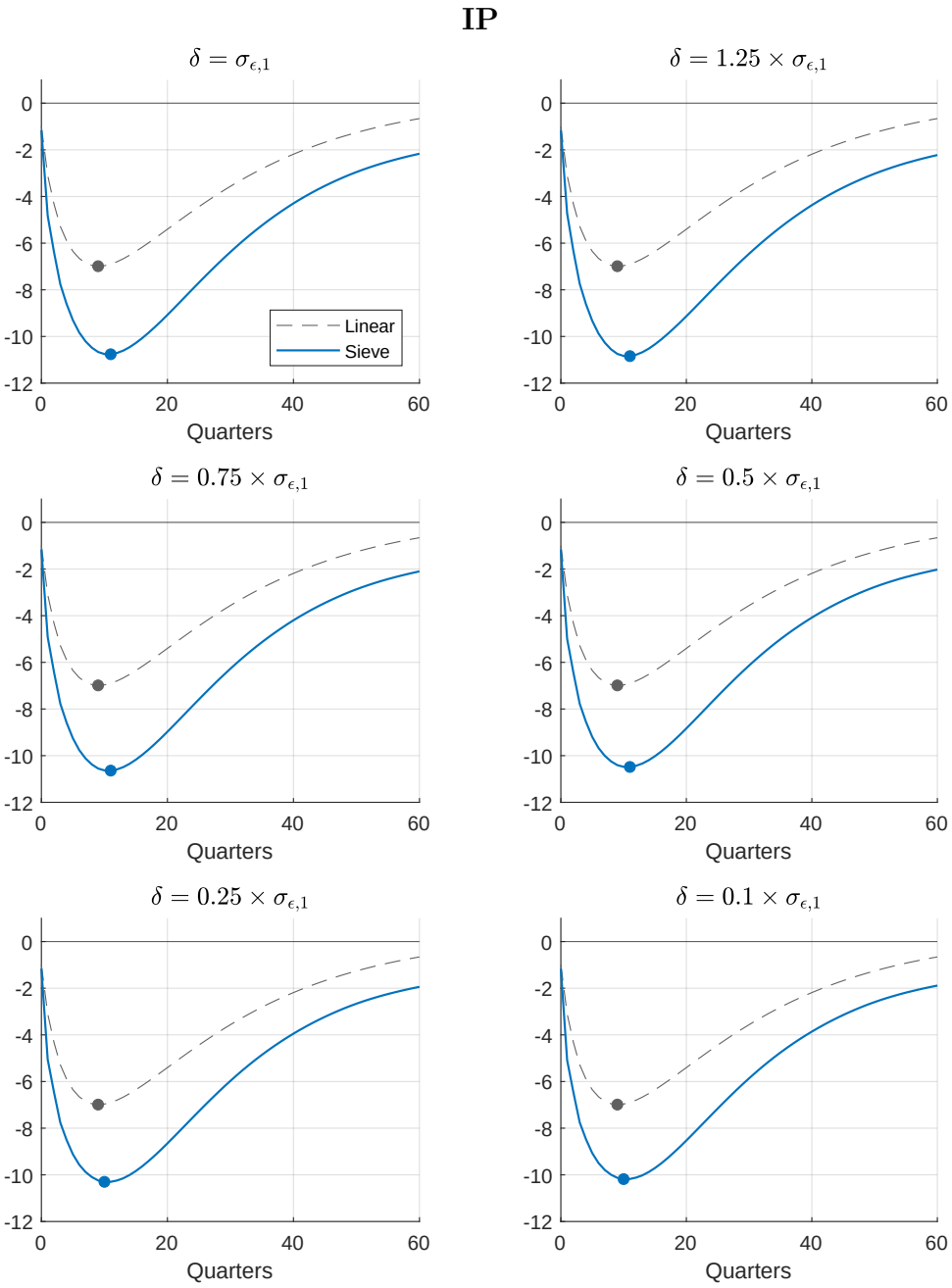


Figure D.11: Relative changes in the industrial production impulse responses function when the size of the shock is reduced from that used in Figure 4. The standard deviation of $\equiv \epsilon_{1t}$ is $\sigma_{\epsilon,1} \approx 0.0389$. Linear IRFs are re-scaled such that for all values of δ the linear response at $h = 0$ is one in absolute value. Nonlinear IRFs are re-scaled by δ times the linear response scaling factor.

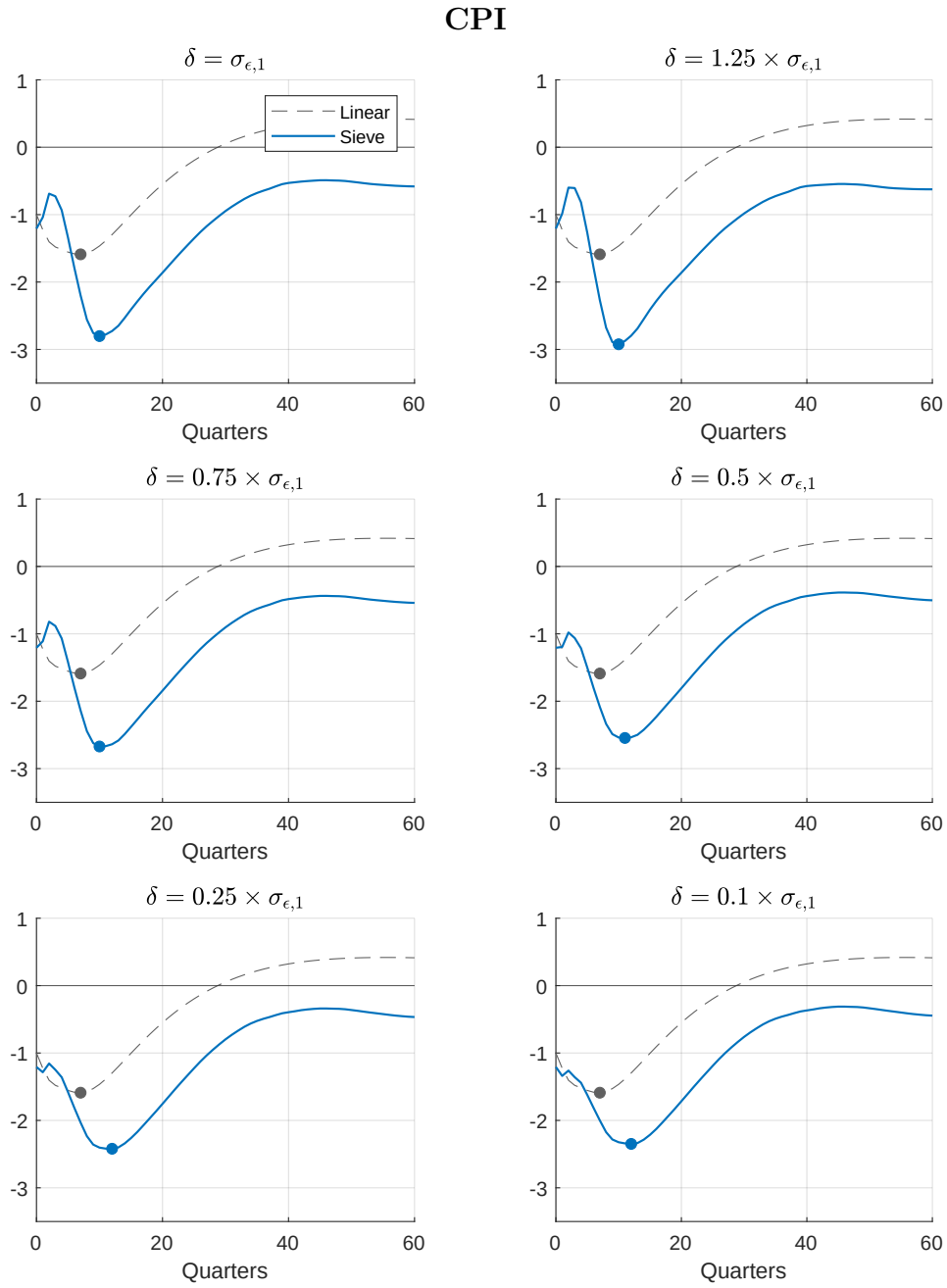


Figure D.12: Relative changes in the CPI impulse responses function when the size of the shock is reduced from that used in Figure 4. The standard deviation of $\equiv \epsilon_{1t}$ is $\sigma_{\epsilon,1} \approx 0.0389$. Linear IRFs are re-scaled such that for all values of δ the linear response at $h = 0$ is one in absolute value. Nonlinear IRFs are re-scaled by δ times the linear response scaling factor.