Sign restricted Smooth Transition VAR models*

Martin Bruns†,‡ Michele Piffer‡,§

September 8, 2016

Abstract

We develop two algorithms to identifying Smooth Transition Vector Autoregressive models (STVARs) using sign restrictions. The aim is to contribute to the literature by showing that STVARs can be identified using approaches that differ from the recursive identification, which is instead the one almost exclusively used for such models. We show an application to the study of monetary policy shocks in expansions and in recessions. We find no observationally equivalent structural representation alternative to the recursive identification in which monetary shocks are more effective in expansions. This suggests that monetary shocks are more effective in recessions rather than in expansions, confirming previous findings from the literature.

JEL classification: C32, E52.
Keywords: Nonlinear models, set identification, monetary policy shocks.

---

*We are thankful to Helmut Lütkepohl for helpful comments and suggestions.
†Freie Universität Berlin, Department of Economics.
‡German Institute for Economic Research (DIW Berlin).
§Corresponding author e-mail: m.b.piffer@gmail.com
1 Introduction

The economic literature employs Structural Vector Autoregressive models (SVARs) to address a very wide set of research questions. Standard applications of these models include the analysis of the effects of monetary policy shocks, fiscal shocks, uncertainty shocks, news shocks, and others (see Ramey, 2016 for a survey). Since the crucial step in these models consists of mapping structural economically meaningful shocks into reduced form VAR innovations, extensive research has been devoted to the development of identification strategies that achieve this step. These include the popular recursive identification, identification through sign restrictions, long run restrictions, heteroskedasticity, external instruments, and others (Kilian, 2013).

In their linear form, SVAR models cannot address selected research questions. For example, linear SVARs cannot study whether the effects of a structural shock depends on when the shock hits the economy, or whether the effects of a shock depend on which additional shocks hit the economy in the future. For this reason, the literature has developed extensions of the baseline model to a nonlinear framework that allows to model nonlinearities in the data. One of these nonlinear model is the Smooth Transition Vector Autoregressive (STVAR) model. STVARs have been widely used in applied work since the work by Anderson and Vahid (1998) and Auerbach and Gorodnichenko (2012), who generalized to a multivariate case the earlier contribution by Granger and Terasvirta (1993) on univariate specifications.

While, as mentioned above, linear SVAR models are identified in the literature using a very rich battery of candidate identifying strategies, nonlinear models tend to be identified using almost exclusively the recursive identification. This choice can be rationalized through the rich set of technicalities that nonlinear models involve at several stages of the analysis, and which impose additional challenges in adapting nonrecursive identification strategies to nonlinear models. However, since the literature on linear models highlights several limitations of the recursive identification, the almost exclusive use of the recursive identification in nonlinear models emerges as an
important limitation that requires being addressed.

This paper offers one step toward filling this gap by providing two algorithms that allow for the identification of structural STVAR models using sign restrictions. Sign restrictions have become extensively used in applied work following their introduction by Faust (1998), Canova and De Nicoló (2002) and Uhlig (2005), and offer a flexible identification approach that has been applied to several types of research questions. We develop the analysis of sign restrictions using identifying restrictions on the impact effect of the structural shocks, or equivalently using the so-called $B$ specification according to the terminology by Lütkepohl (2005). The approach can be generalized to a richer set of sign restrictions, and to set identification in general, including restrictions on the elasticities between variables and lagged effects.

We build our analysis on the work by Auerbach and Gorodnichenko (2012) and specify the general smooth transition model in the regressive form. We then depart from Auerbach and Gorodnichenko (2012) by not imposing a unique mapping of structural shocks into reduced form shocks at each point in time. The first algorithm developed abstracts from estimation uncertainty and computes nonlinear impulse responses that highlight how the effect of shocks across time changes when comparing observationally equivalent representations of the data. The second algorithm accounts for both estimation and identification uncertainty. Both algorithms depart from the contribution by Koop et al. (1996) and allow for the generalization of impulse responses across a new dimension, namely identification uncertainty, in addition to the dimensions considered by Koop et al. (1996), which are the past and the future horizons. In the model considered, the transition variables driving the evolution of the system is endogenous to the system, as in Caggiano et al. (2014) and Berger and Vavra (2014), who use the recursive identification to study uncertainty shocks and fiscal shocks, respectively.

After proposing sign restrictions for STVARs in a general framework, we show an illustrative application to monetary policy shocks in a 3-variate model. To focus on
the contribution of the paper, we apply the algorithms to a research question that has attracted attention within the literature, namely whether monetary policy shocks are more effective in expansions or in recessions. This question was initially addressed by Weise (1999) using a STVAR model identified recursively, and then more recently by Tenreyro and Thwaites (2016) using a univariate smooth transition model. As in these papers, we are interested in whether a monetary policy shock has stronger effects on prices and output during expansions or during recessions. We find that there exists no observationally equivalent model for which a monetary contraction, as identified by our approach, yields a stronger negative effect on prices and output during expansions than during recessions. This result is in line with the finding that monetary policy is more effective during recessions.

We are aware of only one paper that identifies a STVAR model using an identification approach that differs from the recursive one, namely Bolboaca and Fischer (2016). The authors use a STVAR model for the analysis of news shocks in expansions and recessions, and identify such shocks using restrictions on the forecast error variance decompositions, building on the identification of these shocks by Barsky and Sims (2011) in a linear framework. We differ from their contributions by using an identification strategy that can be applied to a wider set of structural shocks of interest. The paper also relates indirectly to Lütkepohl and Netsunajev (2014). They use a linear VAR model, postulate a univariate smooth transition in the reduced form variance of the model, and identify it using heteroskedasticity. We differ from their contribution by allowing for a smooth transition in the full model, and by allowing for the analysis to study how impulse responses differ across time.

The paper is organized as follows. Section 2 presents the model, its estimation and identification and 2 algorithms to construct and summarize nonlinear impulse responses. Section 3 presents an application of the model to the case of monetary policy shocks in recessions and expansions. Section 4 concludes.
2 The model

This section discusses STVAR models in a general framework, and outlines how we achieve set identification within this type of models.

2.1 The reduced form model

The reduced form model is written as

\[ y_t = g(z_{t-1}) \Pi_1 x_{t-1} + (1 - g(z_{t-1})) \Pi_2 x_{t-1} + u_t, \]  
\[ u_t \sim N(0, \Omega_t), \]  
\[ \Omega_t = g(z_{t-1}) \Omega_1 + (1 - g(z_{t-1})) \Omega_2, \]  
\[ g(z_{t-1}) = \frac{1}{1 + e^{\gamma(z_{t-1} - c)}} \quad \gamma > 0. \]

In equation (1a), \( y_t \) is a \( k \times 1 \) vector containing the variables of the model, \( x_{t-1} = (1, y_{t-1}, ..., y_{t-p})' \) is a \( kp + 1 \times 1 \) vector containing the regressor for the constant terms and the \( p \) lags of the variables of the model, and \( g(z_{t-1}) \) is a scalar transition function explained below. The coefficients for the constant terms and the autoregressive components of the models are contained in the \( k \times kp + 1 \) matrices \( \Pi_1 \) and \( \Pi_2 \), while \( u_t \) represents the \( k \times 1 \) reduced form shocks. The reduced form shocks are assumed to be normally distributed as from equation (1b), with a time varying covariance matrix \( \Omega_t \) whose evolution across time is given by equation (1c).

The evolution of the parameters of the model are ruled by the logistic transition function \( g(z_{t-1}) \) in equation (1d). \( g(z_{t-1}) \), which takes values in the interval \([0, 1]\), drives the model as a time varying linear convex combination of models (2a) and (2b), with

\[ y_t = \Pi_1 x_{t-1} + u_t \quad u_t \sim N(0, \Omega_1), \]  
\[ y_t = \Pi_2 x_{t-1} + u_t \quad u_t \sim N(0, \Omega_2). \]
Model (1) moves between equations (2a) and (2b) depending on the evolution of the transition variable \( z_{t-1} \). In the analysis in this paper, \( z_{t-1} \) is a function of \( \{y_{t-1}, y_{t-2}, \ldots, y_{t-m}\} \) and is hence endogenous to the system through the last \( m \) observations of \( y_t \), excluding the contemporaneous observations. Given \( \gamma > 0 \), \( g(z_{t-1}) \) is an increasing function in \( z_{t-1} \). The slope parameter \( \gamma \) and the location parameter \( c \) control for the shape of the transitions and the midpoint of the transition, respectively.\(^1\) In the limit case of \( z_{t-1} \to -\infty \) and \( z_{t-1} \to \infty \), the system converges to the linear models (2a) to model (2b). The restriction that the transition variables is the same for equation (1a) and (1c) can be abandoned, as in Teräsvirta et al. (2014) and in Bolboaca and Fischer (2015).

As discussed in Auerbach and Gorodnichenko (2012), and as summarized in Appendix A of this paper, the estimation of model (1) can be achieved by means of the Metropolis-Hastings algorithm applied to the likelihood function of the model. In short, the algorithm involves iterating over \( \Omega_1 \) and \( \Omega_2 \) and computing the corresponding \( \Pi_1 \) and \( \Pi_2 \) using the Generalized Least Square estimator until the likelihood-function reaches a maximum. This procedure delivers estimates for \( \Pi_1, \Pi_2, \Omega_1 \) and \( \Omega_2 \), and hence it delivers estimates for the set \( \{\Omega_t\}_{t=l}^T \) through equation (1c), with \( l = max\{m, p\} \), given \( m \) the moving average window and \( p \) lag length of the VAR.

In several applications in the literature, \( c \) and \( \gamma \) are calibrated rather than estimated. We follow this approach in Section 3.

\(^1\)Higher values of \( \gamma \) imply more abrupt transitions from one regime to the other, given a variation in \( z_{t-1} \). Higher values of \( c \) imply that the first regime receives more and the second regime receives less weight. In the limit case of \( \gamma \to \infty \), the model converges to the Threshold VAR with \( c \) representing the threshold beyond which an abrupt regime switch occurs.
2.2 Identification

The reduced form shocks of the model are viewed as time varying linear combinations of the underlying structural shocks, according to equation

\[ u_t = B_t \epsilon_t. \]  

(3)

Matrix \( B_t \) is the matrix of impulse vectors at time \( t \) and contains \( k \) columns (the impulse vectors) associated with the \( k \) structural shocks. We adopt the normalization \( V(\epsilon_t) = I_k \), in accordance to which the structural shocks \( \epsilon_t \) have a time invariant variance-covariance matrix, which is set to the identity matrix. The invariance of \( V(\epsilon_t) \) over time is without loss of generality and is not intended as a restriction on the shocks driving the data.\(^2\) Under the normalizazion used, the covariance restrictions for the identification of the model are given by \( B_t B_t' = \Omega_t, \forall t \).

Model (1) is typically identified in the literature using the recursive identification (Auerbach and Gorodnichenko, 2012, Caggiano et al., 2014). This consists of using \( B_t = B_t^c \), with \( B_t^c \) the Cholesky decomposition of \( \Omega_t \). Since \( \Omega_t \) is identified, and since the Cholesky decomposition is unique up to sign convention of the shocks, \( B_t^c \) is identified. The set \( \{B_t^c\}_{t=1}^T \) can then be used to compute time varying impulse responses under the recursive identification.

We depart from exact identification and use the set \( \{B_t^c\}_{t=1}^T \) to obtain identification through sign restrictions. Candidate matrices \( B_t \) are explored using equation

\[ B_t = B_t^c Q_t, \]  

(4)

\(^2\)This can be seen by rewriting the data generating process as \( u_t = \hat{B}_t \hat{\epsilon}_t \) with \( V(\hat{\epsilon}_t) = D_t \), \( D_t \) time varying and diagonal. Model (3) is then obtained through the normalization \( u_t = \hat{B}_t \hat{\epsilon}_t = \hat{B}_t M_t^{-1} M_t \hat{\epsilon}_t = B_t \epsilon_t \), with \( M_t = D_t^{-0.5} \), \( B_t = \hat{B}_t M_t^{-1} \) and \( \epsilon_t = M_t \epsilon_t \). By construction, \( V(\epsilon_t) = I \).

Impulse responses to a one standard deviation shock to variable \( j \) are then obtained through the impulse vector \( \hat{B}_{t,j} \sqrt{\hat{d}_{t,j}} \equiv B_{t,j} \cdot 1 \), with \( \hat{B}_{t,j} \) and \( B_{t,j} \) the \( j \) column of \( \hat{B}_t \) and \( B_t \), respectively, and \( \hat{d}_{t,j} \) the \( j, j \) element of \( D_t \). Accordingly, we can assess the effect of one standard deviation shocks without restricting the standard deviation to be time invariant. Whether a change over time in the response to a one standard deviation shock is driven by a different impact effect, a different size of the shock given or a combination of the two remains unidentified.
with $Q_t$ orthogonal matrices. In this general specification, $Q_t$ is allowed to differ across time, although a time invariant $Q$ can also be considered. Under the latter restriction, $B_t$ remains time varying, due to the time varying nature of $B_t^c$. To simplify notation, the derivations in this section use the additional restriction of $Q_t = Q, \forall t$, which can be easily abandoned adopting the algorithms accordingly.

The use of orthogonal matrices to explore sign restrictions for SVARs is standard practice in the literature (see, for example, Kilian and Murphy, 2012). However, doing so in a nonlinear framework raises new challenges, because it adds a dimension of uncertainty, namely identification uncertainty, upon a model that already inherits from its nonlinearities several dimensions of uncertainty. To highlight how we address this problem, we now briefly discuss in general terms how the computation of impulse responses changes when moving from a linear to a nonlinear framework. We then build on this discussion to develop two algorithms that implement sign restrictions in the nonlinear framework considered.

### 2.3 Computing impulse responses

Impulse responses are used in the literature to study the evolution of the variables of a model in response to a shock that hits the model. While impulse responses can be viewed as the outcome of different types of conceptual experiments, in the present paper we view impulse responses as a tool addressing the following thought experiment: “Given some initial point in time, how will the evolution of the variables of the model differ depending on whether, in addition to some other shocks, the economy is also hit by an additional structural shock of interest?” We now develop this thought experiment and outline how it is affected by the joint combinations of nonlinearities and set identification.

In a stationary linear SVAR model of the type

$$y_t = \Pi x_t + B\epsilon_t,$$

(5)
with \( x_t = (1, y_{t-1}, \ldots, y_{t-p})' \) and with known parameter values, the impulse responses \( \phi_{i,h} \) of \( y_t \) at horizons \( h = 0, \ldots, H \) to a shock of size \( \delta \) to variable \( i \) can be computed following six steps:

1. select a starting point \( \bar{x} \), which in the linear model consists of the unitary vector for the constant terms and the values for \( p \) vectors of the lagged dependent variables;

2. draw a vector of random shocks at each point \( h \) and summarize these pseudo structural shocks in \( \{ \epsilon^O_h \}_{h=0}^H \);

3. compute pseudo shocks \( \{ \epsilon^I_h \}_{h=0}^H \) that augment \( \{ \epsilon^O_h \}_{h=0}^H \) with a shock of magnitude \( \delta \) to equation \( i \) by setting \( \epsilon^I_h = \epsilon^O_h + e_i \delta, \ h = 0, \) and \( \epsilon^I_h = \epsilon^O_h, \ h > 0, \) given \( e_i \) the \( k \times 1 \) selection vector of zeros with value 1 in row \( i \);

4. compute pseudo reduced form shocks \( \{ u^O_h \}_{h=0}^H \) and \( \{ u^I_h \}_{h=0}^H \) as \( u^I_h = B \epsilon^I_h \) and \( u^O_h = B \epsilon^O_h \), respectively;

5. generate pseudo data \( \{ y^O_h \}_{h=0}^H \) and \( \{ y^I_h \}_{h=0}^H \) using \( \bar{x} \) as the starting point and recursively feeding into the model the pseudo shocks \( \{ u^O_h \}_{h=0}^H \) and \( \{ u^I_h \}_{h=0}^H \), respectively;

6. compute impulse responses as \( \phi_h = y^I_h - y^O_h, \ h = 0, \ldots, H, \) and collect them into the \( k \times H + 1 \) matrix \( \Phi_i \), or more formally \( \Phi_i(\bar{x}, \{ \epsilon^O_h \}, \delta). \)

As already discussed in the literature (Lütkepohl, 2005, section 2.3), in a linear model it holds that \( \Phi_i(\bar{x}, \{ \epsilon^O_h \}, \delta) = \Phi_i \delta. \) Put it differently, impulse responses do not depend on the past (the initial point \( \bar{x} \)), nor on the present and the future (the pseudo shocks \( \{ \epsilon^O_h \}_{h=0}^H \)), and they are a scalar multiple of \( \delta \). Hence, in a linear SVAR, the effects of a shock do not depend on when the shock hits the economy, nor on which other shocks hit the economy in addition to the shock of interest. Thanks to these properties, impulse responses can be computed with the convenient close form solution provided by the moving average representation of the model.
The independence of the impulse responses from the past, present and future, as well as their proportionality to the shock of interest, do not generally hold in a non-linear setting. For example, in the STVAR model with an endogenous transition variable discussed in this paper and widely used in the literature, the size of the shock given, the starting point and the pseudo shocks affect the endogenous evolution of the system, and, through \( g(z_{t-1}) \), affect how past values map into future values. This means that in a nonlinear model the effect of a shock is not necessarily proportional to the shock given, and it depends, in general, on the point in time in which the shock is given, and on which other shocks hit the economy within the horizon considered. Impulse responses are still computed as outlined above, adjusting for the specific nonlinear model used. The heterogeneity in the responses according to initial points and present and future shocks is then typically summarized using the non-linear impulse response functions proposed by Koop et al. (1996), which average numerically across initial points and present and future shocks.\(^3\)

In this paper, the computation of impulse responses faces an additional degree dimension. In our framework, since the decomposition of \( \{\Omega_t\}_{t=1}^T \) into \( \{B_t\}_{t=1}^T \) is not unique, identification uncertainty enters step 4 of the above computation. This implies that impulse responses are defined as \( \Phi_i(\bar{x}, \{\epsilon^O_h\}, \delta, Q) \) rather than \( \Phi_i(\bar{x}, \{\epsilon^O_h\}, \delta) \), because not only they are subject to uncertainty on the past, present and future, but they depend on the mapping of structural shocks into reduced form shocks through \( Q \), the orthogonal matrix in equation (4). Since there is typically an entire set \( \{Q_i\}_{i=1}^n \) of orthogonal matrices that satisfies identifying restrictions, the computation of impulse responses involves a new dimension across which to generalize impulse responses. For this reason, we extend to identification uncertainty the nonlinear impulse response.

\(^3\)We do not use the term ”Generalized Impulse Response Function (GIRF)” because there is some disagreement in the literature about the dimension along which the generalization is being carried out. While some researchers use this term to express that this concept can be applied to linear as well as non-linear models, others, e.g. Pesaran and Shin (1998), argue that GIRFs are more general than traditional impulse response functions even in the linear case. We will use the term ”nonlinear impulse responses” in order to avoid confusion.
functions by Koop et al. (1996).

2.4 Algorithms

We propose two algorithms to compute impulse responses in light of identification uncertainty. The first algorithm abstracts from estimation uncertainty. It allows addressing the question whether the results of the analysis conducted under the popular recursive identification are robust to observationally equivalent representations, in the spirit of Uhlig (2005). The second algorithm augments the first one by allowing for estimation uncertainty. It builds on the Bayesian approach by Arias et al. (2014), although it remains frequentist in spirit. We now briefly outline the algorithms and then provide a discussion.

The first algorithm requires some estimates for the reduced form parameters $\Pi_1, \Pi_2, \Omega_1, \Omega_2$ and a calibration of $c$ and $\gamma$, which then imply values for the set $\{\Omega\}_{t=l}^T$ and for the set $\{B_t^c\}_{t=l}^T$. It consists of six steps:

1. draw an orthogonal matrix $Q$ from the algorithm by Rubio-Ramirez et al. (2010) and use it to compute the set $\{B_t\}_{t=l}^T$ from equation (4). If all $B_t$ in $\{B_t\}_{t=l}^T$ satisfy the restrictions, proceed to the following steps, otherwise draw a new $Q$;

2. for $t = l$ (the first usable observation, given the lags of the model and the width of the moving average window of the transition variable) draw $J$ sets of pseudo shocks $\{\epsilon_{h}^{O} \}_{h=0}^{H} j, j = 1, ..., J$ from a multivariate standard normal distribution and construct the corresponding shocks $\{\epsilon_{h}^{L} \}_{h=0}^{H} j, j = 1, ..., J$. Then take the initial point $\bar{x} = \lambda_t$, where $\lambda_t$, defined as $(1, y_{t-1}, y_{t-2}, ..., y_0)'$, includes lags entering equation (1a) and entering the computation of the transition variable. Compute, for each $j$, the impulse responses $\Phi_t(\bar{x}, [\{\epsilon_{h}^{O}\}]_j, \delta, Q)$. This gives $J$ sets of impulse responses that have in common the size and the timing of the shock ($\delta$ and $t = l$, respectively), but that differ by the remaining shocks $([\{\epsilon_{h}^{O}\}]_j)$;
3. for \( t = l \), summarize the uncertainty regarding the \( J \) sets of present and future shocks by computing the median target by Fry and Pagan (2011). Given \( Q \), this delivers one set \( \Phi_i(\bar{x}, [\{c_t^O\}]_{mt}, \delta, Q) \) of impulse responses, which are the impulse responses associated with the pseudo shocks \([\{c_t^O\}]_{mt}\) corresponding to the median target model;

4. repeat steps 2 and 3 for the remaining \( t = l + 1, \ldots, T \) periods, always drawing new pseudo shocks and using \( \bar{x} = \lambda_t \) as starting point, and collect the median target response of each period, for a total of \( T - l + 1 \) impulse responses;

5. repeat steps 1 to 3 a desired number of times drawing new matrices \( Q \), obtaining, for each period \( t = l, \ldots, T \), as many impulse responses as orthogonal matrices that satisfy the restrictions.

The second algorithm accounts for estimation uncertainty and consists of four steps:

1. draw one set of reduced form parameters from the Monte Carlo Chain mentioned in Section 2.1 and discussed in Appendix A;

2. compute the set \( \{B^c_t\}_{t=l}^{T} \) corresponding to the reduced form parameters drawn. Draw one orthogonal matrix \( Q \) using the approach from step 1 in Algorithm 1 and compute the corresponding set \( \{B_t\}_{t=l}^{T} \);

3. if all \( \{B_t\}_{t=l}^{T} \) satisfy the restrictions, draw one set of pseudo shocks \( \{c_t^O\}_{h=0}^{H} \) for each period \( t = l, \ldots, T \) from a multivariate standard normal distribution, construct the corresponding shocks \( \{c_t^I\}_{h=0}^{H} \) and compute the impulse response \( \Phi_i(\bar{x}, [\{c_t^O\}], \delta, Q) \) using \( \bar{x} = \lambda_t \) as initial condition. This delivers one impulse response for each \( t = l, \ldots, T \), for a total of \( T - l + 1 \) sets of impulse responses that differ according to when the shock was given;

4. repeat steps 1 to 4 a desired number of times;
The first algorithm allows the isolation of identification uncertainty, because it displays, for each period \( t = l, \ldots, T \), how the median target impulse response across possible future shocks differs depending on which orthogonal matrix is used to map structural shocks into reduced form innovations. In Section 3, this algorithm is used to compare the results to the recursive identification, which corresponds to the case in which the matrix \( Q \) is set to the identity matrix. The results from the algorithm can be reported in many ways. In Section 3 we collect impulse responses according to whether the period \( t \) in which the shock was given belongs to the first or to the second regime from model (1), depending on the value of \( g(z_{t-1}) \) corresponding to \( \bar{x} = x_{t-1} \). We also report the endogenous evolution of the transition function.

The second algorithms mainly differs from the first one by drawing only one set of pseudo shocks for each candidate draw of \( Q \). Drawing from the algorithms a sufficiently large number of times ensures that the extractions reflect not only estimation and identification uncertainty, but also uncertainty on the shocks that will hit the economy within the horizon of the impulse response. Both algorithms can be easily extended to a time varying \( Q \) matrix by drawing a new \( Q \) for each history and verifying that the identifying restrictions are satisfied in all time periods.

For the first algorithm, there is a hard trade-off between computational feasibility and the inclusion of estimation uncertainty: It would be desirable to have a measure of estimation uncertainty, but this would involve re-estimating the model many times, which, given its computational intensity, is not feasible. Instead, we illustrate model uncertainty within the two regimes of interest by reporting the median target impulse responses across those observations, which belong to each regime. For the second algorithm, a fully-fledged Bayesian approach would be desirable. It involves estimating the posterior of structural parameters directly and avoids the shortcut of drawing from the retained draws of the Metropolis-Hastings chain as an approximation of the reduced form posterior and then combining it with a draw from \( Q \). This would yield a proper posterior distribution allowing computation of error bands reflecting estimation
and model uncertainty as well as variation over time. Given the non-linearity of the model as well as the issue of set-identification, drawing from the posterior of structural parameters poses challenges. In the present state of the paper we employ Monte Carlo methods.

3 An application to monetary policy shocks in recessions and in expansions

We now show an application of the methodology proposed in the previous section. We apply the identification approach discussed so far to a STVAR model for the study of monetary policy shocks on US data. Following Weise (1999), Lo and Piger (2005) and Tenreyro and Thwaites (2016), we address the question whether monetary policy shocks are more effective during recessions or during expansions. These monetary policy shocks are identified by imposing standard sign restrictions.

3.1 The Model

For the baseline specification of the model we use three variables: the effective Federal Funds Rate (FFR), Industrial Production (IP) and the Consumer Price Index (CPI). The last two variables enter the model in log differences. This choice of variables is standard within small VAR models on monetary policy, see for instance Gerlach and Smets (1995) and Sims et al. (2008). We use monthly data for the period 1964M11 through 2015M6. As transition variable $z_{t-1}$ we use a moving average on the endogenous variable IP, as in Auerbach and Gorodnichenko (2012). At each time $t$, the moving average is computed using the 12 months until $t$, excluded.

In the baseline specification of the model we calibrate rather than estimate the parameters $\gamma$ and $c$ that discipline the transition variable. This choice is dictated by computational convenience and follows a standard practice in the use of STVARs in
applied works, as for example Bachmann and Sims (2012), Berger and Vavra (2014) and Caggiano et al. (2014). For the baseline specification, we follow their work by setting $c = 0$ and by calibrating $\gamma = 1.5$ such that the system is in a recessionary period $\alpha = 14\%$ of the time, where a period is labelled as recessionary if $g(z_{t-1}) < 1 - \alpha$, with $\alpha$ estimated from the NBER recessions. In the baseline specification of the model we follow this conventional approach in order to improve comparability of the results and focus on the contribution of the present paper, which is on the identification of the model.

Conditioning on parameter values for $\gamma$ and $c$, the model simplifies by becoming linear in a transformation of variables. In particular, rewrite the model as

$$y_t = g(z_{t-1})\Pi_1 x_{t-1} + (1 - g(z_{t-1}))\Pi_2 x_{t-1} + u_t,$$  

$$(6a)$$

$$= [\Pi_1 \Pi_2] \begin{pmatrix} g(z_{t-1})x_{t-1} \\ (1 - g(z_{t-1}))x_{t-1} \end{pmatrix} + u_t$$

$$(6b)$$

$$= \Pi n_t + u_t$$

$$(6c)$$

$$\Omega_t = g(z_{t-1})\Omega_1 + (1 - g(z_{t-1}))\Omega_2.$$  

$$(6d)$$

The $k \times 2(1 + kp)$ matrix $\Pi$ includes all the autoregressive parameters of the model, while the new variable $n_t$ stacks the weighted constant and lagged variables corresponding to the two regimes underneath each other.

We follow Faust (1998) and Benati (2008) by imposing that a monetary policy shock is the only shock, which on impact increases the federal funds rate, decreases the CPI and decreases industrial production.

### 3.2 Results

The results from our baseline specification are presented in Figure 2 and 3. They show impulse responses in expansion and in recession, together with the difference between the two, to a one standard deviation monetary policy shock.
**Figure 1:** Impulse responses to a monetary tightening - Algorithm 1

**Figure 2:** Note: Algorithm 1: Generalized Impulse Responses to a 1 standard deviation Monetary Policy Shock. The left and middle panel show the GIRFs in expansions and recessions, respectively. Blue line: GIRF from recursive identification. Black lines: GIRF corresponding to different Q matrices. The right panel shows the difference between the two regimes, where the upper black line represents the pointwise maximum across models, the lower black line shows the pointwise minimum across models and the blue line the results from a recursive identification.
Figure 3: Impulse responses to a monetary tightening - Algorithm 2

Note: Generalized Impulse Responses to a 1 standard deviation Monetary Policy Shock. The left and middle panel show the GIRFs in expansions and recessions, respectively. Black solid line: Median GIRF. Black dashed lines: 68% pointwise bands around the median. Blue solid line: GIRF resulting from recursive identification. Blue dashed lines: 68% bands around this GIRF. The right panel shows the difference between the two regimes.
Figure 2 shows the result using Algorithm 1, which abstracts from estimation uncertainty. The left two columns plot four randomly chosen $Q$ matrices to illustrate the extent of model uncertainty. The first two rows show that, while both CPI and IP fall by construction, the model uncertainty surrounding this effect is considerable: For the randomly drawn $Q$ matrices it ranges from $-0.05$ to $-0.2$ percentage points for the case of CPI and from $-0.1$ to $-0.6$ percentage points for the case of IP. The initial drop of CPI and IP is not necessarily implied by the recursive identification. In fact, when using recursive identification, there is evidence of the so-called "price puzzle", i.e. an increase in prices following a monetary contraction. This phenomenon has already been noted by e.g. Uhlig (2005) and is avoided when using sign restrictions. The last column shows that there exists no observationally equivalent model for which a monetary contraction, as identified by our approach, yields a stronger negative effect on prices and output during expansions than during recessions. This result is in line with the finding that monetary policy is more effective during recessions also found by Tenreyro and Thwaites (2016) and Lo and Piger (2005). The last row shows the response of the transition function. It suggests that a negative monetary policy shocks triggers a transition toward the recessionary state.

Figure 3 shows the result using Algorithm 2, which incorporates estimation uncertainty. The bands around the responses thus represent estimation and model uncertainty as well as variation over time. The results concerning the effects on CPI and IP do not change much compared with the first algorithm. The major source of uncertainty seems to be model uncertainty.

The validity of the estimation hinges on the distinction between expansions and recessions. If one wrongly attributes an observation to a certain regime, this will bias the estimates and therefore invalidate the inference. This is why Figure 4 plots the transition function against recessionary periods as identified by the NBER. The transition function lies below the threshold value of 20 percent in most periods which the NBER labels recessions. An exception is the recessionary period in the mid-70s,
which the NBER considers to start earlier than our model does. Overall, there is evidence that recessionary periods as identified by our model overlap with the NBER labelling most of the time.

4 Conclusions

In the present paper we set up a SVAR model with smooth transition in mean and variance and propose two algorithms to partially identify structural shocks using sign restrictions. The advantages of this model are that its reduced form is relatively easy to estimate, that the combination with sign restrictions allows for the quantification of model uncertainty, and that it offers an alternative to the recursive identification, which in some models can be considered inadequate.

As an illustration we show how the STVAR model with sign restrictions can be applied to the question of monetary policy effectiveness across the business cycle, an issue which has been studied, among others, by Weise (1999), Lo and Piger (2005) and Tenreyro and Thwaites (2016). We confirm their finding that monetary policy has a stronger effect on prices and industrial production during recessions. However, the size of this effect is subject to considerable model uncertainty, which previous studies have not accounted for. The effect of a one standard deviation contractionary monetary policy shock on prices varies between -0.05 and -0.2 percentage points on impact and the effect on industrial production varies between -0.1 and -0.6 percentage points.
Appendix

Reduced Form Estimation

The model is estimated using Maximum-Likelihood methods. The log-likelihood of this model is given by:

\[
\log L(\Psi) = \text{const} + 0.5 \sum_{t=1}^{T} \log |\Omega_t| - 0.5 \sum_{t=1}^{T} u_t' \Omega_t^{-1} u_t
\]

with \(\Psi = \{\Pi_E, \Pi_R, \Omega_E, \Omega_R, \}\) the parameter space over which to maximize the likelihood and \(u_t = y_t - g(z_{t-1})\Pi_1 x_{t-1} + (1 - g(z_{t-1}))\Pi_2 x_{t-1}\) the reduced form errors. Note that \(\gamma\) and \(c\), the parameters governing the shape of the transition function, are omitted here because they are calibrated in the baseline. Due to the high non-linearity of this system, maximization of the likelihood follows an iterative Metropolis-Hastings procedure.

The key insight is that conditional on \(\{\Omega_E, \Omega_R\}\) the model is linear and so that \(\{\Pi_E, \Pi_R\}\) can be estimated via generalised least squares. The procedure then iterates on \(\{\Omega_E, \Omega_R\}\) until an optimum is reached. It has been put forward by (CH2003) and more recently applied by (AG2012)

Since (7) can possess multiple local maxima, the choice of the starting value is crucial. We choose as starting values for both \(\Omega_E\) and \(\Omega_R\) the covariance matrix resulting from a homoskedastic smooth transition model, which can be estimated via OLS. Formally, we set the initial candidate set of parameters \(\Theta(0) = \{\hat{\Omega}_{OLS}, \hat{\Omega}_{OLS}\}\).

Each chain element \(n\) then proceeds in the following steps:

In the first step, we draw a candidate set of parameters \(\Theta(n)\) as \(\Theta(n) = \Psi(n) + \psi(n)\), where \(\Psi(n)\) is the current state and \(\psi(n)\) is random noise. When adding \(\psi(n)\), one has to ensure that the resulting elements of \(\Theta(n)\) are both positive-semi-definite because they represent covariance matrices. That is why we first Cholesky-decompose the current elements of \(\Theta(n)\), then add normally distributed noise to the non-zero elements and
re-construct both elements by post-multiplying the resulting term with its transpose.

In the second step, we assess whether the new set of candidate parameters yields a likelihood-improvement. Formally, we take as the $n + 1$ chain element:

$$
\Phi_{i}^{(n+1)} = \begin{cases} 
\Theta^{(n)} & \text{with probability min}\{1, \exp[\log L(\Theta^{(n)}) - \log \Psi^{(n)}]\}, \\
\Psi^{(n)} & \text{otherwise.}
\end{cases}
$$

where $L(\Phi_{i}^{(n)})$ is the value of the objective function at the current state of the chain and $L(\Theta^{(n)})$ is the value of the objective function using the candidate vector of parameter values. If $\Theta^{(n)}$ is accepted, store $\Psi^{(n)} = \{\Pi_{1}^{(n)}, \Pi_{2}^{(n)}, \Omega_{1}^{(n)}, \Omega_{2}^{(n)}\}$ and $L(\Theta^{(n)})$.

In our application, we iterate 20,000 times and scale the variance of the noise $\psi^{(n)}$ to obtain an acceptance rate of roughly 0.3 as proposed in (Gelman et al. 2004). Convergence of the chain is suggested by a relatively stable evolution of the resulting chain of likelihood values. We then drop the first 18,000 chain elements and use the remaining 2,000 for inference.
Figure 4: Transition Function. Sample: 1962M8-2015M6. Solid line: Transition Function with calibrated $c$ and $\gamma$, taking a 12 month Moving Average of the growth rate of Industrial Production as Transition Variable. Dashed line: Cutoff value below which the system is considered to be in recession. Shaded Region: NBER. recessions
References


Bolboaca, M. and S. Fischer (2016). News shocks: different effects in boom and recession?


