Gaussian Mixture Approximations of Impulse Responses and The Non-Linear Effects of Monetary Shocks*

Regis Barnichon
CREI, Universitat Pompeu Fabra, CEPR

Christian Matthes
Federal Reserve Bank of Richmond

June 2016

(first draft: March 2014)

Abstract

This paper proposes a new method to estimate the (possibly non-linear) dynamic effects of structural shocks by using Gaussian basis functions to parametrize impulse response functions. We apply our approach to the study of monetary policy and obtain two main results. First, regardless of whether we identify monetary shocks from (i) a timing restriction, (ii) sign restrictions, or (iii) a narrative approach, the effects of monetary policy are highly asymmetric: A contractionary shock has a strong adverse effect on unemployment, but an expansionary shock has little effect. Second, an expansionary shock may have some expansionary effect, but only when the labor market has some slack. In a tight labor market, an expansionary shock generates a burst of inflation and no significant change in unemployment. *JEL classifications: C14, C32, C51, E32, E52*

*We would like to thank Luca Benati, Francesco Bianchi, Christian Brownlees, Fabio Canova, Tim Cogley, Davide Debortoli, Jordi Gali, Yuriy Gorodnichenko, Eleonora Granziera, Oscar Jorda, Thomas Lubik, Jim Nason, Kris Nimark, Mikkel Plagborg-Moller, Giorgio Primiceri, Ricardo Reis, Barbara Rossi, Mark Watson, Yanos Zylberberg and seminar participants at the Barcelona GSE Summer Forum 2014, the 2014 NBER/Chicago Fed DSGE Workshop, William and Mary college, EUI Workshop on Time-Varying Coefficient Models, Oxford, Bank of England, NYU Alumni Conference, Society for Economic Dynamics Annual Meeting (Warsaw), Universitaet Bern, Econometric Society World Congress (Montreal), the Federal Reserve Board, the 2015 SciencesPo conference on Empirical Monetary Economics and the San Francisco Fed for helpful comments. The views expressed here do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. Any errors are our own.

1 Introduction

There now exists a relatively broad consensus on the average effect of monetary policy on economic activity, and it is generally accepted that a monetary contraction (expansion) leads to a decline (increase) in output.

However, there is still little agreement about possible asymmetric or non-linear effects of monetary policy, and two questions at the core of monetary policy making are largely unsettled.[1] First, does monetary policy have asymmetric effects on economic activity? As captured by the string metaphor, does contractionary monetary policy have a much stronger effect –being akin to pulling on a string– than an expansionary shock –being akin to pushing on a string–? Second, does the effect of monetary policy vary with the state of the business cycle? For instance, does the central bank have more room to stimulate economic activity (without raising inflation) during recessions?

Providing answers to these questions has been difficult in part for one important technical reason: the standard approach to identify the dynamic effect of shocks relies on structural Vector-Autoregressions (VARs),[2] which are linear models. While VARs can accommodate certain types of non-linearities, some questions, such as the asymmetric effect of a monetary shock, cannot be answered within a VAR framework.

This paper proposes a new method to estimate the (possibly non-linear) dynamic effects of structural shocks. Instead of assuming the existence of a VAR representation, our approach consists in working directly with the structural moving-average representation of the economy. Then, to make the estimation of the moving-average representation feasible, we parametrize the impulse response functions with Gaussian basis functions.

Our approach builds on two premises: (i) any mean-reverting impulse response function can be approximated by a mixture of Gaussian basis functions, and (ii) a small number (one or two)

[1]For instance, while Cover (1992) finds evidence of asymmetric effects, Ravn and Sola (1996, 2004) and Weise (1999) instead find nearly symmetric effects. And while Lo and Piger (2005) and Santoro et al. (2014) conclude that monetary policy has stronger effects during recessions, Tenreyro and Thwaites (2015) conclude the opposite.

[2]See e.g., Christiano, Eichenbaum, and Evans (1999) and Uhlig (2005).

of Gaussian functions can already capture a large variety of impulse response functions, and notably the typical impulse responses found in empirical or theoretical studies. For instance, the impulse response functions to monetary shocks are often found (or theoretically predicted) to be monotonic or hump-shaped (e.g., Christiano, Eichenbaum and Evans 1999, Walsh 2010). In such cases, a single Gaussian function can already provide an excellent approximation of the impulse response function.

Thanks to the small number of free parameters allowed by a Gaussian Mixture Approximation (GMA), it is possible to directly estimate the structural moving average model from the data, i.e., directly estimate the impulse response functions.[3] In turn, the parsimony of the approach allows us to estimate more general non-linear models.

We conduct a number of Monte-Carlo simulations to illustrate the performance of our approach in finite sample, first for linear models, then for non-linear models. In a linear model, we show that a GMA model can generate more accurate impulse response estimates (in a mean-squared error sense) than a well-specified VAR model. In a simulation with asymmetry and state-dependence, we find that a GMA model can accurately detect the presence of non-linearities and deliver good estimates of the magnitudes of the non-linearities.

We use our GMA approach to estimate the non-linear effects of monetary shocks. Our benchmark identification scheme is a recursive identification scheme, whereby monetary policy shocks can only affect macro variables with a one period lag (Christiano, Eichenbaum and Evans, 1999). However, to emphasize that GMAs can easily accommodate other structural identification schemes, we also consider two alternative identification schemes: (i) a set identification scheme based on sign restrictions,[4] and (ii) a narrative identification scheme where a series of monetary shocks has been previously identified from narrative accounts (Romer and Romer, 2002).

Consistent with the string metaphor, our findings point towards the existence of strong

[3]Another advantage of using Gaussian basis functions is that prior elicitation can be much easier than with Bayesian estimation of standard VARs, because the coefficients to be estimated are directly interpretable as features of impulse responses.

[4]See e.g., Faust (1998), Canova and De Nicolo (2002), Uhlig (2005), Amir Ahmadi and Uhlig (2015).

asymmetries in the effects of monetary shocks, and Bayesian model comparison strongly favors a GMA model with asymmetry over a linear VAR model. Regardless of whether we identify monetary shocks from a recursive ordering, from sign restrictions or from a narrative approach, we find that a contractionary shock has a strong adverse effect on unemployment, larger than implied by linear estimates, while an expansionary shock has little effect on unemployment.[5] Although our evidence for inflation is more uncertain, the behavior of inflation suggests that the asymmetric response of unemployment could be due to the presence downward price/wage rigidities, because inflation displays a more marked price puzzle following a contractionary shock than following an expansionary shock.[6]

We also find that the effect of a monetary shock depends on the state of the business cycle at the time of the intervention: an expansionary shock can have some expansionary effect, but only when the labor market has some slack. In a tight labor market, an expansionary shock generates no significant drop in unemployment but leads to a burst of inflation, consistent with a standard Keynesian narrative.

Although our use of Gaussian basis functions to model and estimate impulse response functions is new in the economics literature, our approach can be cast in the broader context of the machine (supervised) learning literature in that we project the function to be estimated on the space spanned by a dictionary of basis functions (see Hastie, Tibshirani and Friedman, 2009). In basis functions methods, the number of basis functions is often too large for empirical purposes, and the complexity of the model is typically controlled through a combination of restriction, selection and/or regularization methods. Our approach, which consists in using a limited number of basis functions, uses both selection and restriction to control the complexity of the model.[7]

[5]This finding is interesting in the context of the current debate on the appropriate timing of the lift-off of the policy rate from its (close to) zero level in most developed economies. Our estimates suggest that an inappropriate (i.e., too strong or too early) increase in the policy rate could be a lot more costly (in terms of economic activity) than conventional (linear) estimates suggest.

[6]See e.g., Morgan (1993) for a discussion of the effect of downward price rigidity on asymmetric effects of monetary policy.

[7]It uses selection in the sense that our algorithm scans the dictionary of possible basis functions to find the Gaussian basis functions that best fit the data (in a maximum likelihood sense), and it uses restriction in

In economics, our parametrization of impulse responses relates to an older literature on distributed lag models and in particular the Almon (1965) lag specification, in which the successive weights, i.e., the impulse response function in our context, are given by a polynomial function.[8] Our use of Gaussian basis functions relates to a large applied mathematics literature that relies on radial basis functions (of which Gaussian functions are one example) to approximate arbitrary multivariate functions (e.g., Buhmann, 2003) or to approximate arbitrary distributions using a mixture of Gaussian distributions (Alspach and Sorenson 1971, 1972, McLachlan and Peel, 2000). Although Gaussian basis functions provide a more natural and more parsimonious way than polynomials to approximate mean-reverting impulse response functions, our approach is general and other basis functions are possible. For instance, the inverse quadratic function, which is also a popular radial basis function, could be used to parametrize impulse response functions.[9] Finally, our approach shares with the non-parametric econometrics literature (e.g., Racine, 2008) the insight that mixtures of Gaussian kernels can approximate very general shapes, although we use that insight in a very different manner.

The economic literature has so far tackled the estimation of non-linear effects of shocks in two main ways.[10]

A first approach estimates non-linear effects by regressing a variable of interest on contemporaneous and lagged values of some independently identified shocks while allowing for possible non-linear effects. In the context of monetary policy, Cover (1992), DeLong and Summers (1988) and Morgan (1993) identify monetary shocks from unanticipated money innovations (obtained from a money supply process regression, following Barro, 1977) and test whether the impulse response function depends on the sign of these innovations. While that

the sense that we restrict ourselves to the class of impulse response functions that can be generated by a few Gaussian basis functions.

[8]Recently, Plagborg-Moller (2016) proposes a Bayesian method to directly estimate the structural moving-average representation of the data by using prior information about the shape and the smoothness of the impulse response.

[9]In fact, in a different context, Jorgenson (1966) suggested that ratios of polynomials, of which the inverse quadratic function is one example, could be used to parametrize distributed lag functions.

[10]A third non-linear approach was recently proposed by Angrist et al. (2013) who develop a semi-parametric estimator to evaluate the (possibly asymmetric) effects of monetary policy interventions. They find asymmetric effects of monetary shocks consistent with our findings.

approach was later abandoned because money supply regressions were suspected to poorly identify monetary shocks, the use of independently identified shocks has been recently revived thanks to the use of narratively identified shocks (Romer and Romer, 2002) and thanks to the Local Projection method pioneered by Jorda (2005).[11] The narrative approach was precisely developed in order to identify exogenous monetary innovations, and Jorda's method can easily accommodate non-linearities in the response function.[12] However, the Local Projection method is limited by efficiency considerations. Indeed, while the Local Projection approach is intentionally model-free –not imposing any underlying dynamic system–, this can come at an efficiency cost (Ramey, 2012), which makes inferences on a rich set of non-linearities (e.g., sign- and state-dependence) difficult. In contrast, by positing that the response function can be approximated by one (or a few) Gaussian functions, our approach imposes strong dynamic restrictions between the parameters of the impulse response function, which in turn allow us to estimate a rich set of non-linearities.[13] Another advantage of our approach is that it can be used for model selection and model evaluation through marginal data density comparisons.

A second strand in the literature has relied on regime-switching VAR models –notably threshold VARs (e.g., Hubrich and Terasvirta, 2013) and Markov-switching VARs (Hamilton, 1989)– to capture certain types of non-linearities.[14,15] However, while regime-switching VARs can capture state dependence (whereby the value of some state variable affects the impulse response functions), they cannot capture asymmetric effects of shocks (whereby the impulse response to a structural shock depends on the sign of that shock). Indeed, with regime-

[11]The combination of Jorda's method with narratively identified shocks was first introduced in the context of fiscal policy by Auerbach and Gorodnichenko (2013) in order to test for the existence of state dependence in the effects of fiscal policy.

[12]Santoro et al. (2014) and Tenreyro and Thwaites (2013) use the Jorda method to estimate the extent of state dependence in the effect of monetary policy.

[13]Naturally, this statement also implies that our results are valid under the assumption that response functions can be well approximated by a few Gaussian functions. In this respect, our approach is best seen as complementing the model-free approach of Jorda (2005).

[14]For examples in the monetary policy literature, see Beaudry and Koop (1993), Thoma (1994), Potter (1995), Kandil (1995), Koop, Pesaran and Potter (1996), Koop and Potter, (1998), Ravn and Sola (1996, 2004), Weise (1999), Lo and Piger (2005).

[15]Another prominent class of non-linear VARs includes models with time-varying coefficients and/or time-varying volatilities (e.g., Primiceri, 2005).

switching VAR models, it is assumed that the economy can be in a finite number of regimes, and that each regime corresponds to a different set of VAR coefficients. However, if the true data generating process features asymmetric impulse responses, a new set of VAR coefficients would be necessary each period, because the (non-linear) behavior of the economy at any point in time depends on all structural shocks up to that point. As a result, such asymmetric data generating process cannot generally be approximated by a small number of state variables such as in threshold VARs or Markov-switching models. In contrast, by working directly with the structural moving-average representation, GMA models can easily capture asymmetric impulse response functions (as well as state dependence).

Section 2 describes how we approximate impulse responses using mixtures of Gaussians, Section 3 discusses the key steps of the estimation methodology; Section 4 generalizes our approach to non-linear models; Section 5 presents Monte Carlo simulations to evaluate the performance of our approach in finite sample, first for linear models, then for non-linear models; Section 6 applies GMA to the study of the non-linear effects of monetary shocks using US data; Section 7 concludes.

2 Gaussian Mixture Approximations

This section presents a new method to estimate impulse responses using Gaussian Mixture Approximations (GMA) of the structural moving-average representation of the economy. Although the use of GMAs was motivated in the introduction by the need to model and estimate certain types of non-linearities, the intuition and benefits of GMA models can be understood in a linear context, and this section introduces GMAs in a linear context. We postpone the modeling and estimation of non-linearities to Section 4.

2.1 A structural moving average representation

Our starting point is a structural moving-average model of the economy, in which the behavior of a system of macroeconomic variables is dictated by its response to past and present structural shocks. Specifically, denoting $\boldsymbol{y_t}$ an $L \times 1$ vector of stationary macroeconomic variables, the economy is described by

$$\boldsymbol{y_t} = \sum_{k=0}^{K} \boldsymbol{\Psi}_k \boldsymbol{\varepsilon_{t-k}} \tag{1}$$

where boldface letters indicate vectors or matrices, $\boldsymbol{\varepsilon_t}$ is the vector of structural innovations with $E\boldsymbol{\varepsilon_t} = \mathbf{0}$ and $E\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_t}' = \mathbf{I}$, and K is the number of lags, which can be finite or infinite. Throughout the text, we omit the intercepts for ease of exposition, but all estimated models include intercepts. The matrices $\{\boldsymbol{\Psi}_k\}_{k=0}^{K}$ capture the impulse responses to shocks, and as a normalization, we posit that $\boldsymbol{\Psi}_0$ has positive entries on the diagonal, i.e., $\boldsymbol{\Psi}_{0,\ell\ell} \geq 0$, $\forall \ell \in \{1, .., L\}$. For now, the model is linear, and the $\boldsymbol{\Psi}_k$ matrices are fixed.

If (1) is invertible and admits a VAR representation, the model can be estimated from a VAR on $\boldsymbol{y_t}$ (provided some structural identifying assumption, such as the recursive ordering of $\boldsymbol{\Psi}_0$). However, assuming the existence of a VAR representation can be restrictive. In particular, in a non-linear world where $\boldsymbol{\Psi}_k$ depends on the value of $\boldsymbol{\varepsilon_{t-k}}$ (for instance, when the impulse response function varies with the sign of the shock), the existence of a VAR is compromised. Thus, in this paper, we propose an alternative method that side-steps the need to invert (1), i.e., we propose a method that side-steps the need for a VAR representation.

2.2 Gaussian Mixture Approximations of impulse response functions

Rather than looking for a VAR representation of the dynamic system (1), our aim is to directly estimate (1), the moving-average representation of the economy. Because the number of free parameters $\{\boldsymbol{\Psi}_k\}_{k=0}^{K}$ in (1) is very large or possibly infinite, our strategy consists in parameterizing the impulse response functions, and more precisely in using mixtures of Gaussian functions to approximate each impulse response function.

2.2.1 Theoretical background

Our parametrization of the impulse response functions builds on the following theorem, which states that any integrable function can approximated with a sum of Gaussian functions.

Theorem 1 *Let f be a bounded continuous function on $\mathbb{R}$ that satisfies $\int_{-\infty}^{\infty} f(x)^2 dx < \infty$. There exists a function f_N defined by*

$$f_N(x) = \sum_{n=1}^{N} a_n e^{-(\frac{x-b_n}{c_n})^2}$$

with a_n, b_n, $c_n \in \mathbb{R}$ for $n \in \mathbb{N}$, such that the sequence $\{f_N\}$ converges pointwise to f on every interval of $\mathbb{R}$.

Proof. See Appendix. ■

Denote $\psi(k)$ a representative element of matrix $\mathbf{\Psi}_k$, so that $\psi(k)$ is the value of the impulse response function ψ at horizon k.

Motivated by Theorem 1, our approach will consist in approximating the impulse response function ψ with a sum of Gaussian functions, that is

$$\psi(k) \simeq \sum_{n=1}^{N} a_n e^{-(\frac{k-b_n}{c_n})^2}, \quad \forall k \in (0, K] \tag{2}$$

with a_n, b_n, $c_n \in \mathbb{R}$.[16]

Since our strategy consists in approximating impulse response functions with mixtures of Gaussians, we refer to this class of models as Gaussian Mixture Approximations (GMA), with a $GMA(N)$ denoting a GMA with N Gaussian basis functions.

2.2.2 Intuition and Motivation

Before describing the estimation of GMA models, it is instructive to first intuitively discuss the benefits of our approach over traditional VARs.

The advantage of our approach, and its use for studying the (possibly non-linear) effects of policy, will rest on the fact that, in practice, only a very small number of Gaussian basis

[16] The GMA parametrization of ψ may or may not include the contemporaneous impact coefficient, that is one may choose to use the approximation (2) for $k > 0$ or for $k \geq 0$. In this paper, we treat $\psi(0)$ as a free parameter for additional flexibility.

functions are needed to approximate a typical impulse response function, allowing for efficiency gains and opening the door to estimating non-linearities.

Intuitively, impulse response functions of stationary variables are often found (or theoretically predicted) to be monotonic or hump-shaped (e.g., Christiano, Eichenbaum, and Evans, 1999).[17] In such cases, a single Gaussian function can already provide a good approximate description of the impulse response. To illustrate this observation, Figure 1 plots the impulse response functions of unemployment, the price level and the fed funds rate to a monetary shock estimated from a standard VAR specification,[18] along with the corresponding $GMA(1)$, the Gaussian approximations with only *one* Gaussian function, i.e., using the approximation

$$\psi(k) \simeq a e^{-\frac{(k-b)^2}{c^2}}. \tag{3}$$

We can see that a $GMA(1)$ already does a good job at capturing the impulse responses implied by the VAR.[19] With a $GMA(2)$, the impulse responses are virtually on top on those of the VAR (Figure 1). For illustration, Figure 2 plots the Gaussian basis functions used for each impulse response in the GMA(2) case.

In both cases, the number of free parameters is manageable. For instance, in this 3 variables example, a GMA(1) only has 27 parameters (9 impulse responses times 3 parameters per impulse response, ignoring intercepts) to capture the whole set of impulse responses $\{\boldsymbol{\Psi}_k\}_{k=1}^K$, while a GMA(2) has 48 free parameters $(9 * 3 * 2 = 48)$.[20]

This relatively small number of free parameters in turn allows us to directly estimate the impulse response functions from the vector moving-average representation (1). This point is at the core of our GMA approach, because being able to directly work with the moving-average

[17] In New-Keynesian models, the impulse response functions are generally monotonic or hump-shaped (see e.g., Walsh, 2010).

[18] See Section 6 for the exact specification of the SVAR behind Figure 1. The VAR is specified with unemployment, PCE inflation and the fed funds rate. The impulse response for the price level is calculated from the response of inflation.

[19] In Figure 1, the parameters of the GMA (the a, b and c coefficients) were set to minimize the discrepancy (sum of squared residuals) between the two sets of impulse responses.

[20] For comparison, a corresponding quarterly VAR with 3 variables and 4 lags has $4 * 3^2 = 36$ free parameters, and a monthly VAR with 12 lags has $12 * 3^2 + 6 = 108$ free parameters.

representation will allow us to estimate models in which shocks can have non-linear effects.

To conclude this intuition section, we comment on a particularly interesting case: the $GMA(1)$ model, which has two additional advantages: (i) ease of interpretation, and (ii) ease of prior elicitation.

In a $GMA(1)$ model like (3), the a, b and c coefficients can be easily interpreted, because the impulse response function is summarized by three parameters –the peak effect, the time to peak effect, and the persistence of the impulse response–, which are generally considered the most relevant characteristics of an impulse response function.[21] As illustrated in Figure 3, parameter a is the height of the impulse-response, which corresponds to the maximum effect of a unit shock, parameter b is the timing of this maximum effect, and parameter c captures the persistence of the effect of the shock, as the amount of time τ required for the effect of a shock to be 50% of its maximum value is given by $\tau = c\sqrt{\ln 2}$.

Then, the ease of interpretation of the a, b and c parameters in turn makes prior elicitation easier than in standard VARs, in which the VAR coefficients have a less direct economic interpretation.

3 Bayesian estimation

To estimate our model, we use a Bayesian approach, which is particularly well suited for models that only approximate the true DGP (Fernandez-Villaverde and Rubio-Ramirez, 2004). In particular, Bayes factors will allow us to evaluate GMA models against VAR models, even though the two classes of models are non-nested.[22] Bayesian model comparison will also offer us a natural way to select the order of the GMA model, i.e., the number of Gaussian basis functions used in the approximation.

In this section, we describe the implementation and estimation of GMA models. We first

[21]For instance, when comparing the effects of monetary shocks across different specifications, Coibion (2012) focuses on the peak effect of the monetary shock, which in a GMA(1) model is simply parameter a.

[22]Bayes factors are functions of the marginal data densities for the two models that are being compared. Since marginal data densities can be rewritten as products of one-step ahead forecast densities, Bayes factors also offer insights about the relative forecasting abilities of the two models that are being compared.

describe how we construct the likelihood function by exploiting the prediction-error decomposition, discuss structural identification, then present the estimation routine based on a multiple-block Metropolis-Hasting algorithm, discuss prior elicitation, the determination of the order of the GMA and identification issues related to fundamentalness. We conclude by discussing how to deal with non-stationary data.

3.1 Constructing the likelihood function

We now describe how to construct the likelihood function $p(\boldsymbol{y^T}|\boldsymbol{\theta})$ of a sample of size T for the moving-average model (1) with parameter vector $\boldsymbol{\theta}$ and where a variable with a superscript denotes the sample of that variable up to the date in the superscript.

To start, we use the prediction error decomposition to break up the density $p(\boldsymbol{y^T}|\boldsymbol{\theta})$ as follows:[23]

$$p(\boldsymbol{y^T}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\boldsymbol{y_t}|\boldsymbol{\theta}, \boldsymbol{y^{t-1}}). \tag{4}$$

To calculate the one-step-ahead conditional likelihood function needed for the prediction error decomposition, we assume that all innovations $\{\boldsymbol{\varepsilon_t}\}$ are Gaussian with mean zero and variance one,[24] and we note that the density $p(\boldsymbol{y_t}|\boldsymbol{\theta}, \boldsymbol{y^{t-1}})$ can be re-written as $p(\boldsymbol{y_t}|\boldsymbol{\theta}, \boldsymbol{y^{t-1}}) = p(\boldsymbol{\Psi}_0\boldsymbol{\varepsilon_t}|\boldsymbol{\theta}, \boldsymbol{y^{t-1}})$ since

$$\boldsymbol{y_t} = \boldsymbol{\Psi}_0\boldsymbol{\varepsilon_t} + \sum_{k=1}^{K} \boldsymbol{\Psi}_k\boldsymbol{\varepsilon_{t-k}}. \tag{5}$$

Since the contemporaneous impact matrix is a constant, $p(\boldsymbol{\Psi}_0\boldsymbol{\varepsilon_t}|\boldsymbol{\theta}, \boldsymbol{y^{t-1}})$ is a straightforward function of the density of $\boldsymbol{\varepsilon_t}$.

To recursively construct $\boldsymbol{\varepsilon_t}$ as a function of $\boldsymbol{\theta}$ and $\boldsymbol{y^t}$, we need to uniquely pin down the values of the components of $\boldsymbol{\varepsilon_t}$ from equation (5), that is we need that $\boldsymbol{\Psi}_0$ is invertible. We impose this restriction by assigning a minus infinity value to the likelihood whenever $\boldsymbol{\Psi}_0$ is not invertible. It is also at this stage that we impose the identifying restriction that we describe

[23]To derive the conditional densities in decomposition (4), our parameter vector $\boldsymbol{\theta}$ thus implicitly also includes the K initial values of the shocks: $\{\boldsymbol{\varepsilon}_{-K}...\boldsymbol{\varepsilon}_0\}$. We will keep those fixed throughout the estimation and discuss alternative initializations below.

[24]The estimation could easily be generalized to allow for non-normal innovations such as t-distributed errors.

next. Finally, to initialize the recursion, we set the first K innovations $\{\boldsymbol{\varepsilon_j}\}_{j=-K}^{0}$ to zero.[25,26]

3.2 Structural identifying assumptions

Model (1) is under-identified without additional restrictions. In our application of GMAs to the study of monetary policy, we will use as our benchmark a recursive identification scheme (Christiano, Eichenbaum and Evans, 1999). However, to emphasize that GMAs can easily accommodate other structural identification schemes, we will also consider two popular schemes to identify monetary shocks: (i) the narrative identification scheme where a series of monetary shocks has been previously identified from narrative accounts (Romer and Romer, 2002), and (ii) a set identification scheme based on sign restrictions (Uhlig, 2005).[27] We describe the implementation of these identification schemes next.

Short-run restrictions Short-run restrictions consist in restrictions on $\boldsymbol{\Psi}_0$, which are straightforward to implement in a GMA model.

Short-run restrictions in a fully identified model consists in imposing $\frac{L(L-1)}{2}$ restrictions on $\boldsymbol{\Psi}_0$ (of dimension $L \times L$), and a common approach is to impose that $\boldsymbol{\Psi}_0$ is lower triangular, so that the different shocks are identified from a timing restriction. This identifying scheme is popular in the case of monetary policy, where monetary shocks are assumed to only affect macro variables with a one period lag (Christiano, Eichenbaum and Evans, 1999).

In a partially identified model, one can impose a timing restriction for one shock only. In the case of the monetary model considered in section 6, this will amount to ordering the monetary policy variable last and imposing that $\boldsymbol{\Psi}_0$ has its last column filled with 0 except for the diagonal coefficient. The submatrix $\tilde{\boldsymbol{\Psi}}_0$ made of the first $(L-1)$ rows and $(L-1)$ columns of $\boldsymbol{\Psi}_0$ is then left unrestricted, apart from invertibility to ensure that equation (5) defines a unique shock vector $\boldsymbol{\varepsilon}_t$ (as described in section 3.1).

[25] Alternatively, we could use the first K values of the shocks recovered from a structural VAR.

[26] When K, the lag length of the moving average (1), is infinite, we truncate the model at some horizon K, large enough to ensure that the lag matrix coefficients $\boldsymbol{\Psi}_K$ are "close" to zero. Such a K exists since the variables are stationary.

[27] In Barnichon and Matthes (2016), we discuss how to impose other identification schemes.

Narrative identification In a narrative identification scheme, a series of shocks has been previously identified from narrative accounts. For that case, we can proceed as with the recursive identification, because the use of narratively identified shocks can be cast as a partial identification scheme. If one orders the narratively identified shocks series first in $\boldsymbol{y}_t$, we can assume that $\boldsymbol{\Psi}_0$ has its first row filled with 0 except for the diagonal coefficient, which implies that the narratively identified shock does not react contemporaneously to other shocks (as should be the case if the narrative shocks were correctly identified).

Sign restrictions Set identification through sign restrictions consists in imposing sign-restrictions on the sign of the $\boldsymbol{\Psi}_k$ matrices, i.e., the impulse response coefficients at different horizons. Again, because a GMA model works directly with the moving average representation and the $\boldsymbol{\Psi}_k$ matrices, imposing sign-restrictions is straightforward to implement in a GMA model. One can impose sign-restrictions on only the impact coefficients (captured by $\boldsymbol{\Psi}_0$, which could be left as a free parameter in this case) and/or sign restrictions on the impulse response over a specific horizon (captured by the $\{a_n, b_n, c_n\}$ GMA coefficients that model $\boldsymbol{\Psi}_k$). To implement parameter restrictions on $\boldsymbol{\Psi}_0$ and/or $\{a_n, b_n, c_n\}$, we assign a minus infinity value to the likelihood whenever the restrictions are not met.

More generally, in line with the insights from Baumeister and Hamilton (2015), the implementation of sign-restrictions can take the form of priors on the coefficients of $\boldsymbol{\Psi}_0$ and on the $\{a_n, b_n, c_n\}_{n=1}^{N}$ coefficients.[28]

3.3 Estimation routine

To estimate our model, we use a Metropolis-within-Gibbs algorithm (Robert & Casella 2004, Haario et al., 2001) with the blocks given by the different groups of parameters in our model

[28]More generally, because GMAs work directly with the structural moving-average representation, the parameters to be estimated can be interpreted as "features" of the impulse responses, and one could envision set identification schemes through shape restrictions (see e.g., Lippi and Reichlin, 1994 for an early application of this idea). For instance, one could posit priors on the location of the peak effect, posit priors on the persistence of the effect of the shock, among other possibilities. See Plagborg-Moller (2016) for a related idea.

(there is respectively one block for the a parameters, one block for the b parameters, one block for the c parameters and one block for the constant and other parameters).

To initialize the Metropolis-Hastings algorithm in an area of the parameter space that has substantial posterior probability, we follow a two-step procedure: first, we estimate a standard VAR using OLS on our data set, calculate the moving-average representation, and we use the impulse response functions implied by the VAR as our starting point. More specifically, we calculate the parameters of our GMA model to best fit the VAR-based impulse response functions.[29] Second, we use these parameters as a starting point for a simplex maximization routine that then gives us a starting value for the Metropolis-Hastings algorithm.

3.4 Prior elicitation

We use (loose) Normal priors centered around the impulse response functions obtained from the benchmark (linear) VAR. Specifically, we put priors on the a, b and c coefficients that are centered on the values for a, b and c obtained by matching the impulse responses obtained from the VAR, as described in the previous paragraph.

Specifically, denote $a^0_{ij,n}$, $b^0_{ij,n}$ and $c^0_{ij,n}$, $n \in \{1, N\}$ the values implied by fitting the GMA(N) to the VAR-based impulse response of variable i to shock j. The priors for $a_{ij,n}$, $b_{ij,n}$ and $c_{ij,n}$ are centered on $a^0_{ij,n}$, $b^0_{ij,n}$ and $c^0_{ij,n}$, and the corresponding standard-deviations are set as follows: $\sigma_{ij,a} = 10$, $\sigma_{ij,b} = K$ and $\sigma_{ij,c} = K$ (recall that K is the length of the moving-average).[30] While there is clearly some arbitrariness in choosing the tightness of our priors, it is important to note that they are sufficiently loose to let us explore a large class of alternative specifications.[31] More generally, the use of informative priors is not critical for

[29]Specifically, we set the parameters of our model (the a, b and c coefficients) to minimize the discrepancy (sum of squared residuals) between the two sets of impulse responses.

[30]Going back to our intuitive interpretation of the three parameters of a Gaussian basis function in Section 2, note that these priors are very loose. This is easy to see for a and b. For c, recall that $c\sqrt{\ln 2}$ is the the half-life of the effect of a shock. If $c = K$, this already corresponds to very persistent impulse response functions, since $K\sqrt{\ln 2} = 38$ quarters.

[31]For our monetary policy application, we verified that the prior did not influence our conclusions by using uninformative priors: We estimated both the asymmetric GMA model and the asymmetric and state dependent GMA model with improper flat priors, and we obtained very similar results.

our approach, and we could have used improper uniform priors, but the use of proper priors allows us to compute posterior odds ratios, which are important to select the order of the moving-average and to compare different GMA models.

3.5 Choosing N, the number of Gaussian basis functions

To choose N, the order of the GMA model, we use posterior odds ratios (assigning equal probability to any two model) to compare models with increasing number of mixtures. We select the model with the highest posterior odds ratio.[32]

3.6 Fundamentalness

In a linear moving average model, different representations (i.e., different sets of coefficients and innovation variances) can exhibit the same first two moments, so that with Gaussian-distributed innovations, the likelihood can display multiple peaks, and the moving average model is inherently underidentified. Since a GMA model works off directly with the moving-average representation, it cannot distinguish between invertible (also called "fundamental") and non-invertible representations. By using the VAR-based impulse responses as starting values, we implicitly focus on the invertible part of the parameter space.[33,34]

[32]This approach can be seen as analogous to the choice of the parameter lag in VAR models. While the Wold theorem shows that any covariance-stationary series can be written as a VAR(∞), one must select a finite lag order p that reasonable approximate the VAR(∞) (e.g., Canova, 2007). The usual approach is to use information criteria such as AIC and BIC, which is similar to our present approach. Just as in the case of lag length choice in a VAR (where this is rarely, if ever, done), we could alternatively treat N as a discrete parameter. We choose to use one value for N at a time to highlight how different choices for N affect estimated impulse responses.

[33]Since a VAR is obtained by inverting the fundamental moving-average representation, it automatically selects the fundamental representation (e.g., Lippi and Reichlin, 1994).

[34]An alternative estimation procedure to handle both invertible and non-invertible representations would be to use the Kalman filter with priors on the K initial values of the shocks $\{\varepsilon_{-K}...\varepsilon_0\}$, as recently proposed by Plagborg-Moller (2016). However, unlike our proposed approach, this procedure would be difficult to implement in non-linear models. Note also that the non-uniqueness of the moving average representation was proven for linear models (under Gaussian shocks). When we consider non-linearities, the non-uniqueness of the moving-average representation is not guaranteed anymore, and identification may be easier. In practice (and in Monte-Carlo simulations), the likelihood did not display multiple peaks when we allowed for asymmetry or state-dependence.

3.7 Dealing with non-stationary data

As can be seen from Theorem 1, GMA models can only capture impulse response functions that are bounded and integrable, which restricts our approach to stationary series. If the data are non-stationary, we can (i) allow for a deterministic trend in equation (1) and/or (ii) first-difference the data, and then proceed exactly as described above.

If a deterministic trend is suspected, we allow for a polynomial trend in each series, and we jointly estimate the parameters of the impulse responses (the $\boldsymbol{\Psi}_k$ coefficients) and the polynomial parameters.

If a stochastic trend is suspected, we can transform the data into stationary series by differencing the data. Importantly, the presence of co-integration does not imply that a GMA model in first-difference is misspecified.[35] After estimation, one can even test for co-integration by testing whether the matrix sum of moving-average coefficients ($\sum\limits_{k=1}^{K}\sum\limits_{l=0}^{k}\boldsymbol{\Psi}_l$) is of reduced rank (Engle and Yoo, 1987).

4 Gaussian Mixture Approximations of non-linear models

We now generalize the moving average model (1) by allowing for asymmetry and state-dependence, and we show how GMA models can easily accommodate such non-linearities.

4.1 A non-linear moving-average model

In this section, we generalize model (1) by allowing the economy to respond non-linearly to shocks, and we consider the model

$$\boldsymbol{y_t} = \sum_{k=0}^{K} \boldsymbol{\Psi}_k(\boldsymbol{\varepsilon_{t-k}}, \boldsymbol{z_{t-k}})\boldsymbol{\varepsilon_{t-k}} \tag{6}$$

[35] The reason is that a GMA model directly works with the moving-average representation and does not require inversion of the moving-average, unlike VAR models.

where $\boldsymbol{\varepsilon_t}$ is again the vector of structural innovations with $E\boldsymbol{\varepsilon_t} = \mathbf{0}$ and $E\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_t'} = I$, and $\boldsymbol{z_t}$ is a vector of stationary macroeconomic variables that can be a function of past variables of $\boldsymbol{y_t}$ or a function of variables exogenous to $\boldsymbol{y_t}$. As a normalization, we posit that $\boldsymbol{\Psi}_0$ has positive entries on the diagonal, i.e., $\boldsymbol{\Psi}_{0,\ell\ell}(\boldsymbol{\varepsilon_t}, \boldsymbol{z_t}) \geq 0$, $\forall \ell \in \{1, .., L\}$, $\forall t \in \{1, .., T\}$.

Model (6) is a *non-linear* vector moving average representation of the economy, because in contrast to (1), the matrix of lag coefficients $\boldsymbol{\Psi}_k(\boldsymbol{\varepsilon_{t-k}}, \boldsymbol{z_{t-k}})$ is no longer constant. Instead, the coefficients of matrix $\boldsymbol{\Psi}_k$ can depend on the values of the structural innovations $\boldsymbol{\varepsilon_{t-k}}$ and on the values of the macroeconomic variables in $\boldsymbol{z_{t-k}}$.

With $\boldsymbol{\Psi}_k$ a function of $\boldsymbol{\varepsilon_{t-k}}$, the impulse response functions to a given structural shock depend on the value of the shock at the time of shock. For instance, a positive shock may trigger a different impulse response than a negative shock.

With $\boldsymbol{\Psi}_k$ a function of $\boldsymbol{z_{t-k}}$, the impulse response functions to a structural shock depend on the value of the macroeconomic variables in $\boldsymbol{z}$ at the time of that shock. For instance, the response function may be different depending on the state of the business cycle (recession or expansion) at the time of the shock.

Because of its non-linear nature (6) does not admit a VAR representation, and the model cannot be recovered from a VAR.[36] Instead, our GMA approach directly works with the moving-average representation and can easily accommodate non-linearities. Moreover, the parametrization offered by Gaussian mixture approximations can ensure that the dimensionality of the problem remains reasonable. We now discuss in more details two cases of non-linear behavior that a GMA model can easily handle: (i) asymmetry and (ii) state-dependence.

[36]Regime-switching VAR models can capture certain types of non-linearities such as state dependence (whereby the value of some state variable affects the impulse response functions), but they cannot capture asymmetric effects of shocks (whereby the impulse response to a structural shock depends on the sign of that shock). With regime-switching VAR models, it is assumed that the economy can be in a finite number of regimes, and that each regime corresponds to a different set of VAR coefficients. However, if the true data generating process features asymmetric impulse responses, a new set of VAR coefficients would be necessary each period, because the (non-linear) behavior of the economy at any point in time depends on all structural shocks up to that point. As a result, such asymmetric data generating process cannot generally be approximated by a small number of state variables such as in threshold VARs or Markov-switching models.

4.1.1 Asymmetric effects of shocks

To allow for asymmetries, we let $\boldsymbol{\Psi}_k$ depend on the sign of the structural shock, i.e., we let $\boldsymbol{\Psi}_k$ take two possible values: $\boldsymbol{\Psi}_k^+$ or $\boldsymbol{\Psi}_k^-$. Specifically, a model that allows for asymmetric effects of shocks would be

$$\boldsymbol{y_t} = \sum_{k=0}^{K} \left[\boldsymbol{\Psi}_k^+ (\boldsymbol{\varepsilon_{t-k}} \odot \mathbf{1}_{\boldsymbol{\varepsilon_{t-k}>0}}) + \boldsymbol{\Psi}_k^- (\boldsymbol{\varepsilon_{t-k}} \odot \mathbf{1}_{\boldsymbol{\varepsilon_{t-k}<0}})\right] \tag{7}$$

with $\boldsymbol{\Psi}_k^+$ and $\boldsymbol{\Psi}_k^-$ the lag matrices of coefficients for, respectively, positive and negative shocks and $\odot$ denoting element-wise multiplication.

Denoting $\psi_{ij}^+(k)$, the i-row j-column coefficient of $\boldsymbol{\Psi}_k^+$ (that is, the impulse response of variable j to a positive shock i), a GMA(N) model would then be

$$\psi_{ij}^+(k) = \sum_{n=1}^{N} a_{ij,n}^+ e^{-\left(\frac{k-b_{ij,n}^+}{c_{ij,n}^+}\right)^2}, \quad \forall k \in (0, K] \tag{8}$$

with $a_{ij,n}^+$, $b_{ij,n}^+$, $c_{ij,n}^+$ some constants to be estimated. A similar expression would hold for $\psi_{ij}^-(k)$.

4.1.2 Asymmetric and state-dependent effects of shocks

With asymmetry and state dependence, $\boldsymbol{\Psi}_k^+$ becomes $\boldsymbol{\Psi}_k^+(\boldsymbol{z_{t-k}})$, i.e., the impulse response to a positive shock depends on the indicator vector $\boldsymbol{z_t}$ (and similarly for $\boldsymbol{\Psi}_k^-$).

For simplicity, let us consider the case where the vector of indicator variables $\boldsymbol{z}$ is a scalar z. Using a GMA(N) model, the impulse response function following a positive innovation (ψ_{ij}^+) can be parametrized as

$$\psi_{ij}^+(k) = (1 + \gamma_{ij}^+ z_{t-k}) \sum_{n=1}^{N} a_{ij,n}^+ e^{-\left(\frac{k-b_{ij,n}^+}{c_{ij,n}^+}\right)^2}, \quad \forall k \in (0, K] \tag{9}$$

with γ_{ij}^+, $a_{ij,n}^+$, $b_{ij,n}^+$ and $c_{ij,n}^+$ parameters to be estimated. An identical functional form holds for ψ_{ij}^{-}.

In this model, the amplitude of the impulse response depends on the state of the business cycle at the time of the shock. In (9), the amplitude of the impulse response is a function of the indicator variable z_t. Such a specification allows us to test whether, for instance, an expansionary policy has a stronger effect on output in a recession than in an expansion.

Note that in specification (9), the state of the cycle is allowed to stretch/contract the impulse response, but the shape of the impulse response is fixed (because a, b and c are all independent of z_t). While one could allow for a more general model in which all variables a, b and c depend on the indicator variable, specification (9) has two advantages. First, with limited sample size, it will typically be necessary to impose some structure on the data, and imposing a constant shape for the impulse response is a natural starting point.[37] Second, specification (9) generalizes trivially to GMAs of any order. The order of the GMA only determines the shape of the impulse response with higher order allowing for increasingly complex shapes. Then, for a given shape, the γ coefficient can stretch or expand the impulse response depending on the state of the cycle.[38]

4.2 Bayesian estimation of non-linear GMA models

The Bayesian estimation of non-linear GMA models proceeds similarly to linear GMA models, but the construction of the likelihood involves one additional complication that we briefly mention here and describe in detail in the Appendix.

The additional complication comes from the fact that one must make sure that the system $\boldsymbol{\Psi_0}(\boldsymbol{\varepsilon_t}, \boldsymbol{z_t})\boldsymbol{\varepsilon_t} = \mathbf{u}_t$ has a unique solution vector $\boldsymbol{\varepsilon_t}$ given a set of model parameters and given some vector $\mathbf{u}_t$. With the contemporaneous impact matrix $\boldsymbol{\Psi_0}$ a function of $\boldsymbol{\varepsilon_t}$, a unique so-

[37]Importantly, this assumption is easy to relax or to evaluate by model comparison using posterior odds ratios.

[38]Note the parallel and difference between (9) and a *varying coefficient model.* A varying coefficient model (e.g., Hastie and Tibshirani, 1993) is a (locally) linear model, whose coefficients are allowed to vary smoothly with some third variable z_t. In (9), the use of a *finite* sum of Gaussian basis functions (independent of z_t) plays a similar role to smoothness in varying coefficient models by restricting the shape of the impulse response and disciplining the estimates. Then, the effect of the third variable z_t is captured by letting the scale of the impulse response be a linear function of z_t.

lution is a priori not guaranteed. However, we show in the Appendix that there is a unique solution when we allow the identified shocks to have with asymmetric and/or state dependent effects in (i) the (full or partial) recursive identification scheme, (ii) the narrative identification scheme, and (iii) the sign-restriction identification scheme under the restriction that $sgn(\det \mathbf{\Psi_0^+}) = sgn(\det \mathbf{\Psi_0^-})$.

Compared to the linear case, the non-linear models require some initial values and prior distribution for the parameters controlling the non-linearities. As initial guesses, we set the parameters capturing asymmetry and state dependence to zero (i.e., no non-linearity).[39] This approach is consistent with the starting point of this paper: structural shocks have linear effects on the economy, and we are testing this hypothesis against the alternative that shocks have some non-linear effects. We then center the priors for these parameters at zero with flat (but proper) priors.

5 Monte Carlo simulations

In this section, we conduct a number of Monte-Carlo simulations to illustrate the working of GMA models as well as to evaluate their performances in finite sample. We first evaluate the performances of GMA models in the linear case, and we then evaluate the ability of GMA models to detect (i) asymmetry alone and (ii) asymmetry and state-dependence.

Importantly, in all our Monte Carlo exercises, the estimated GMA models will be misspecified and only approximate the true Data Generating Process (DGP). We follow this strategy for two reasons. First, we want to be conservative and stack the odds against our proposed method. Second, this strategy is consistent with the idea that a GMA is meant to approximate the true DGP. By focusing on the approximate shape of the impulse response and thereby economizing on degrees of freedom, a GMA may (i) provide better estimates of the impulse responses in short sample, –a classical example of the bias-variance trade-off–, and (ii) be able

[39]An alternative would be to obtain initial estimates about possible non-linear effects. One option could be to combine Jorda's (2005) local projection method (which can accommodate non-linearities) with the structural shocks recovered from the VAR in order to get first estimates of the non-linear impulse responses.

to detect non-linearities. One goal of these simulation exercises is to evaluate whether this can indeed be the case.

To simulate data, we proceed as follows. We first estimate a structural VAR on US data (using a recursive identification scheme), invert it to obtain a set of impulse responses $\left\{\hat{\mathbf{\Psi}}_k\right\}_{k=0}^{\infty}$, and we modify these baseline impulse responses to introduce non-linearities, in particular asymmetry or state dependence. From these impulse responses, we generate simulated data from

$$\boldsymbol{y_t} = \sum_{k=0}^{\infty} \hat{\mathbf{\Psi}}_k(\boldsymbol{\varepsilon_{t-k}}, \boldsymbol{z_{t-k}})\boldsymbol{\varepsilon_{t-k}} \tag{10}$$

with $\boldsymbol{\varepsilon_t}$ Normally distributed, $E\boldsymbol{\varepsilon_t} = 0$ and $E\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_t'} = \mathbf{I}$.

In each scenario, we use 50 Monte-Carlo replications with a sample size $T = 200$, which roughly corresponds to the sample size available for the US.

5.1 Linear model

Our first simulation is meant to illustrate the workings of Gaussian mixture approximations in the linear case. Our goal is not to claim that GMAs are superior to VARs but instead to convey that GMAs can provide a useful alternative approach, especially in short samples.

The DGP is obtained from estimating the quarterly VAR(4) considered previously with the unemployment rate, the PCE inflation rate and the federal funds rate over 1959-2007. The impulse response functions to a monetary shock can be seen in Figure 1.

For each simulated dataset, we estimate (i) a GMA(2), and (ii) a VAR(4), and we evaluate the Mean-Square Error (MSE) of the estimated impulse response function over the horizons $k = 1...25$.[40] Importantly, we stack the odds in favor of the VAR and against the GMA model, because the estimated VAR is a correctly specified model.

The first row of Table 1 presents the average MSEs over the simulations. For unemployment and inflation, the GMA(2) is respectively 25 percent and 50 percent more accurate on average

[40]Specifically, we report $MSE = \sum_{k=1}^{25}(\hat{\psi}(k) - \psi(k))^2$ where $\hat{\psi}$ is the estimated impulse response function and ψ is the true function.

than the VAR. For the fed funds rate, the MSE is small in both cases, but again with a slight advantage for the GMA.[41] Table 1 also presents the average length and coverage rate of the confidence bands capturing the 95 percent posterior probability and compares it with the confidence bands implied by a Bayesian VAR with loose, but proper, Normal-Wishart priors. We report the average length and coverage rate at the time of the peak effect of the shock of the variable of interest. We can see that the average lengths are smaller for the GMA than for the VAR, while the coverage rate of the GMA remains good.

5.2 Non-linear models

We now evaluate the performances of GMA models in detecting non-linearities. For the DGP, we start from a VAR with (log) GDP, inflation and the fed funds rate, where we detrend GDP with a quadratic trend. Although we could have used the same VAR as previously, we preferred this one, because the price puzzle is more substantial in this specification (Figure 4), so that the Monte-Carlo exercise will be a more stringent test on a GMA(1) model that cannot capture the oscillating pattern in inflation. Again, the goal of the exercise is to assess whether a GMA model that only approximates the main feature of the impulse responses can still recover non-linearities.

Asymmetry

We first consider a DGP where the impulse response functions to monetary shocks depend on the sign of the shock. To introduce asymmetry, we modify the impulse responses $\left\{\hat{\mathbf{\Psi}}_k\right\}_{k=0}^{\infty}$ to make them depend on the sign of the monetary shock, and Figure 4 plots the asymmetric impulse response functions. For realism, the level of asymmetry that we simulate is chosen to roughly match the magnitude of the asymmetry we later find in US data. Note that we do not impose asymmetry for the response of the fed funds rate. This is done to test whether our procedure incorrectly reports the existence of asymmetry when there is none.

[41] Intuitively, the reason for the superior performances of GMA is the fact that the VAR often shows counterfactual oscillation patterns. In contrast, the GMA(2) is disciplined by its stricter parametrization.

We estimate a GMA(1) with asymmetry on each set of simulated data, and Table 2 presents summary statistics for $a^+ - a^-$, which captures the amount of peak asymmetry for each one of the three variables in the model.

A number of results emerge. First, as shown by the frequency of rejection of zero coefficient for $a^+ - a^-$, the algorithm can detect asymmetry when it exists (case of output and inflation, first row of Table 2), even when the impulse response is not generated by one Gaussian, and even when, as with inflation, there is a strong oscillating pattern that cannot not captured by a one Gaussian approximation.[42] This is encouraging, because it supports our motivating idea that by approximating the most important feature of an impulse response, one can detect important non-linearities. Moreover, the algorithm does not detect asymmetry when there is none (case of the fed funds rate). Second, looking at the mean and standard-deviation of the estimates across Monte-Carlo replications (second row of Table 2), we can see that the algorithm under-estimates the amount of asymmetry (both for output and inflation). This indicates that in our empirical application on US data, our algorithm may under-estimate the magnitude of asymmetry present in the data. Third, the dispersion (third row) in the estimates across the Monte-Carlo replications is reasonably small, while the coverage rate of the posterior distribution – the frequency with which the true value lies within 90 percent of the posterior distribution–, is also good (fourth row).

Asymmetry and state dependence

We now consider a DGP where the impulse response functions to monetary shocks depend on the sign of the shock as well as the state of the business cycle. We introduce asymmetry exactly as in the previous exercise, but in addition, we posit that there is state dependence for output in response to a positive shock, i.e., $\gamma^+_{\mathsf{gdp}} \neq 0$ in (9), where the indicator variable z_t is the US unemployment rate.[43] Again, the value of γ^+_{gdp} is chosen to be of the same order of

[42] Specifically, the 90 percent posterior probability of $a^+ - a^-$ excludes zero for output and inflation respectively 94 and 90 percent of the time.

[43] We could have used any indicator, but we wanted an indicator that has the same time series properties as the one we use on US data. We thus chose to use the US unemployment rate, which is the indicator we used in

magnitude as our later empirical findings with US data, and we set $\gamma^{+}_{\mathsf{gdp}} = 1$.

We estimate a GMA(1) with asymmetry and state dependence on each set of simulated data, and Table 3 summarizes the results. A number of results emerge. First, the algorithm is very successful at detecting state dependence in output and the fact that $\gamma^{+}_{\mathsf{gdp}} \neq \gamma^{-}_{\mathsf{gdp}}$ (first set of columns in Table 3). In the 50 Monte-Carlo replications, we detect $\gamma^{+}_{\mathsf{gdp}} \neq \gamma^{-}_{\mathsf{gdp}}$ in all samples but one (first row). The algorithm also estimates the values of $\gamma^{+}_{\mathsf{gdp}} - \gamma^{-}_{\mathsf{gdp}}$ without bias (second row), with reasonable dispersion (third row) and with good coverage (fourth row). Importantly, the algorithm detects no state dependence when there is none (case of inflation), as can be seen from the close to zero frequency of rejection of zero coefficient. Second, the algorithm can still pick up the existence of asymmetry for output and inflation ($\alpha^{+} - \alpha^{-} \neq 0$, second set of columns). With a larger number of free parameters, estimation is more uncertain, but we can still detect the existence of asymmetry in more than 80 percent of cases. Finally, looking at the estimates for $\gamma^{+}_{\mathsf{gdp}}$ and $\gamma^{-}_{\mathsf{gdp}}$ separately, the algorithm estimates the value of $\gamma^{+}_{\mathsf{gdp}}$ –the magnitude of the non-linearity– with a downward bias, which seems to translate into an upward bias for $\gamma^{-}_{\mathsf{gdp}}$, although that bias is not significant over the 50 Monte-Carlo replications (last four columns of Table 3).

6 The non-linear effects of monetary shocks

In this section, we apply our proposed GMA approach and study the non-linear effects of monetary shocks. We consider a model of the US economy in the spirit of Primiceri (2005), where $\boldsymbol{y}_t$ includes the unemployment rate, the PCE inflation rate and the federal funds rate. As in Primiceri (2005), monetary policy affects the economy with a lag, and the matrix $\boldsymbol{\Psi}_0$ has its last column filled with 0 except for the diagonal coefficient. The data cover 1959Q1 to 2007Q4, and we exclude the latest recession where the fed funds rate was constrained at zero and no longer captured variations in the stance of monetary policy.[44] When constructing the

the application section.

[44] While we use quarterly data as in Primiceri (2005), we also conducted our estimation using monthly data. Results were very similar.

likelihood, we consider a moving-average model with $K = 45$, chosen to be large enough such that the lag matrix coefficients $\mathbf{\Psi}_k$ are close enough to zero for $k > K$.[45] For GMA models, we leave the non-zero coefficients of the contemporaneous impact matrix $\mathbf{\Psi}_0$ as free parameters.

As a preliminary test, we start by checking that a linear GMA model performs well against a standard VAR model. Then, we present the non-linear impulse response functions obtained from a non-linear GMA with asymmetry alone first, and then with asymmetry and state dependence.

6.1 The linear case: VAR versus GMA

First, we evaluate our GMA approach by doing a simple model comparison between a linear GMA(1) and a regular VAR with 4 lags.

Table 4 reports the (log) marginal data densities for the GMA and the VAR, so that a model comparison can be readily obtained by computing the Bayes factor (obtained by taking the exponential of the difference in (log) marginal data densities) after positing equal priors for the two competing models. Encouragingly for our approach, Bayesian model comparison favors the more parsimonious GMA(1) with a Bayes factor of about 400.

6.2 The asymmetric effects of monetary shocks

We now estimate an asymmetric GMA model in which the impulse responses to monetary shocks depend on the sign of the shock. As detailed in the methodology section, to choose the appropriate order of the GMA model, we consider models with an increasing number of Gaussian basis functions. As shown in columns (3) to (5) of Table 4, Bayesian model comparison favors a GMA(2) , and from now on we will report and discuss the results obtained using a GMA(2).

We can see that Bayesian model comparison strongly favors a model with asymmetry in the impulse responses to monetary shocks: the (log) marginal data density of an asymmetric

[45] As a robustness check, we consider a higher moving-average lag-length with $K = 55$. Results were identical.

GMA(2) is respectively 20 log-points larger than the linear (symmetric) GMA model and 25 log-points larger than the VAR model, which imply Bayes factors of respectively about 10^8 and 10^{11}.

Figure 5 plots the impulse responses (in percentage points) of unemployment, the price level and the federal funds rate to a one standard-deviation monetary shock. The thick lines denote the impulse response functions implied by the posterior mode, and the error bands are the 5th and 95th posterior percentiles.[46] When comparing impulse responses to positive and negative shocks, it is important to keep in mind that the impulse responses to expansionary monetary shocks (a decrease in the fed funds rate) were multiplied by -1 in order to ease comparison across impulse responses. With this convention, when there is no asymmetry, the impulse responses are identical in the upper panels (responses to a contractionary monetary shock) and in the bottom panels (responses to an expansionary monetary shock).

The evidence for asymmetry is striking: following a contractionary monetary shock, which represents a 70 basis points increase in the fed funds rate, unemployment increases by about 0.15 percentage points (ppt), whereas a (linear) VAR implies only a 0.10 ppt increase. In contrast, following an expansionary monetary shock (a 70 basis points decrease in the fed funds rate), the response of unemployment is small (a decline of 0.04 percentage points) and non-significantly different from zero. Figure 6 plots the posterior distribution of the difference in impulse responses between positive and negative shocks. This figure can be seen as a point-wise test of difference in impulse responses at different horizons. The 90 percent posterior interval of the difference in impulse responses of unemployment is substantially above zero for horizons 3 to 10, in line with the conclusion from the Bayes factors that the data support a model with asymmetric impulse responses to monetary shocks.[47]

Although the error bands are too large to be conclusive, the response of the price level also displays an interesting asymmetric pattern: the price level appears more sticky following a

[46]To be specific, this figure and subsequent figures show paths of the moving average coefficients ψ_k.

[47]In the case of the GMA(1) model, an alternative test for asymmetry is a Wald-type test on $a^+ - a^-$. This test (not shown) gives a similar conclusion: for unemployment, the 90 percent posterior interval of $a^+ - a^-$ excludes zero.

contractionary shock –displaying a larger price puzzle– than following an expansionary shock for which the price level drops on impact and displays no price puzzle. This is exactly the pattern one would expect if downward price (or wage) rigidity was responsible for the asymmetric response of unemployment.[48]

We also find asymmetry in the response of the fed funds rate to a monetary shock, but it is relatively mild. A monetary shock generates a slightly more persistent increase in the fed funds rate than its expansionary counterpart. This can be seen in the bottom right panel of Figure 5 where the response of the fed funds rate is slightly more short-lived following an expansionary shock.[49]

Robustness to identification assumptions

To show the robustness of our findings as well as to highlight how GMAs can accommodate other identification schemes, we now present asymmetric impulse response functions obtained with two alternative identification schemes: (i) a narrative approach, and (ii) sign restrictions.

Narrative approach We first evaluate the presence of asymmetry using monetary shocks identified through the narrative approach by Romer and Romer (2004) and extended until 2007 by Coibion et al. (2012). As pointed out by Coibion (2012), the advantage of the narrative procedure is that one should be able to more precisely identify the effects of monetary shocks than with a relatively small model like the one considered above, since the Romer and Romer measure controls for much of the endogenous fluctuations in the interest rate as well as the Fed's information set.

[48]The existence of downward wage rigidity is supported empirically by the scarcity of nominal wage cuts relative to nominal wage increases (e.g., Card and Hyslop, 1997).

[49]One way to gauge how much of the asymmetric response of unemployment can be explained by the asymmetric response of the fed funds rate is to proceed as in the government spending multiplier literature (e.g., Ramey and Zubairy, 2014) and to compute the total change in unemployment relative to the total change in the fed funds rate, that is to compute the multiplier $m = \sum_{k=0}^{K} \psi_k^U / \sum_{k=0}^{K} \psi_k^{ffr}$ for respectively positive and negative shocks. After "controlling" for the total change in the fed funds rate, the asymmetry is still present with $m^+ = .24 > m^- = .12$ with m^+ the multiplier associated with a contractionary shock (an increase in the fed funds rate) and m^- the multiplier associated with an expansionary shock.

We estimate an asymmetric GMA(2) model with 4 variables included in the following order: the Romer and Romer shocks, unemployment, inflation and the fed funds rate, and we posit that the contemporaneous matrix $\mathbf{\Psi}_0$ has its first row filled with 0 except for the diagonal coefficient, which implies that the narratively identified shock does not react contemporaneously to other shocks. This restriction is innocuous if the narrative shocks were correctly identified.

Figure 7 plots the asymmetric impulse responses to an innovation to the Romer and Romer shocks. Confirming our previous results, unemployment displays a very asymmetric response: there is no significant movement in unemployment following an expansionary shock, but there is a large increase following a contractionary shock.

Sign restrictions We also evaluate the presence of asymmetry using monetary shocks identified through sign restrictions. We posit that monetary shocks are the only shocks that raise the fed funds rate and lower inflation. We use a GMA(1) specification, so that the sign restrictions for inflation and the fed funds rate are imposed over the whole horizon.[50] As initial guess in our optimization routine, we use the structural impulse responses implied by a Cholesky ordering, and we use flat priors with $a \in [-10, 10]$ (as well as for the intercepts and the coefficients of $\mathbf{\Psi}_0$), $b \in [0, K]$ and $c \in [0, K]$.[51]

Figure 8 plots the asymmetric impulse responses to a monetary shock. Again, the evidence for asymmetry is very strong: while a contractionary shock raises unemployment significantly, an expansionary shock generates a much smaller (and non-significant) change in unemployment. Interestingly, the response of the price level is also strongly asymmetric with a strong price response following an expansionary shock, but only a weak response following a contractionary shock.[52] In other words, following a contractionary shock, quantities react, while following an expansionary shock, prices react. This asymmetry is consistent with downward

[50]Other identification schemes are possible, and a GMA(2) would allow us to impose the sign restriction over a specific horizon. We also experimented with imposing the additional restriction that the unemployment increases following a contractionary monetary shock. The estimated impulse responses were similar.

[51]The latter prior variance imposes that the effect of a shock can have a half-life as large as $K\sqrt{\ln 2} = 38$ quarters (recall $K = 45$ in our monetary application), which represents an extremely persistent impulse response.

[52]A similar pattern could be seen with the two previous identification schemes, but the asymmetry in the price response is most striking (and highly significant) with sign restrictions.

price (or wage) rigidity playing a role in the asymmetric response of unemployment.

6.3 The asymmetric and state-dependent effects of monetary shocks

In this section, we enrich our model by allowing the effects of monetary policy to depend on both the sign of the shock and the state of the business cycle. Intuitively, we would like to test whether monetary policy is more powerful at stimulating the economy in a period of economic slack, and whether an expansionary shock is more likely to generate inflation in a tight labor market. We thus estimate model (9) with a GMA(2), and we use last period's unemployment rate as cyclical indicator (z_t).[53] To put results into perspective, Figure 9 plots the unemployment rate (i.e., the indicator variable z_t) along with the identified monetary shocks.

Table 4 shows that Bayes model comparison strongly favors the model with asymmetry *and* state dependence over all the other models.

To visualize the effects of the state of the cycle on the impulse responses, Figure 10 shows how the peak effect of a monetary shock on unemployment or inflation depends on the state of the business cycle at the time of the shock.[54] The first two rows plot the peak responses of unemployment and inflation to contractionary and expansionary shocks. The left quadrants depict how the peak effect of a contractionary shock varies as we move from a tight labor market (unemployment at 4 percent) to a slack labor market (unemployment at 8 percent), and the right quadrants plot the same thing for an expansionary shock. The blues line depict estimates from our non-linear GMA model, and the thick dashed line represents the VAR estimate. Since the VAR is linear, that latter estimate is a horizontal line as the peak effect of monetary policy is independent of the state of the business cycle. Finally, the last row

[53] As an alternative, we also experienced with the unemployment rate detrended with an HP-filter ($\lambda = 10^5$). The latter specification was used to make sure that our results were not driven by slow moving trends (e.g., due to demographics) in the unemployment rate, which could make the unemployment rate a poor indicator of the amount of economic slack (see e.g. Barnichon and Mesters, 2015). We obtained similar results.

[54] To be specific, denote $\psi(k, z)$ the value of an impulse response function to a shock ε at horizon k when the indicator variable takes the value z at the time of the shock. Figure 10 plots the function f defined by $f(z) = sgn(\varepsilon) \max_{k \in [0,K]} |\psi(k, z)|$.

of Figure 10 plots histograms of the distributions of respectively contractionary shocks and expansionary shocks over the business cycle. This information is meant to get a sense of the range of unemployment over which we identify the coefficients capturing state dependence.

We first discuss the response of unemployment. The real effect of a contractionary shock (top left quadrant) increases with the unemployment rate: in a tight labor market, a (one standard-deviation) contractionary shock increases unemployment by about 0.13 percentage point (at the peak effect), but in a slack labor market, the same contractionary shock increases unemployment by about 0.18 percentage point (at the peak effect). Regarding the real effect of an expansionary shock (top right quadrant), the evidence is not very strong, but our estimates suggest some mild state dependence going in the same direction: the higher the unemployment rate, the larger the real effect of an expansionary policy. For instance, the 90th posterior probability bands start including the VAR point estimate, when the unemployment rate rises above 7 percent. The asymmetry in the real effects of expansionary and contractionary shocks remains however, and an expansionary shock is always considerably less potent than its contractionary counterpart.

We now turn to the response of inflation, depicted in the second row of Figure 10. While there is no evidence of state dependence for contractionary shocks, we find strong evidence that expansionary shocks generate a substantial rise in inflation when the unemployment rate is low: with an unemployment rate at 4 percent, an expansionary shock generates a peak increase in inflation of about 4 basis points (roughly twice as large as implied by the VAR point estimates). In contrast, with an unemployment rate at 8 percent, an expansionary shock has no effect on inflation. Interestingly, this finding is consistent with a standard Keynesian narrative, according to which a monetary authority trying to expand an economy already above potential would only achieve higher inflation through increased price/wage pressures.

7 Conclusion

This paper proposes a new method to estimate the (possibly non-linear) dynamic effects of structural shocks by using Gaussian basis functions to approximate impulse response functions. We apply our approach to the study of monetary policy and find that the effect of a monetary intervention depends strongly on the sign of the intervention. A contractionary shock has a strong adverse effect on output, larger than implied by linear estimates, but an expansionary shock has, on average, no significant effect on output. Interestingly, and while the evidence for inflation is more uncertain, the behavior of inflation is consistent with asymmetry emerging (at least in part) out of downward price/wage rigidities, because inflation displays a more marked price puzzle following a contractionary shock than following an expansionary shock. Finally, the effect of a monetary shock also depends on the state of the business cycle at the time of the intervention: An expansionary shock during a time a low unemployment generates not significant drop in unemployment but leads to a burst of inflation, consistent with a standard Keynesian narrative.

Although this paper studies non-linearities in the effect of monetary policy, Gaussian Mixture Approximations of the impulse responses may be useful in many other contexts, and we showed how our approach can be used with other identification schemes. Looking forward, our method could be used to estimate the non-linear effects of other important shocks where the existence of asymmetry or state-dependence remains an important and unresolved question; notably fiscal policy shocks (Auerbach and Gorodnichenko, 2012, Ramey and Zubairy, 2014) or credit supply shocks (Gilchrist and Zakrajsek, 2012). Moreover, the parametrization offered by GMA models and the associated efficiency gains may be useful even for linear models, where the sample size is small and/or the data are particularly noisy.

References

[1] Almon, S. "The distributed lag between capital appropriations and expenditures," Econometrica, 33, January, 178-196, 1965

[2] Amir Ahmadi P. and H. Uhlig. "Sign Restrictions in Bayesian FaVARs with an Application to Monetary Policy Shocks," NBER Working Paper, 2015

[3] Angrist, J., Jorda O and G. Kuersteiner. "Semiparametric estimates of monetary policy effects: string theory revisited," NBER Working Paper, 2013

[4] Alspach D. and H. Sorenson H. "Recursive Bayesian Estimation Using Gaussian Sums," Automatica, Vol 7, pp 465-479, 1971.

[5] Alspach D. and H. Sorenson H. "Nonlinear Bayesian Estimation Using Gaussian Sum Approximations," IEEE Transactions on Automatic Control, Vol 17-4, August 1972.

[6] Auerbach A and Y Gorodnichenko. "Measuring the Output Responses to Fiscal Policy," American Economic Journal: Economic Policy, vol. 4(2), pages 1-27, 2012

[7] Auerbach A and Y Gorodnichenko. "Fiscal Multipliers in Recession and Expansion." In Fiscal Policy After the Financial Crisis, edited by Alberto Alesina and Francesco Giavazzi, pp. 63–98. University of Chicago Press, 2013.

[8] Barnichon R, and G. Mesters, "On the Demographic Adjustment of Unemployment," Working Paper, 2015.

[9] Barnichon R. and C. Matthes. "Imposing structural identifying restrictions in Gaussian Mixture Approximation (GMA) models," Working Paper, 2016

[10] Barro, Robert J., "Unanticipated Money Growth and Unemployment in the United States," American Economic Review, LXVII, 101-15, 1977.

[11] Baumeister, C. and J. Hamilton, "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information," Econometrica, 83(5), 1963-1999, 2015.

[12] Beaudry, Paul, and Gary Koop. "Do Recessions Permanently Change Output?" Journal of Monetary Economics 31 (1993), 149 - 63.

[13] Blanchard, O. and D. Quah. "The Dynamic Effects of Aggregate Demand and Supply Disturbances," American Economic Review, 79(4), pages 655-73, September 1989.

[14] Buhmann, Martin D, Radial Basis Functions: Theory and Implementations, Cambridge University Press, 2003.

[15] Canova, F. "Methods for Applied Macroeconomic Research," Princeton University Press, 2007.

[16] Canova, F. and G. De Nicolo, "Monetary Disturbances Matter for Business Fluctuations in the G-7", Journal of Monetary Economics, 49, 1131-1159, 2002

[17] Card, D. and D. Hyslop, "Does Inflation Greases the Wheels of the Labor Market?" in Reducing Inflation: Motivation and Strategy, C. Romer and D. Romer, eds., University of Chicago Press, 1997.

[18] Casella, G. and R. L. Berger, Statistical Inference, Duxbury, 2002.

[19] Cover, J. "Asymmetric Effects of Positive and Negative Money-Supply Shocks," The Quarterly Journal of Economics, Vol. 107, No. 4, pp. 1261-1282, 1992.

[20] Coibion, O. "Are the Effects of Monetary Policy Shocks Big or Small?," American Economic Journal: Macroeconomics, vol. 4(2), pages 1-32, April, 2012.

[21] Coibion, O., Y. Gorodnichenko, L. Kueng, and J. Silvia, "Innocent Bystanders? Monetary Policy and Inequality in the US," NBER Working Papers 18170, National Bureau of Economic Research, Inc, 2012.

[22] Christiano, L., M. Eichenbaum, and C. Evans. "Monetary policy shocks: What have we learned and to what end?," Handbook of Macroeconomics, volume 1, chapter 2, pages 65-148, 1999.

[23] DeLong B and L. Summers, "How Does Macroeconomic Policy Affect Output?," Brookings Papers on Economic Activity, Economic Studies Program, The Brookings Institution, vol. 19(2), pages 433-494, 1988.

[24] Engle, R. and B. Yoo, "Forecasting and testing in co-integrated systems," Journal of Econometrics, vol. 35(1), pages 143-159, May 1987.

[25] Faust, J., "The Robustness of Identified VAR Conclusions about Money", Carnegie-Rochester Series on Public Policy, 49, 207-244, 1998.

[26] Fernandez-Villaverde, J. and J. Rubio-Ramirez, "Comparing dynamic equilibrium models to data: a Bayesian approach," Journal of Econometrics, vol. 123(1), pages 153-187, November, 2004.

[27] Gertler M and P Karadi, "Monetary Policy Surprises, Credit Costs, and Economic Activity," American Economic Journal: Macroeconomics, vol. 7(1), pages 44-76, January 2015.

[28] Gilchrist, S and E Zakrajsek, "Credit Spreads and Business Cycle Fluctuations," American Economic Review, American Economic Association, vol. 102(4), pages 1692-1720, June 2012.

[29] Haario, H., E. Saksman, and J. Tamminen, "An adaptive Metropolis algorithm," Bernoulli 7, no. 2, 223–242, 2001.

[30] Hamilton, J. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," Econometrica 57, 357-384, 1989.

[31] Hastie, T. and Tibshirani, R. "Varying-coefficient models," J. Roy. Statist. Soc. Ser. B 55, 757–796, 1993.

[32] Hastie, T., R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer 2009.

[33] Hubrich, K. and T. Terasvirta. "Thresholds and smooth transitions in vector autoregressive models," Advances in Econometrics "VAR Models in Macroeconomics, Financial Econometrics, and Forecasting", Vol. 31, 2013.

[34] Jorda O., "Estimation and Inference of Impulse Responses by Local Projections," American Economic Review, pages 161-182, March 2005.

[35] Koop G, M. Pesaran and S. Potter "Impulse response analysis in nonliner multivariate models," Journal of Econometrics, 74 119-147, 1996.

[36] Koop G and S. Potter "Dynamic asymmetries in US unemployment," Journal of Business & Economic Statistics Volume 17, Issue 3, 1999.

[37] Koreyaar J. *Mathematical Methods*, Vol. 1, pp 330-333. Academic Press, New York, 1968.

[38] Lippi, M. & Reichlin, L. "VAR analysis, nonfundamental representations, Blaschke matrices," Journal of Econometrics 63(1), 307–325, 1994.

[39] Lippi, M., Reichlin, L., "Diffusion of technical change and the decomposition of output into trend and cycle," Review of Economic Studies 61 (1) (206), 19–30, 1994b.

[40] Lo, M. C., and J. Piger "Is the Response of Output to Monetary Policy Asymmetric? Evidence from a Regime-Switching Coefficients Model," Journal of Money, Credit and Banking, 37(5), 865–86, 2005.

[41] McLachlan, G. and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2000.

[42] Morgan, D. "Asymmetric Effects of Monetary Policy." Federal Reserve Bank of Kansas City Economic Review 78, 21-33, 1993.

[43] Plagborg-Moller, P. "Bayesian Inference on Structural Impulse Response Functions," Working Paper, 2016.

[44] Potter, S. "A non-linear approach to US GNP," Journal of Applied Econometrics," Vol 10 109-125, 1995

[45] Primiceri, G. "Time Varying Structural Vector Autoregressions and Monetary Policy," Review of Economic Studies, vol. 72(3), pages 821-852, 2005.

[46] Racine, J. "Nonparametric Econometrics: A Primer," Foundations and Trends in Econometrics, now publishers, vol. 3(1), pages 1-88, March 2008.

[47] Ravn M and M Sola. "A Reconsideration of the Empirical Evidence on the Asymmetric Effects of Money-supply shocks: Positive vs. Negative or Big vs. Small," Archive Discussion Papers 9606, Birkbeck, 1996.

[48] Ravn M. and M. Sola. "Asymmetric effects of monetary policy in the United States," Review, Federal Reserve Bank of St. Louis, issue Sep, pages 41-60, 2004.

[49] Ramey V. "Comment on "Roads to Prosperity or Bridges to Nowhere? Theory and Evidence on the Impact of Public Infrastructure Investment"," NBER Chapters, in: NBER Macroeconomics Annual 2012, Volume 27, pages 147-153.

[50] Ramey V. and S. Zubairy. "Government Spending Multipliers in Good Times and in Bad: Evidence from U.S. Historical Data ," Working Paper, 2014.

[51] Robert, Christian P. and George Casella "Monte Carlo Satistical Methods" Springer, 2004.

[52] Romer, C., and D. Romer. "A New Measure of Monetary Shocks: Derivation and Implications," American Economic Review 94 (4): 1055–84, 2004

[53] Santoro, E, I. Petrella, D. Pfajfar and E. Gaffeo ”Loss Aversion and the Asymmetric Transmission of Monetary Policy,” Journal of Monetary Economics, 68 19-35, 2014.

[54] Swanson, E. and J. Williams. “Measuring the Effect of the Zero Lower Bound on Medium– and Longer–Term Interest Rates.” American Economic Review 104 (10): 3154–85, 2014.

[55] Tenreyro, S., and G. Thwaites (2015): “Pushing on a string: US monetary policy is less powerful in recessions,” Working Paper.

[56] Thoma, M. ”Subsample Instability and Asymmetries in Money-Income Causality.” Journal of Econometrics 64, 279-306, 1994

[57] Uhlig, H. ”What are the effects of monetary policy on output? Results from an agnostic identification procedure,” Journal of Monetary Economics, vol. 52(2), pages 381-419, March 2005.

[58] Walsh C. Monetary Theory and Policy, 3nd. ed., The MIT Press, 2010.

[59] Weise, C. ”The Asymmetric Effects of Monetary Policy: A Nonlinear Vector Autoregression Approach,” Journal of Money, Credit and Banking, vol. 31(1), pages 85-108, February 1999.

Appendix A1: Proof of Theorem 1

Following Alspach and Sorenson (1971, 1972) in the context of approximating distributions, the problem of approximating a function f can be considered within the context of delta families of positive types.

Delta families are families of functions which converge to a delta function as a parameter characterizing the family converges to a limit value.

Let $\{\delta_\lambda\}$ be a family of functions on the interval $]-\infty, +\infty[$ which are integrable over every interval. $\{\delta_\lambda\}$ forms a delta family of positive type if the following conditions are satisfied:

1. For every constant $\gamma > 0$, δ_λ tends to zero uniformly for $\gamma \leq |x| \leq \infty$ as $\lambda \to \lambda_0$
2. There exist s in $\mathbb{R}$ so that $\int_{-s}^{s} \delta_\lambda(x)dx \longrightarrow 1$ as λ tends to some limit value λ_0
3. $\delta_\lambda(x) \geq 0$ for all x and λ

Defining

$$\delta_\lambda(x) \equiv G_\lambda(x) = \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{x^2}{\lambda^2}}, \tag{11}$$

it is easy to see that the Gaussian functions $\{G_\lambda\}$ form a delta family of positive type as $\lambda \to 0$ (i.e., $\lambda_0 = 0$). That is, the Gaussian function tends to the delta function as the variance tends to zero.[55]

We can then make use of the following theorem.

Theorem: The sequence $\{f_\lambda\}$ which is formed by the convolution of δ_λ and f

$$f_\lambda(x) = \int_{-\infty}^{+\infty} \delta_\lambda(x-u) f(u) du \tag{12}$$

converges uniformly to f as $\lambda \to \lambda_0$ for x on every interval $[x_0, x_1]$ of $\mathbb{R}$.

Proof. See Korevaar (1968). ■

[55] Note that this proof can be easily applied to other functions (such as the inverse quadratic function $x \to \frac{1}{1+(\frac{x}{\lambda})^2}$) that form a delta family of a positive type, so that our approach is not restricted to Gaussian functions.

Using (11) in (12), the function f_λ given by

$$f_\lambda(x) = \int_{-\infty}^{+\infty} G_\lambda(x-u)f(u)du \tag{13}$$

converges uniformly to f as $\lambda \to 0$ for x in some arbitrary interval $[x_0, x_1]$ of $\mathbb{R}$.

Next, we want to approximate (13) with a Riemann sum. To do so, first rewrite f_λ as

$$f_\lambda(x) = \underbrace{\int_{-\infty}^{-s} G_\lambda(x-u)f(u)du}_{=A(\lambda,x)} + \int_{-s}^{+s} G_\lambda(x-u)f(u)du + \underbrace{\int_{s}^{+\infty} G_\lambda(x-u)f(u)du}_{=B(\lambda,x)} \tag{14}$$

for $s > 1$.

Note that for any $s > 1$, we have

$$\begin{aligned} 0 &\leq \int_s^{+\infty} G_\lambda(u)du \\ &\leq \frac{1}{\sqrt{2\pi\lambda^2}} \int_s^{+\infty} e^{-\frac{u}{\lambda^2}} du \text{ since } u^2 > u \text{ for any } u \text{ in } [s, +\infty[,\ s > 1 \\ &\leq \left[\frac{-\lambda^2}{\sqrt{2\pi\lambda^2}} e^{-\frac{u}{\lambda^2}}\right]_s^{+\infty} = \frac{|\lambda|}{\sqrt{2\pi}} e^{-\frac{s}{\lambda^2}} \xrightarrow[\lambda \to 0]{} 0 \end{aligned}$$

which shows that $\forall s > 1$, $\lim\limits_{\lambda -> 0} \int_s^{+\infty} G_\lambda(u)du = 0$. Symmetrically, we can show $\lim\limits_{\lambda -> 0} \int_{-\infty}^{-s} G_\lambda(u)du = 0$.

Going back to (14), we have

$$0 \leq |B(\lambda, x)| \leq M \int_{-\infty}^{x-s} G_\lambda(t)dt$$

where $M = \sup\limits_{x \in \mathbb{R}} |f(x)|$. Since $x \in [x_0, x_1]$, we can choose an $s > 1$ such that $x - s < -1$, so that we can apply the previous result and get

$$\lim_{\lambda \to 0} |B(\lambda, x)| = 0. \tag{15}$$

Proceeding symmetrically, we have $\lim_{\lambda \to 0} |A(\lambda, x)| = 0$.

Finally, since the function $u \mapsto G_\lambda(x-u)f(u)$ is continuous over $[-s, s]$, we can approximate $\int_{-s}^{+s} G_\lambda(x-u)f(u)du$ with a Riemann sum. Denoting

$$f_{\lambda,N}(x) = \sum_{n=1}^{N} G_\lambda(x - \xi_n) f(\xi_n) (\xi_n - \xi_{n-1})$$

where $\xi_n = -s + n\frac{2s}{N}$, we get that

$$\lim_{N \to \infty} f_{\lambda,N}(x) = \int_{-s}^{+s} G_\lambda(x-u)f(u)du. \tag{16}$$

Denoting $a_n = f(\xi_n)(\xi_n - \xi_{n-1})$, $b_n = \xi_n$ and $c_n = \lambda$, using (16), (15) in (14) and combining with (13), we get that

$$\lim_{\lambda \to 0} \left(\lim_{N \to \infty} f_{\lambda,N}(x) \right) = f(x)$$

which completes the proof.

Appendix A2: Identifying restrictions in non-linear Moving-Average models

We now detail how to impose the different identifying restrictions used in the paper. We only discuss the non-linear model $\boldsymbol{y_t} = \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k(\boldsymbol{\varepsilon_{t-k}}, \boldsymbol{z_{t-k}})\boldsymbol{\varepsilon_{t-k}}$, since it includes the simpler linear model $\boldsymbol{y_t} = \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k \boldsymbol{\varepsilon_{t-k}}$.

As described in the main text, we impose the identifying restriction when we construct the likelihood, so that constructing the likelihood and imposing identifying restrictions are intimately linked, and we thus describe them jointly. To recursively construct the likelihood at time t, one must ensure that the shock vector $\boldsymbol{\varepsilon_t}$ is uniquely determined given a set of model parameters and the history of variables up to time t. As described in the main text, in order to construct the likelihood recursively, the system of equations

$$\boldsymbol{\Psi}_0(\boldsymbol{\varepsilon_t}, \boldsymbol{z_t})\boldsymbol{\varepsilon_t} = \mathbf{u}_t \tag{17}$$

need to have a unique solution vector $\boldsymbol{\varepsilon_t}$ given $\mathbf{u}_t = \boldsymbol{y_t} - \sum_{k=0}^{K} \boldsymbol{\Psi}_k(\boldsymbol{\varepsilon_{t-k}}, \boldsymbol{z_{t-k}})\boldsymbol{\varepsilon_{t-1-k}}$. That is, we must ensure that there is a one-to-one mapping from $\boldsymbol{\varepsilon_t}$ to $\boldsymbol{\Psi}_0(\boldsymbol{\varepsilon_t}, \boldsymbol{z_t})\boldsymbol{\varepsilon_t}$. In the linear case, this means that we must ensure $\boldsymbol{\Psi}_0$ is invertible. In the non-linear case, ensuring that the shock vector $\boldsymbol{\varepsilon_t}$ is uniquely determined becomes more complicated, *when* we allow $\boldsymbol{\Psi}_0$ to depend on the sign of the shock or on some state variable.[56]

Consider first the consequences of allowing for state dependence, i.e., when $\boldsymbol{\Psi}_k$ depends on the value of the indicator vector $\boldsymbol{z_t}$, so that the likelihood also depends on the value of the indicator vector $\boldsymbol{z_t}$. Technically, constructing the likelihood of this specification is a straightforward extension of the linear case, when $\boldsymbol{z_t}$ is a function of lagged values of $\boldsymbol{y}_t$. To see that, note that we use the prediction-error decomposition to construct the likelihood function. We build a sequence of densities for $\boldsymbol{y_t}$ that conditions on past values of $\boldsymbol{y_t}$. Thus, conditional on past values of $\boldsymbol{y}_t$, $\boldsymbol{z_t}$ is known, and as long as $\boldsymbol{\Psi}_0(\boldsymbol{z_t})$ is invertible, there is (one-to-one) mapping from $\boldsymbol{\varepsilon_t}$ to $\boldsymbol{\Psi}_0\boldsymbol{\varepsilon_t}$, and the likelihood can be recursively constructed.[57]

Consider now the consequences of allowing for asymmetry, i.e., when $\boldsymbol{\Psi}_k$ depends on the sign of $\boldsymbol{\varepsilon_t}$. A complication arises when one allows $\boldsymbol{\Psi}_0$ to depend on the sign of the shock *while also* imposing identifying restrictions on $\boldsymbol{\Psi}_0$. The complication arises, because with asymmetry, the system of equations $\boldsymbol{\Psi}_0(\boldsymbol{\varepsilon_t})\boldsymbol{\varepsilon_t} = \mathbf{u}_t$ need not have a unique solution vector $\boldsymbol{\varepsilon_t}$, because $\boldsymbol{\Psi}_0(\boldsymbol{\varepsilon_t})$, the impact matrix, depends on the sign of the shocks, i.e., on the vector $\boldsymbol{\varepsilon_t}$.

In this appendix, we show how to address the issue when we allow the identified shocks

[56] Note that if the impact matrix $\boldsymbol{\Psi}_0$ is a constant and does not depend on $\boldsymbol{\varepsilon_t}$ or $\boldsymbol{z_t}$ (so that $\boldsymbol{\Psi}_k$ depends on $\boldsymbol{\varepsilon_t}$ or $\boldsymbol{z_t}$ only for $k > 0$), then one can construct the likelihood just as in the linear case, because as long as $\boldsymbol{\Psi}_0$ is invertible, there is (one-to-one) mapping from $\boldsymbol{\varepsilon_t}$ to $\boldsymbol{\Psi}_0\boldsymbol{\varepsilon_t}$, and $\boldsymbol{\varepsilon_t}$ is uniquely defined from $\mathbf{u}_t$.

[57] If we wanted to use an indicator function that was not a function of the history of endogenous variables $\boldsymbol{y}^{t-1}$, this would also be possible by using a quasi-likelihood approach. That is, we would build a likelihood function that not only conditions on the parameters, but also the sequence of indicators $\boldsymbol{z_t}$. This would in general not be efficient because the joint density of $\boldsymbol{z_t}$ and $\boldsymbol{y_t}$ could carry more information about the parameters in our model than the conditional density we advocate using. As long as $\boldsymbol{z_t}$ is highly correlated with elements of (functions of) $\boldsymbol{y}_t$, this loss in efficiency will likely be small.

to have asymmetric and state dependent effects on the impulse response functions. We successively consider each identification scheme used in the paper: (i) recursive ordering, (ii) narrative identification, and (iii) sign restrictions.

1. Recursive identification scheme

It will be convenient to adopt the following conventions for notation:

- Denote $y_{\ell,t}$ the ℓth variable of vector $\boldsymbol{y_t}$ and denote $\boldsymbol{y_t}^{<\ell} = (y_{\ell,t}, ..., y_{\ell-1,t})'$ the vector of variables ordered before variable $y_{\ell,t}$ in $\boldsymbol{y_t}$. Similarly, we can define $\boldsymbol{y_t}^{\leq\ell}$ or $\boldsymbol{y_t}^{>\ell}$.

- For a matrix $\boldsymbol{\Gamma}$ of size $L \times L$ and $(i,j) \in \{1, ..., L\}^2$, denote $\boldsymbol{\Gamma}^{<i,<j}$ the $(i-1) \times (j-1)$ submatrix of Γ made of the first $(i-1)$ rows and $(j-1)$ columns. Similarly, we denote $\boldsymbol{\Gamma}^{>i,>j}$ the $(L-i) \times (L-j)$ submatrix of $\boldsymbol{\Gamma}$ made of the last $(L-i)$ rows and $(L-j)$ columns. In the same spirit, we denote $\boldsymbol{\Gamma}^{i,<j}$ the submatrix of $\boldsymbol{\Gamma}$ made of the ith row and the first $(j-1)$ columns. $\boldsymbol{\Gamma}^{i,<j}$ is in fact a row vector. A combination of these notations allows us to denote any submatrix of $\boldsymbol{\Gamma}$. Finally, d enote $\boldsymbol{\Gamma}_{ij}$ the ith row jth column element of $\boldsymbol{\Gamma}$.

With these notations, we can now state the recursive identifying assumption

Assumption 1 (Partial recursive identification) *The contemporaneous impact matrix $\boldsymbol{\Psi_0}$ of dimension $L \times L$ is of the form*

$$\boldsymbol{\Psi_0} = \begin{bmatrix} \underset{(\ell-1)\times(\ell-1)}{\boldsymbol{\Psi_0}^{<\ell,<\ell}} & \underset{(\ell-1)\times 1}{\mathbf{0}^{<\ell,\ell}} & \underset{(\ell-1)\times(L-\ell)}{\mathbf{0}^{<\ell,>\ell}} \\ \underset{1\times(\ell-1)}{\boldsymbol{\Psi_0}^{\ell,<\ell}} & \underset{1\times 1}{\boldsymbol{\Psi}_{0,\ell\ell}} & \underset{1\times(L-\ell)}{\mathbf{0}^{\ell,>\ell}} \\ \underset{(L-\ell)\times(\ell-1)}{\boldsymbol{\Psi_0}^{>\ell,<\ell}} & \underset{(L-\ell)\times 1}{\boldsymbol{\Psi_0}^{>\ell,\ell}} & \underset{(L-\ell)\times(L-\ell)}{\boldsymbol{\Psi_0}^{>\ell,>\ell}} \end{bmatrix}.$$

with $\ell \in \{1, .., L\}$, $\boldsymbol{\Psi_0}^{<\ell,<\ell}$ and $\boldsymbol{\Psi_0}^{>\ell,>\ell}$ matrices of full rank and $\mathbf{0}$ denoting the $L \times L$ zero matrix.

Assumption 1 states that the shock of interest $\varepsilon_{\ell,t}$, ordered in ℓth position in $\boldsymbol{\varepsilon_t}$, affects the variables ordered from 1 to $\ell-1$ with a one period lag, and that the first ℓ variables in $\boldsymbol{y_t}$ do not react contemporaneously to shocks ordered after $\varepsilon_{\ell,t}$ in $\boldsymbol{\varepsilon_t}$. For instance, in Primiceri (2005)'s monetary model used in section 6, the policy rate is ordered last, and the recursive identification scheme states that shocks to the policy rate do not affect unemployment and inflation contemporaneously, i.e., that the last column of $\boldsymbol{\Psi_0}$ is filled with zeros except for the diagonal element.

We first consider a model with only asymmetry and then a model with asymmetry and state dependence.

1.1 Asymmetric impulse response functions

Proposition 1 *Consider the non-linear moving average model defined in (6) with*

$$\boldsymbol{\Psi}_k(\boldsymbol{\varepsilon_{t-k}}) = \left[\boldsymbol{\Psi}_k^+ 1_{\varepsilon_{\ell,t-k}>0} + \boldsymbol{\Psi}_k^- 1_{\varepsilon_{\ell,t-k}<0}\right], \;\; \forall k \in \{0,..,K\}, \; \forall t \in \{1,..,T\} \tag{18}$$

with $\ell \in \{1,..,L\}$, $\varepsilon_{\ell,t}$, the ℓth structural shock in $\boldsymbol{\varepsilon_t}$ and with $\boldsymbol{\Psi}_0$ satisfying Assumption 1. Then, given $\{\boldsymbol{y_t}\}_{t=1}^T$, given the model parameters and given K initial values of the shocks $\{\boldsymbol{\varepsilon}_{-K}...\boldsymbol{\varepsilon}_0\}$, the series of shocks $\{\boldsymbol{\varepsilon_t}\}_{t=1}^T$ is uniquely determined.

Proof. The key to Proposition 1 is to show that the sign of the monetary shock $\varepsilon_{\ell,t}$ is uniquely pinned down by (17).

We first establish the following lemma:

Lemma 1 *Consider a matrix $\boldsymbol{\Gamma}$ that can be written as*

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and $\mathbf{D}$ are matrix sub-blocks of arbitrary size, with A a non-singular squared *matrix and* $\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ *nonsingular. Then, the inverse of* $\mathbf{\Gamma}$ *satisfies*

$$\mathbf{\Gamma}^{-1}=\begin{pmatrix} \mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}\mathbf{F}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{F}^{-1} \\ -\mathbf{F}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{F}^{-1} \end{pmatrix}$$

with $\mathbf{F}=\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}$.

Proof. Verify that $\mathbf{\Gamma}\mathbf{\Gamma}^{-1}=\mathbf{I}$. ■

We prove Proposition 1 by induction, so that given past shocks $\{\boldsymbol{\varepsilon_{t-1-K}},...,\boldsymbol{\varepsilon_{t-1}}\}$ (and given model parameters $\{\mathbf{\Psi}_k\}_{k=0}^{K}$), we will prove that the system

$$\mathbf{u}_t=\mathbf{\Psi_0}(\varepsilon_{\ell,t})\boldsymbol{\varepsilon_t} \tag{19}$$

with $\mathbf{u}_t=\boldsymbol{y_t}-\sum_{k=0}^{K}\mathbf{\Psi}_k(\varepsilon_{\ell,t})\boldsymbol{\varepsilon_{t-1-k}}$, has a unique solution vector $\boldsymbol{\varepsilon_t}$.

Notice that (19) implies the sub-system with ℓ equations

$$\mathbf{u}_t^{\leq\ell}=\begin{pmatrix} \mathbf{\Psi}_0^{<\ell,<\ell} & \mathbf{0}^{<\ell,1} \\ \mathbf{\Psi}_0^{\ell,<\ell} & \mathbf{\Psi}_{0,\ell\ell}(\varepsilon_{\ell,t}) \end{pmatrix}\boldsymbol{\varepsilon_t^{\leq\ell}} \tag{20}$$

and notice that the matrix in (20) depends on $\varepsilon_{\ell,t}$ only through the scalar $\mathbf{\Psi}_{0,\ell\ell}(\varepsilon_{\ell,t})$. Denoting $\mathbf{A}\equiv\mathbf{\Psi_0}^{<\ell,<\ell}$ a $(\ell-1)\times(\ell-1)$ invertible matrix (from Assumption 1), $\mathbf{C}\equiv\mathbf{\Psi_0}^{\ell,<\ell}$ a $1\times(\ell-1)$ matrix, $\mathbf{B}\equiv\mathbf{0}$ of dimension $(\ell-1)\times 1$, and $D(\varepsilon_{\ell,t})\equiv\mathbf{\Psi}_{0,\ell\ell}(\varepsilon_{\ell,t})$ the (ℓ,ℓ) coefficient of $\mathbf{\Psi}_0$ (a scalar), we can use Lemma 1 to invert the system (20) and obtain

$$\boldsymbol{\varepsilon_t^{\leq\ell}}=\frac{1}{D(\varepsilon_{\ell,t})}\begin{pmatrix} D(\varepsilon_{\ell,t})\mathbf{A}^{-1} & \mathbf{0}^{<\ell,1} \\ -\mathbf{C}\mathbf{A}^{-1} & 1 \end{pmatrix}\mathbf{u}_t^{\leq\ell}. \tag{21}$$

The last row of (21) provides the equation $\varepsilon_{\ell,t}=\frac{1}{D(\varepsilon_{\ell,t})}\begin{pmatrix} -\mathbf{C}\mathbf{A}^{-1} & 1 \end{pmatrix}\mathbf{u}_t$, which defines $\varepsilon_{\ell,t}$. Since the right hand side of that equation only depends on $\varepsilon_{\ell,t}$ through $D(\varepsilon_{\ell,t})$, the sign of the right hand side depends on $\varepsilon_{\ell,t}$ only through the sign of $D(\varepsilon_{\ell,t})=\mathbf{\Psi}_{0,\ell\ell}(\varepsilon_{\ell,t})$. But since

$\boldsymbol{\Psi}_{0,\ell\ell}(\varepsilon_{\ell,t})$, the sign of the contemporaneous effect of the shock $\varepsilon_{\ell,t}$ on variable $y_{l,t}$, is posited to be positive as a normalization, the sign (and the value) of $\varepsilon_{\ell,t}$ is uniquely determined from the last row of (21). Then, with $\boldsymbol{\Psi}_{\mathbf{0}}^{<\ell,<\ell}$ and $\boldsymbol{\Psi}_{\mathbf{0}}^{>\ell,>\ell}$ invertible, (19) has a unique solution vector $\boldsymbol{\varepsilon_t}$. ■

Proposition 1 ensures that the system (17) has a unique solution vector, even when the shock $\varepsilon_{\ell,t}$, identified from a recursive ordering, triggers asymmetric impulse response functions.

With Proposition 1, we can then construct the likelihood recursively. To write down the one-step ahead forecast density $p(\boldsymbol{y_t}|\boldsymbol{\theta}, \boldsymbol{y^{t-1}})$ as a function of past observations and model parameters, we use the standard result (see e.g., Casella-Berger, 2002) that for $\boldsymbol{\Psi}_0$ a function of $\boldsymbol{\varepsilon_t}$, we have

$$p(\boldsymbol{\Psi}_0(\varepsilon_{\ell,t})\varepsilon_{\ell,t}|\boldsymbol{\theta}, \boldsymbol{y^{t-1}}) = J_t p(\boldsymbol{\varepsilon_t})$$

where J_t is the Jacobian of the (one-to-one) mapping from $\boldsymbol{\varepsilon_t}$ to $\boldsymbol{\Psi}_0(\boldsymbol{\varepsilon_t})\boldsymbol{\varepsilon_t}$ and where $p(\boldsymbol{\varepsilon_t})$ is the density of $\boldsymbol{\varepsilon_t}$.[58]

Finally, note that while we considered the case of a partially identified model, we can proceed similarly for a fully identified model with $\boldsymbol{\Psi}_0$ lower triangular and show that the shock vector $\boldsymbol{\varepsilon}_t$ is uniquely determined by (17) even when all shocks have asymmetric effects.

1.2 Asymmetric and state-dependent impulse response functions

We now consider a model with asymmetry and state dependence. For clarity of exposition, we consider the simpler case of a univariate state variable $z_t \in [\underline{z}, \overline{z}]$ with $\underline{z} = \min_{t\in[1,T]}(z_t)$ and $\overline{z} = \max_{t\in[1,T]}(z_t)$. The following proposition establishes the condition under which system (17) has a unique solution even when the identified shock $\varepsilon_{\ell,t}$ has asymmetric and state dependent effects.

[58]In our case with asymmetry, this Jacobian is simple to calculate, but the mapping is not differentiable at $\varepsilon_{\ell,t} = 0$. Since we will never exactly observe $\varepsilon_{\ell,t} = 0$ in a finite sample, we can implicitly assume that in a small neighborhood around 0, we replace the original mapping with a smooth function.

Proposition 2 *Consider the non-linear moving average model defined in (6) with*

$$\boldsymbol{\Psi}_k(\boldsymbol{\varepsilon_{t-k}}, z_{t-k}) = \left[\boldsymbol{\Psi}_k^+(z_{t-k})1_{\varepsilon_{\ell,t-k}>0} + \boldsymbol{\Psi}_k^-(z_{t-k})1_{\varepsilon_{\ell,t-k}<0}\right], \ \forall k \in \{0,..,K\}, \forall t \in \{1,..,T\} \tag{22}$$

with $z_t \in [\underline{z}, \overline{z}]$*,* $\ell \in \{1,..,L\}$*,* $\varepsilon_{\ell,t}$*, the* ℓ*th structural shock in* $\boldsymbol{\varepsilon_t}$*, and with* $\boldsymbol{\Psi}_0$ *satisfying Assumption 1. Then, given* $\{\boldsymbol{y_t}\}_{t=1}^T$*, given the model parameters and given* K *initial values of the shocks* $\{\boldsymbol{\varepsilon}_{-K}...\boldsymbol{\varepsilon}_0\}$*, the series of shocks* $\{\boldsymbol{\varepsilon_t}\}_{t=1}^T$ *is uniquely determined provided that* $sgn\left(\boldsymbol{\Psi}_{0,\ell\ell}^+(z_t)\right) = sgn\left(\boldsymbol{\Psi}_{0,\ell\ell}^-(z_t)\right) > 0$*,* $\forall z_t \in [\underline{z}, \overline{z}]$.

Proof. The proof proceeds exactly as with Proposition 1 and consists in showing that the system $\mathbf{u}_t = \boldsymbol{\Psi_0}(\varepsilon_{\ell,t}, z_t)\boldsymbol{\varepsilon_t}$ determines a unique solution vector $\boldsymbol{\varepsilon_t}$. As with Proposition 1, this is the case as long as $\boldsymbol{\Psi}_{0,\ell\ell}(\varepsilon_{\ell,t}, z_t) > 0$ regardless of the value of z_t. ■

Taking as an example the case of the monetary model from section 6, the restriction in Proposition 2 implies $sgn\left(\boldsymbol{\Psi}_{0,\ell\ell}^+(\underline{z})\right) = sgn\left(\boldsymbol{\Psi}_{0,\ell\ell}^+(\overline{z})\right)$ and similarly for $\boldsymbol{\Psi}_{0,\ell\ell}^-$, so that the coefficient of the impact response of the fed funds rate to a monetary shock is always positive, regardless of the state of the cycle. Note that this restriction is very mild, in that it is in fact an existence condition for the moving average model, since the diagonal coefficients of $\boldsymbol{\Psi}_k$ are posited to be positive as a normalization.

With Proposition 2 in hand, we can then construct the likelihood recursively as described in the previous section.

2. Narrative identification scheme

For a narrative identification scheme, we can use the previous results on recursive identification, since the use of narratively identified shocks can be cast as a partial identification scheme.

Indeed, if one orders the narratively identified shocks series first in $\boldsymbol{y_t}$, we can assume that $\boldsymbol{\Psi}_0$ has its first row filled with 0 except for the diagonal coefficient, which implies that the narratively identified shock does not react contemporaneously to other shocks (as should be the case if the narrative shocks were correctly identified). With Assumption 1 satisfied with

$\ell = 1$, Proposition 1 and 2 then imply that (17) has a unique solution vector $\boldsymbol{\varepsilon}_t$ even when the narratively identified shocks has asymmetric and state dependent effects.

3. Identification from sign restrictions

We now consider the case of a set identification scheme based on sign restrictions. Denote ε_t^r the structural shock of interest identified from sign restrictions. We now establish the conditions under which system (17) has a unique solution vector, first in a model with asymmetry, and second in a model with asymmetry and state dependence.

3.1 Asymmetric impulse response functions

Proposition 3 *Consider the non-linear moving average model defined in (6) with*

$$\boldsymbol{\Psi}_k(\boldsymbol{\varepsilon_{t-k}}) = \left[\boldsymbol{\Psi}_k^+ 1_{\varepsilon_{t-k}^r>0} + \boldsymbol{\Psi}_k^- 1_{\varepsilon_{t-k}^r<0}\right], \;\; \forall k \in \{0,..,K\}, \;\; \forall t \in \{1,..,T\} \tag{23}$$

with ε_t^r the structural shock identified from sign restrictions. Then, given $\{\boldsymbol{y_t}\}_{t=1}^T$, given the model parameters and given K initial values of the shocks $\{\boldsymbol{\varepsilon}_{-K}...\boldsymbol{\varepsilon}_0\}$, the series of shocks $\{\boldsymbol{\varepsilon_t}\}_{t=1}^T$ is uniquely determined provided that $sgn(\det \boldsymbol{\Psi}_0^+) = sgn(\det \boldsymbol{\Psi}_0^-)$.

Proof. Without loss of generality, let us order the variables such that ε_t^r, the shock with asymmetric effects, is ordered last. We can then write $\boldsymbol{\Psi}_0(\varepsilon_t^r)$ (of dimension $L \times L$) as

$$\boldsymbol{\Psi}_0(\varepsilon_t) = \begin{pmatrix} \mathbf{A} & \mathbf{B}(\varepsilon_t^r) \\ \mathbf{C} & D(\varepsilon_t^r) \end{pmatrix}$$

with $\mathbf{A}$ a $(L-1) \times (L-1)$ invertible matrix, $\mathbf{C}$ a $1 \times (L-1)$ matrix, $\mathbf{B}(\varepsilon_t^r)$ a matrix of dimension $(L-1) \times 1$ that depends on ε_t^r, and $D(\varepsilon_t^r) \equiv \boldsymbol{\Psi}_{0,LL}(\varepsilon_t^r)$ a scalar. Notice that only the last column of $\boldsymbol{\Psi}_0$ depends on ε_t^r.

We will make use of the following lemma:

Lemma 2 *Consider the same matrix* Γ *as in Lemma 1. We have*

$$\det \boldsymbol{\Gamma} = \det(\mathbf{A}) \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}).$$

Proof. Rewrite Γ as

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{C} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{A}^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \end{pmatrix}$$

and the lemma follows. ■

Using Lemma 1 and noting that $D(\varepsilon_t^r)$ is a scalar, we have that the inverse of $\boldsymbol{\Psi}_0$ satisfies

$$\boldsymbol{\Psi}_0^{-1} = \frac{1}{D(\varepsilon_t^r) - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}(\varepsilon_t^r)} \begin{pmatrix} \left(D(\varepsilon_t^r) - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}(\varepsilon_t^r)\right) \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\varepsilon_t^r) \\ -\mathbf{C}\mathbf{A}^{-1} & 1 \end{pmatrix}.$$

The last row of the system $\boldsymbol{\varepsilon_t} = \boldsymbol{\Psi}_0^{-1}\mathbf{u}_t$ provides the equation $\varepsilon_t^r = \frac{1}{D(\varepsilon_t^r) - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}(\varepsilon_t^r)} \begin{pmatrix} -\mathbf{C}\mathbf{A}^{-1} & 1 \end{pmatrix} \mathbf{u}_t$, which defines ε_t^r. Since the right hand side of that equation only depends on ε_t^r through $D(\varepsilon_t^r) - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}(\varepsilon_t^r)$, the sign of the right hand side depends on ε_t^r only through the sign of $D(\varepsilon_t^r) - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}(\varepsilon_t^r)$.[59] Using Lemma 2, this means that the sign of the right hand side depends on ε_t^r only through the sign of $\det \boldsymbol{\Psi}_0$. Thus, with $sgn(\det \boldsymbol{\Psi}_0^+) = sgn(\det \boldsymbol{\Psi}_0^-)$, the sign (and value) of ε_t^r is uniquely pinned down, so that with $\mathbf{A}$ invertible, the system (17) has a unique solution vector. ■

Proposition 3 states that the system $\mathbf{u}_t = \boldsymbol{\Psi_0}(\varepsilon_t^r)\boldsymbol{\varepsilon_t}$ determines a unique solution vector $\boldsymbol{\varepsilon_t}$ as long as both $sgn\left(\det \boldsymbol{\Psi}_0^+\right) = sgn\left(\det \boldsymbol{\Psi}_0^-\right)$, i.e., as long as the asymmetry is not too strong. In practice, we impose this restriction by assigning a minus infinity value to the likelihood whenever $sgn(\det \boldsymbol{\Psi}_0^+) \neq sgn(\det \boldsymbol{\Psi}_0^-)$.

[59] In fact, we have $D(\varepsilon_t^r) - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}(\varepsilon_t^r) = \boldsymbol{\Psi}_{0,LL}(\varepsilon_t^r) - \sum_{\ell=1}^{L-1} \left(\mathbf{C}\mathbf{A}^{-1}\right)_\ell \boldsymbol{\Psi}_{0,\ell L}(\varepsilon_t^r)$.

Then, to construct the likelihood, we proceed as described in the recursive identification section by using the fact that there is a one-to-one mapping from $\boldsymbol{\varepsilon_t}$ to $\boldsymbol{\Psi}_0(\boldsymbol{\varepsilon_t})\boldsymbol{\varepsilon_t}$.

3.2 Asymmetric and state-dependent impulse response functions

For clarity of exposition, we consider the simpler case of a univariate state variable $z_t \in [\underline{z}, \overline{z}]$ with $\overline{z} = \max_{t\in[1,T]}(z_t)$ and $\underline{z} = \min_{t\in[1,T]}(z_t)$. With asymmetric and state dependent effects of ε_t^r, we can establish the proposition

Proposition 4 *Consider the non-linear moving average model defined in (6) with*

$$\boldsymbol{\Psi}_k(\boldsymbol{\varepsilon_{t-k}}, z_{t-k}) = \left[\boldsymbol{\Psi}_k^+(z_{t-k})1_{\varepsilon_{t-k}^r>0} + \boldsymbol{\Psi}_k^-(z_{t-k})1_{\varepsilon_{t-k}^r<0}\right], \;\; \forall k \in \{0,..,K\}, \;\; \forall t \in \{1,..,T\} \tag{24}$$

with ε_t^r the structural shock identified from sign restrictions. Then, given $\{\boldsymbol{y_t}\}_{t=1}^T$, given the model parameters and given K initial values of the shocks $\{\boldsymbol{\varepsilon}_{-K}...\boldsymbol{\varepsilon}_0\}$, the series of shocks $\{\boldsymbol{\varepsilon_t}\}_{t=1}^T$ is uniquely determined provided that $sgn(\det \boldsymbol{\Psi}_0^+(z_t)) = sgn(\det \boldsymbol{\Psi}_0^-(z_t))$, $\forall z_t \in [\underline{z}, \overline{z}]$.

Proof. Proceed as in the proof of Proposition 3. ■

Proposition 4 states that the system $\mathbf{u}_t = \boldsymbol{\Psi_0}(\varepsilon_t^r, z_t)\boldsymbol{\varepsilon_t}$ determines a unique solution vector $\boldsymbol{\varepsilon_t}$ as long as $sgn\left(\det \boldsymbol{\Psi}_0^+(z_t)\right) = sgn\left(\det \boldsymbol{\Psi}_0^-(z_t)\right)$ is independent of the value of z_t, i.e., as long as state dependence is not too strong. In practice, we can impose this restriction by assigning a minus infinity value to the likelihood whenever $sgn(\det \boldsymbol{\Psi}_0^{\pm}(\underline{z})) \neq sgn(\det \boldsymbol{\Psi}_0^{\pm}(\overline{z}))$.

Constructing the likelihood then proceeds as described in the previous section on recursive identification.

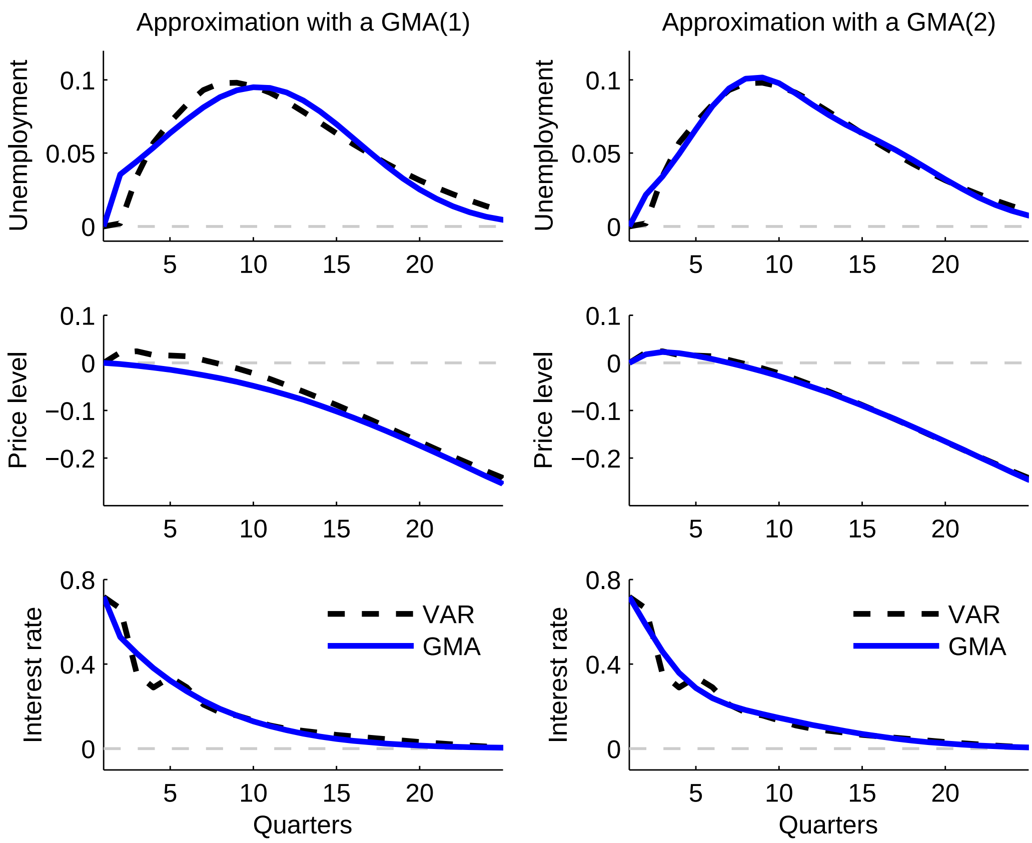

Figure 1: Impulse response functions of the unemployment rate (in ppt), the (log) price level (in percent) and the federal funds rate (in ppt) to a one standard-deviation monetary shock. Impulse responses estimated with a VAR (dashed-line) or approximated using one Gaussian basis function (GMA(1), left-panel, thick line) or two Gaussian basis functions (GMA(2), right panel thick line). Estimation using data covering 1959-2007.

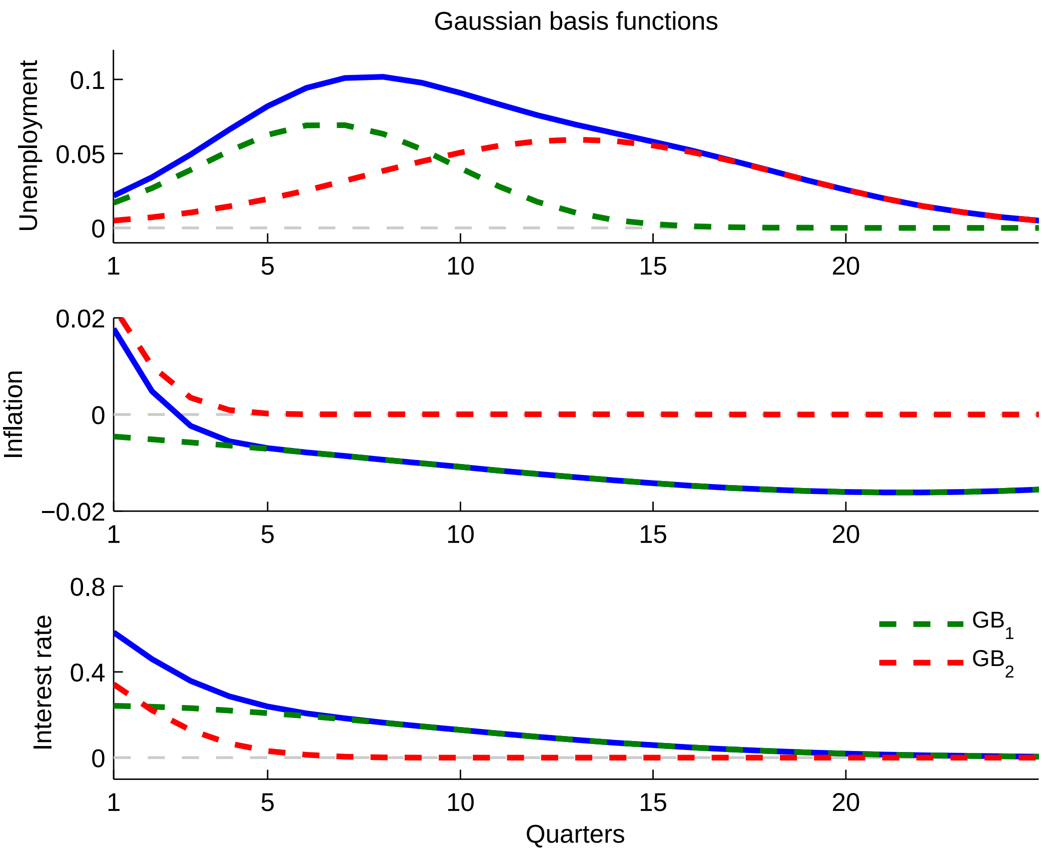

Figure 2: Gaussian basis functions (dashed lines) used by a GMA(2) to approximate the responses of unemployment, inflation and the fed funds rate to a monetary shock. The basis functions are appropriately weighted so that their sum gives the GMA(2) parametrization of the impulse response functions (solid lines) reported in the right-panels of Figure 1.

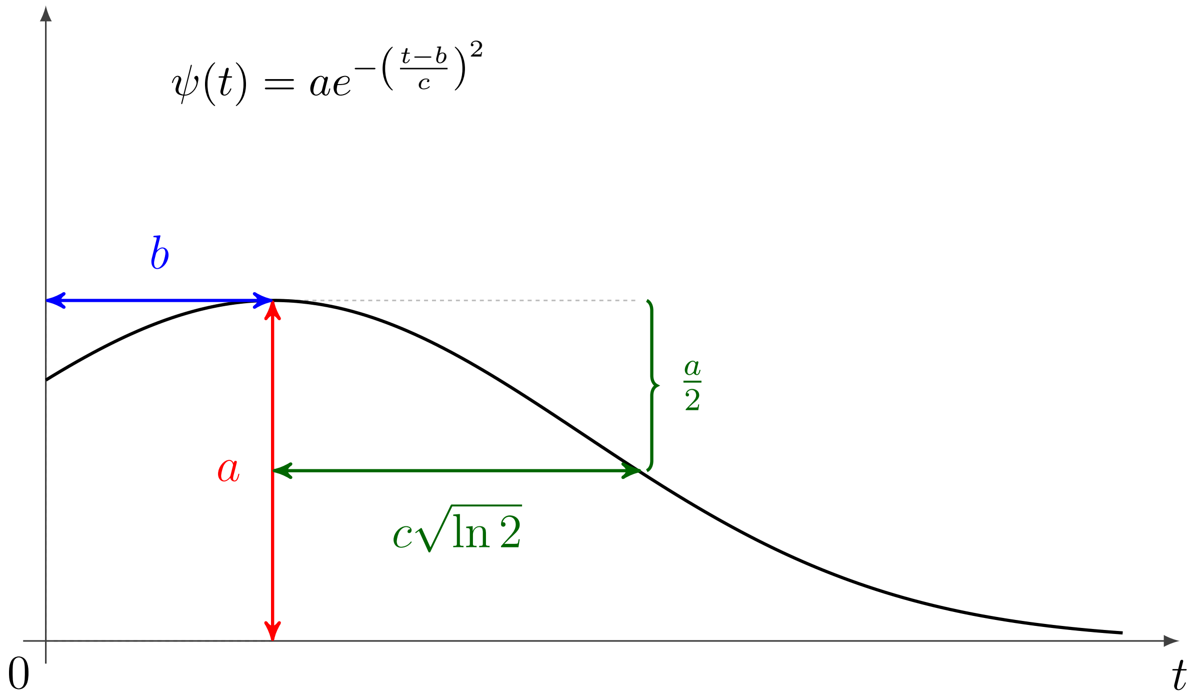

Figure 3: Interpreting an impulse response function with a GMA(1) model.

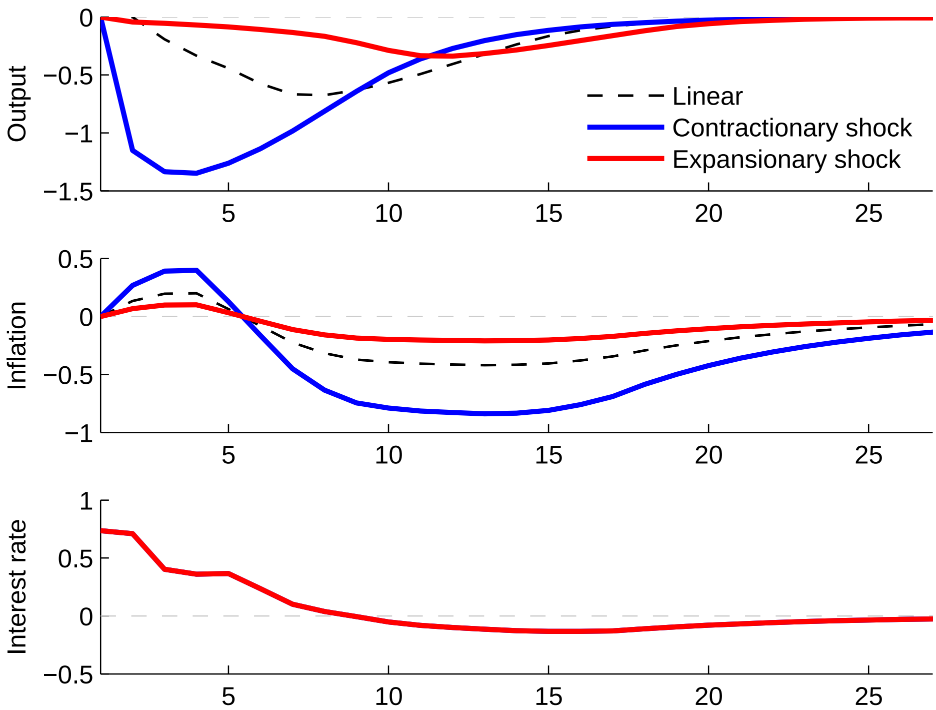

Figure 4: Monte Carlo simulation with asymmetric impulse responses to monetary shocks. The thick blue lines report the simulated impulse responses to a contractionary shock, and the thick red lines report the simulated impulse responses to an expansionary shock (with the responses to an expansionary shock multiplied by -1 for clarity of exposition). The dashed lines are the impulse responses estimated from a VAR over 1959-2007.

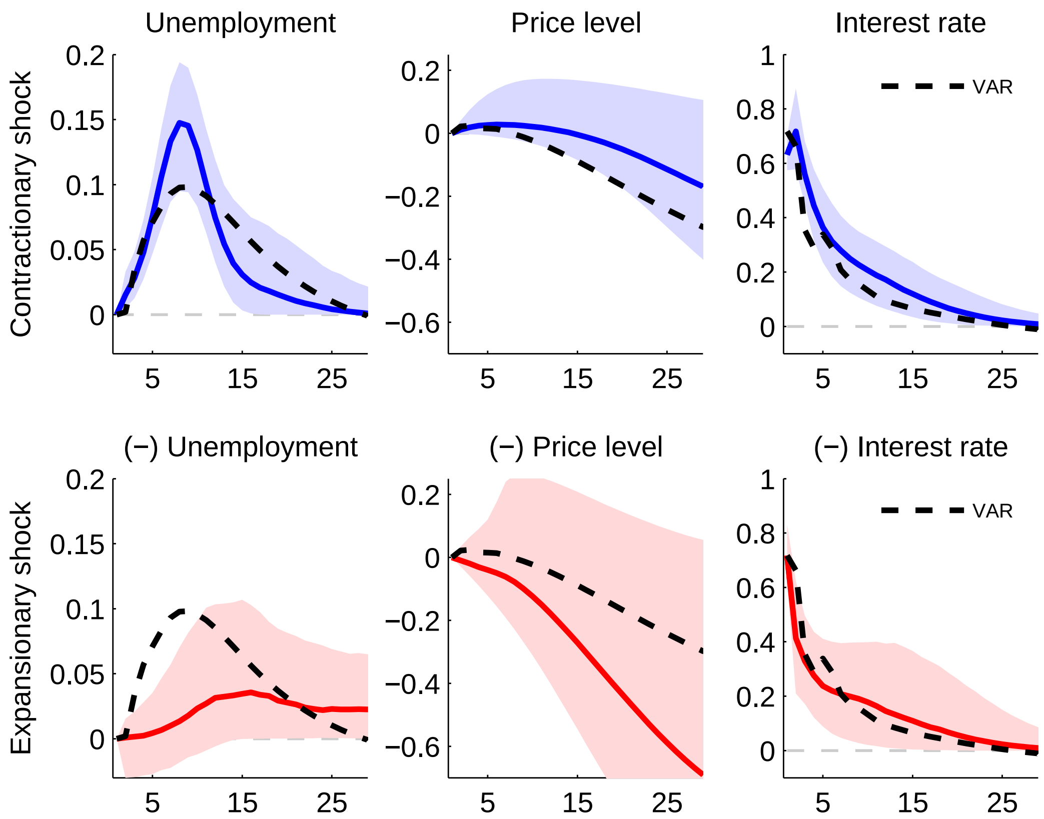

Figure 5: Impulse response functions of the unemployment rate (in ppt), the (log) price level (in percent) and the federal funds rate (in ppt) to a one standard-deviation monetary shock identified from a recursive ordering. Estimation from a VAR (dashed-line) or from a GMA(2) with asymmetry (plain line). Shaded bands denote the 5th and 95th posterior percentiles. For ease of comparison, responses to the expansionary shock are multiplied by -1. Estimation using data covering 1959-2007.

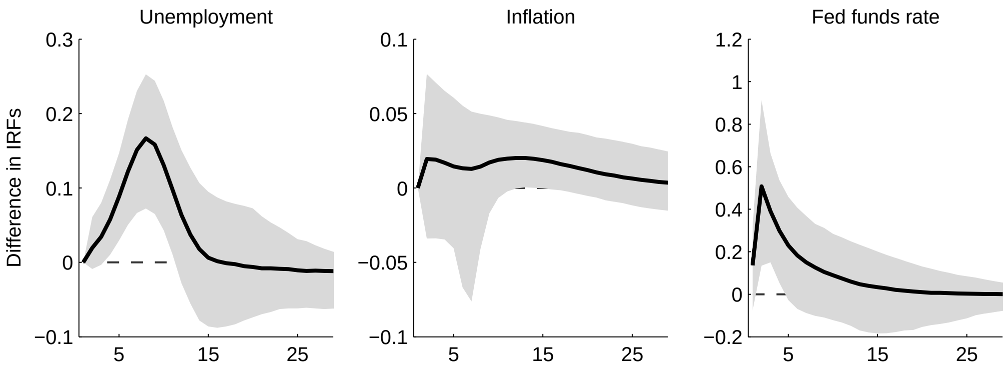

Figure 6: Differences in impulse response functions of the unemployment rate (in ppt), the (log) price level (in percent) and the federal funds rate (in ppt) to a one standard-deviation monetary shock. Shaded bands denote the 5th and 95th posterior percentiles. Estimation using data covering 1959-2007.

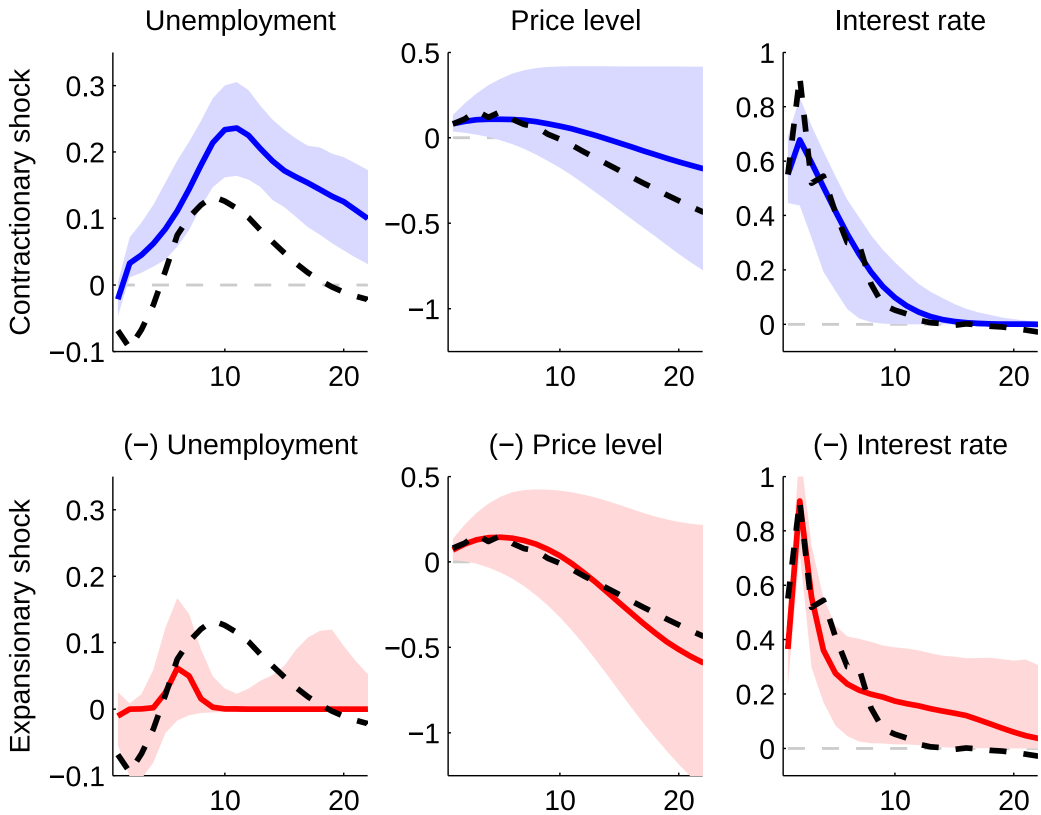

Figure 7: Impulse response functions of the unemployment rate (in ppt), the (log) price level (in percent) and the federal funds rate (in ppt) to a one standard-deviation Romer and Romer monetary shock. Estimation from a VAR (dashed-line) or from a GMA(2) with asymmetry (plain line). Shaded bands denote the 5th and 95th posterior percentiles. For ease of comparison, responses to the expansionary shock are multiplied by -1. Estimation using data covering 1966-2007.

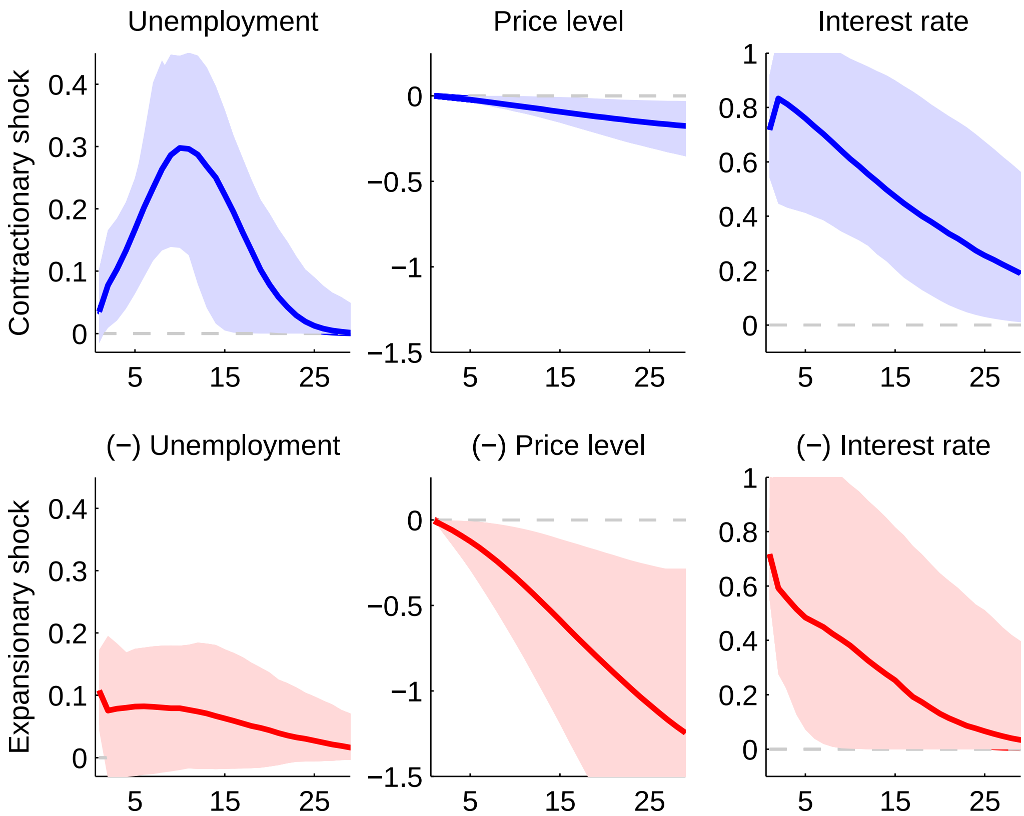

Figure 8: Impulse response functions of the unemployment rate (in ppt), the (log) price level (in percent) and the federal funds rate (in ppt) to a one standard-deviation monetary shock identified with sign restrictions. Estimation from a GMA(1) with asymmetry (plain line). Shaded bands denote the 5th and 95th posterior percentiles. For ease of comparison, responses to the expansionary shock are multiplied by -1. Estimation using data covering 1959-2007.

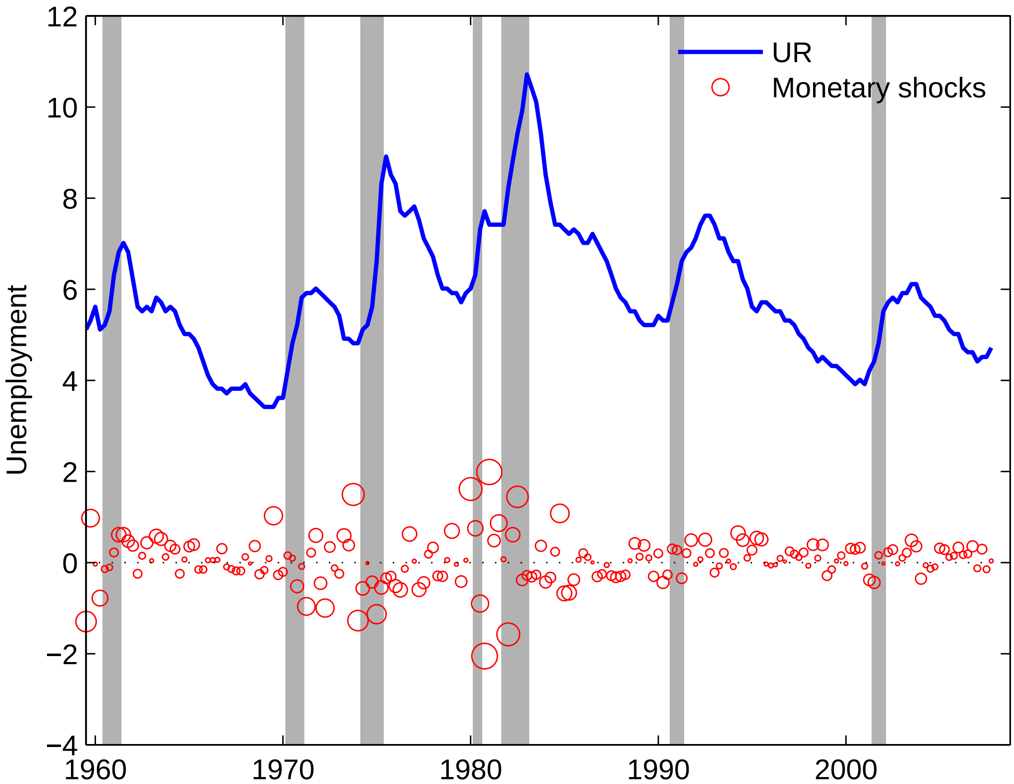

Figure 9: Unemployment rate –the business cycle indicator (solid line, left scale)–, and estimated monetary shocks (circles, right scale) with larger circles indicating larger shocks.

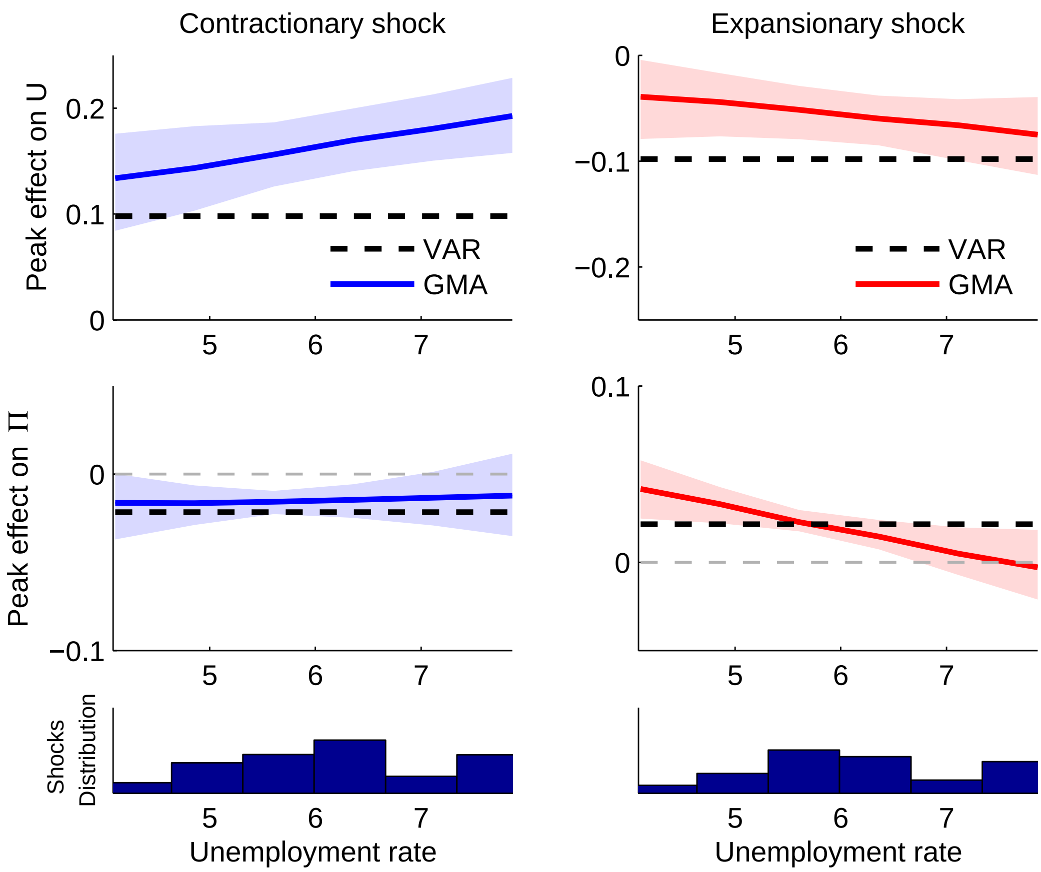

Figure 10: Peak effect of monetary policy on unemployment and inflation (in ppt) as a function of the state of the business cycle (measured with the unemployment rate) for one standard deviation contractionary monetary shocks (left panel) and expansionary monetary shocks (right panel). The dashed lines represent the 5th and 95th posterior percentiles. The thick-dashed line is the linear VAR estimate. The bottom panel plots the distribution of (respectively) contractionary shocks and expansionary shocks over the business cycle. Estimation using data covering 1959-2007.

Table 1: Summary statistics for Monte Carlo simulation with a linear model

	U		π		ffr	
	VAR	**GMA**	**VAR**	**GMA**	**VAR**	**GMA**
MSE	0.057	0.043	0.077	0.041	0.003	0.002
Avg length (at peak effect)	0.16	0.13	0.27	0.11	0.05	0.03
Coverage rate (at peak effect)	0.94	0.83	1	0.78	0.94	0.93

Note: Summary statistics over 50 Monte-Carlo replications. MSE is the mean-squared error of the estimated impulse response function over horizons 1 to 25. Avg length is the average distance between the lower (2.5%) and upper (97.5%) confidence bands at the time of peak effect of the monetary shock. The coverage rate is the frequency with which the true value lays within 95 percent of the posterior distribution. The VAR estimates and confidence bands are obtained from a Bayesian VAR with Normal-Whishart priors. U, π and ffr denote respectively unemployment, inflation and the fed funds rate.

Table 2: Summary statistics for Monte Carlo simulation with asymmetry

	a^+-a^-		
	gdp	**π**	**ffr**
Frequency of rejection of zero coefficient	0.94	0.90	0.08
Mean (true value)	-0.82 (-1.00)	-0.50 (-0.60)	0.03 (0.00)
Std-dev	0.28	0.17	0.12
Coverage rate	0.82	0.86	0.88

Note: Summary statistics over 50 Monte-Carlo replications. For each coefficient of interest, "Frequency of rejection of zero coefficient" is the frequency that 0 lies outside 90 percent of the posterior distribution, and "Coverage rate" is the frequency with which the true value lies within 90 percent of the posterior distribution. gdp, π and ffr denote respectively output, inflation and the fed funds rate.

Table 3: Summary statistics for Monte Carlo simulation with asymmetry and state dependence

	γ^{+}-γ^{-}		α^{+}-α^{-}		γ^{+}		γ^{-}	
	gdp	**π**	**gdp**	**π**	**gdp**	**π**	**gdp**	**π**
Frequency of rejection of zero coefficient	0.96	0.03	0.82	0.80	0.87	0.06	0.20	0.05
Mean (true value)	0.96 (1.00)	0.02 (0.00)	-0.78 (-1.00)	-0.48 (-0.60)	0.71 (1.00)	0.00 (0.00)	-0.21 (0.00)	-0.00 (0.00)
Std-dev	0.26	0.17	0.37	0.23	0.31	0.19	0.23	0.19
Coverage rate	0.84	0.92	0.71	0.70	0.68	0.92	0.65	0.90

Note: Summary statistics over 50 Monte-Carlo replications. For each coefficient of interest, "Frequency of rejection of zero coefficient" is the frequency that 0 lies outside 90 percent of the posterior distribution, and "Coverage rate" is the frequency with which the true value lies within 90 percent of the posterior distribution. gdp and π denote respectively output and inflation.

Table 4: Marginal data densities

	VAR	**GMA(1)**	**GMA(1) Asymmetry**	**GMA(2) Asymmetry**	**GMA(3) Asymmetry**	**GMA(2) Asymmetry State dep.**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
(log) marginal data density	112	118	127	138	107	158

Note: Trivariate models with unemployment, PCE inflation and the fed funds rate estimated over 1959-2007. The VAR estimates and confidence bands are obtained from a Bayesian VAR with Normal-Whishart priors.