ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Jordà, Òscar

Working Paper

Model-free impulse responses

Working Paper, No. 06-8

Provided in Cooperation with:
University of California Davis, Department of Economics

Suggested Citation: Jordà, Òscar (2004) : Model-free impulse responses, Working Paper, No. 06-8, University of California, Department of Economics, Davis, CA

This Version is available at:
https://hdl.handle.net/10419/31369

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Department of Economics
Working Paper Series

Model-Free Impulse Responses

Oscar Jorda
University of California, Davis

June 02, 2004

Paper # 06-8

This paper introduces methods for computing impulse response functions that do not require specification and estimation of the unknown dynamic multivariate system itself. The central idea behind these methods is to estimate flexible local projections at each period of interest rather than extrapolating into increasingly distant horizons from a given model, as it is usually done in vector autoregressions (VAR). The advantages of local projections are numerous: (1) they can be estimated by simple regression techniques with standard regression packages; (2) they are more robust to misspecification; (3) standard error calculation is direct; and (4) they easily accommodate experimentation with highly non-linear and flexible specifications that may be impractical in a multivariate context. Therefore, these methods are a natural alternative to estimating impulse responses from VARs. An application to a simple, closed-economy monetary model suggests that the output loss and inflation effects of an interest rate shock depend on the stage of the business cycle.

UCDAVIS

Department of Economics
One Shields Avenue
Davis, CA 95616
(530)752-0741

http://www.econ.ucdavis.edu/working_search.cfm

August 2003 (this version, January 2004)

Model-Free Impulse Responses*

Abstract

This paper introduces methods for computing impulse response functions that do not require specification and estimation of the unknown dynamic multivariate system itself. The central idea behind these methods is to estimate flexible local projections at each period of interest rather than extrapolating into increasingly distant horizons from a given model, as it is usually done in vector autoregressions (VAR). The advantages of local projections are numerous: (1) they can be estimated by simple regression techniques with standard regression packages; (2) they are more robust to misspecification; (3) standard error calculation is direct; and (4) they easily accommodate experimentation with highly non-linear and flexible specifications that may be impractical in a multivariate context. Therefore, these methods are a natural alternative to estimating impulse responses from VARs. An application to a simple, closed-economy monetary model suggests that the output loss and inflation effects of an interest rate shock depend on the stage of the business cycle.

- *Keywords:* impulse response function, local projection, vector autoregression, nonlinear.
- *JEL Codes:* C32, E47, C53.

1 Introduction

In response to the rigid identifying assumptions used in theoretical macroeconomics during the seventies, Sims (1980) provided what has become the standard in empirical macroeconomic research: vector autoregressions (VARs). Since then, researchers in macroeconomics often compute dynamic multipliers of interest (such as impulse responses and forecast-error variance decompositions) by specifying a VAR, even though the VAR per se is, often times, of no particular interest. However, VAR-based impulse responses are restrictive in a manner seldom recognized. In particular impulse responses are constrained to have the following properties[1] : (1) *symmetry,* responses to positive and negative shocks are mirror images of each other; (2) *shape invariance*, responses to shocks of different magnitudes are scaled versions of one another; (3) *history independence*, the shape of the responses is independent of the local conditional history; and (4) *multidimensionality*, responses are nonlinear functions of high-dimensional parameter estimates which complicate the calculation of standard errors and have the potential of compounding misspecification errors. In addition, a VAR is a representation of a system of linear, stochastic difference equations that may not appropriately represent general economic processes whose solutions are nonlinear stochastic difference equations instead.

Impulse responses (and variance decompositions) are important statistics in their own right and thus avoiding these constraints is a natural empirical objective. This paper introduces methods for computing impulse response functions for a vector time series that do not require specification and estimation of the unknown multivariate dynamic system itself. The central idea behind these methods is to use *local* projections (a term to be defined precisely in the next section) for each period of interest rather than extrapolating from a given model into increasingly distant horizons, as it is usually done in a VAR. The advantages of local projections are numerous: they are disarmingly simple to compute; they are more robust to misspecification; standard error calculation is direct;

[1] The following list of properties is mostly in Koop et al., 1996.

and they easily accommodate experimentation with highly non-linear and flexible specifications. Since estimation of these local projections can be done equation by equation, impulse response functions and their standard errors can be easily calculated with available standard regression packages, thus becoming a natural alternative to estimating impulse responses from VARs.

Although there is now a number of more complex, multivariate econometric models that relax some of the constraints implicit in VARs, systems of dynamic non-linear equations are often difficult to estimate and are impractical for computing impulse responses – there are no closed-form solutions and non-linear forecasts beyond one-period ahead require simulation techniques for their calculation. Instead, this paper argues in favor of divesting the object of interest from the primitive econometric specification of a model into methods for calculating the implied time profiles directly from the data, and therefore, in a manner robust to a wider array of model choices and specifications. The key insight is that most dynamic multivariate models (such as VARs) represent global approximations to the ideal data generation process (DGP) and are optimally designed for one-period ahead forecasting. Meanwhile, impulse responses describe the time profiles of variables at increasingly distant horizons, suggesting that a sequence of local approximations is preferable to a global one. Precursors of some of the ideas discussed below are Cox (1961), Tsay (1993), Lin and Tsay (1996) and Clements and Hendry (1998).

An advantage of calculating impulse responses by local projections is that forecasting accuracy increases as the forecast horizon increases relative to a wide class of model misspecification. Naturally, when the primitive model is correctly specified these projections will be less efficient. However, Monte Carlo evidence will show that this loss in efficiency is rather small. Another advantage of the local projection method is that standard errors for impulse responses are calculated directly from conventional regression output rather than from delta method approximations or with substantial computational effort (such as Monte Carlo, or bootstrap methods). Monte Carlo evidence provides support for these claims. The new methods are applied to a simple system for the output gap, inflation, and the federal funds rate. Such a system has become popular in

the literature that investigates the performance of monetary policy rules (see Galí, 1992, Fuhrer and Moore, 1995a, 1995b, and Taylor, 1999). In evaluating such rules, it is crucial to determine the relative trade-offs between inflation and output embodied by the Phillips curve. Tests of the null of linearity against the alternative of a threshold effect based on Hansen (2000) reveal that the responses of these trade-offs to monetary policy shocks depend on whether the economy is growing above or below potential. In particular, the results suggest that the loss of output due to an increase in interest rates is much smaller when the economy is below potential, a consideration of critical importance in designing an optimal policy response.

2 Impulse Responses by Local Projections: Estimation and Inference

2.1 Estimation

The concept of an impulse response function popularized by Sims' (1980) seminal paper is often and almost exclusively associated with linear multivariate Markov models – such as VARs – and their Wold decomposition. However, impulse responses are statistics that can almost always be calculated from any data generating process (DGP), even from those that do not have a Wold decomposition (see Koop et al. 1996; and Potter, 2000). The more general definition of an impulse response that I adopt in this paper is found in Hamilton (1994) and Koop et al. (1996) and is given by

$$\left.\frac{\partial \mathbf{y}_{t+s}}{\partial \boldsymbol{\delta}_t}\right|_{\boldsymbol{\delta}_t=\mathbf{d}_i} = E(\mathbf{y}_{t+s}|\boldsymbol{\delta}_t = \mathbf{d}_i; X_{t-1}) - E(\mathbf{y}_{t+s}|\boldsymbol{\delta}_t = \underline{\mathbf{0}}; X_{t-1}) \qquad s = 0, 1, 2, ... \tag{1}$$

where the operator $E(.|.)$ denotes the best, mean squared error predictor; $\mathbf{y}_t$ is an $n \times 1$ random vector; $X_{t-1} \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ...)'$; $\underline{0}$ is of dimension $n \times 1$; and D is an $n \times n$ matrix, whose columns $\mathbf{d}_i$ contain the relevant experimental shocks. It is worth clarifying with an example the meaning of these experimental shocks, $\mathbf{d}_i$.

Time provides a natural mechanism for organizing the causal linkages among the variables

in $\mathbf{y}_t$ but it is ineffective for identifying its contemporaneous causal relations. To overcome this deficiency, one common strategy is to assume a Wold-causal order for the elements of $\mathbf{y}_t$ in the triangular factorization of the contemporaneous, variance-covariance matrix (say Ω), conditional on the past. Therefore, if $\Omega = PP'$, where P is lower-triangular, then experimental shocks can be obtained by setting $D = P^{-1}$ and the ith column of D, $\mathbf{d}_i$, then represents the "structural shock" to the ith element in $\mathbf{y}_t$ in the usual parlance of the VAR literature. This type of identification assumption, while common, is not unique. The issue of identification is an important one but it is not the object of this paper. Instead, the paper proceeds by taking D as given by the practitioner's choice of identification assumptions and therefore subsequent results do not depend on this choice.[2]

Instead of calculating impulse responses from a pre-specified, multivariate model, consider computing the multi-step ahead forecasts required in definition (1) from projections of each $\mathbf{y}_{t+s}$ onto the linear space generated by $X_{t-1} \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ...)'$. I will use the term "local projections" to clarify that a different projection is estimated for each horizon s over which the impulse response is calculated, in contrast to a typical VAR, which is a simple projection of $\mathbf{y}_t$ onto X_{t-1}. The term "local projections" is therefore aptly evocative of nonparametric considerations. Local projections for $\mathbf{y}_{t+s}$ can be easily estimated by the sequence of least squares regressions

$$\mathbf{y}_{t+s} = \boldsymbol{\alpha}^s + B_1^{s+1}\mathbf{y}_{t-1} + B_2^{s+1}\mathbf{y}_{t-2} + ... + B_p^{s+1}\mathbf{y}_{t-p} + \mathbf{u}_{t+s}^s \qquad s = 0, 1, 2, ..., h \tag{2}$$

where $\boldsymbol{\alpha}^s$ is an $n \times 1$ vector of constants, the B_i^{s+1} are matrices of coefficients for each lag i and horizon $s+1$ (this timing convention will become clear momentarily). I truncate the projection at lag p, which can be determined by information criteria for each horizon s individually (in principle, there is no restriction that requires that all horizons share the same lag truncation). Naturally, impulse responses can be calculated up to a maximum horizon h that depends on the sample size

[2] For statistically-based methods of structural identification the reader is encouraged to consult Granger and Swanson (1997) and Demiralp and Hoover (2003).

and available degrees of freedom.

According to definition (1), the impulse responses calculated from (2) are

$$\left.\frac{\partial \widehat{\mathbf{y}_{t+s}}}{\partial \boldsymbol{\delta}_t}\right|_{\boldsymbol{\delta}_t=\mathbf{d}_i} = \widehat{B}_1^s \mathbf{d}_i \qquad s = 0, 1, 2, ..., h \tag{3}$$

with the obvious normalization $B_1^0 = I$. The parameters B_1^s are consistently estimated from simple least squares although, as we will see momentarily, the residuals $\mathbf{u}_{t+s}^s$ will not be white noise in general. This, however, poses no difficulty. These residuals will have an unknown, moving-average-type structure involving information dated $t, t+1, ..., t+s$ which by construction is uncorrelated with the regressors $\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-p}$. Expression (2) is reminiscent of the "adaptive forecasts" in Lin and Tsay (1996) or the "dynamic forecasts" in Clements and Hendry (1998) for which proofs of asymptotic consistency and normality are available in Weiss (1991).

Expression (2) describes a system of n linear equations that can be estimated equation by equation without loss of generality (since the regressors are common to all equations and there are no cross-equation restrictions). Therefore, the response of the ith variable at time $t+s$ to an experimental shock assigned to the jth variable is simply calculated from the univariate regression

$$\begin{aligned} y_{i,t+s} = \alpha_i^s + b_{i,1(1)}^{s+1} y_{1,t-1} + ... + b_{i,j(1)}^{s+1} y_{j,t-1} + ... + b_{i,n(1)}^{s+1} y_{n,t-1} + \\ \mathbf{b}_{i(2)}^{s+1} \mathbf{y}_{t-2} + ... + \mathbf{b}_{i(p)}^{s+1} \mathbf{y}_{t-p} + u_{i,t+s}^s \qquad s = 0, 1, 2, ..., h \end{aligned} \tag{4}$$

where α_i^s is the ith element of the vector of constants $\boldsymbol{\alpha}^s$, $b_{i,j(k)}^s$ denotes the (i,j) element of the matrix B_k^s, and $\mathbf{b}_{i(k)}^s$ is the ith row of the matrix B_k^s. The impulse response function thus becomes,

$$\left.\frac{\partial \widehat{y_{i,t+s}}}{\partial \boldsymbol{\delta}_t}\right|_{\boldsymbol{\delta}_t=\mathbf{d}_j} = \widehat{b}_{i,1(1)}^s d_{j,1} + ... + \widehat{b}_{i,j(1)}^s d_{j,i} + ... + \widehat{b}_{i,n(1)}^s d_{j,n} \qquad s = 0, 1, 2, ..., h.$$

and the corresponding normalization $b_{i,j(1)}^0 = 1$.

The local projections described in expressions (2)-(4) are also very useful in calculating the variance decompositions of the forecast error variances. In fact, these are easily calculated as a

by-product of the local projection at each horizon s. By definition, the error in forecasting $\mathbf{y}_t$, s periods into the future is given from expression (2) by

$$\mathbf{y}_{t+s} - E(\mathbf{y}_{t+s}|X_{t-1}) = \mathbf{u}^s_{t+s} \qquad s = 0, 1, 2, ...$$

from which the unnormalized mean squared error (MSE_u) is

$$MSE_u(E(\mathbf{y}_{t+s}|X_{t-1})) = E(\mathbf{u}^s_{t+s}\mathbf{u}^{s\prime}_{t+s}) \qquad s = 0, 1, 2, ..., h$$

The choice experiment D renormalizes the MSE into

$$MSE(E(\mathbf{y}_{t+s}|X_{t-1})) = D^{-1}E(\mathbf{u}^s_{t+s}\mathbf{u}^{s\prime}_{t+s})D'^{-1} \qquad s = 0, 1, 2, ..., h \tag{5}$$

from which the traditional variance decompositions can be calculated by plugging in the usual sample-based equivalents. For comparison, in traditional VARs the unnormalized MSE is

$$MSE(E(\mathbf{y}_{t+s}|X_{t-1})) = E(\mathbf{u}^0_t\mathbf{u}^{0\prime}_t) + \Psi_1 E(\mathbf{u}^0_t\mathbf{u}^{0\prime}_t)\Psi'_1 + ... + \Psi_s E(\mathbf{u}^0_t\mathbf{u}^{0\prime}_t)\Psi'_s \qquad s = 0, 1, 2, ..., h$$

where the Ψ_i and $E(\mathbf{u}^0_t\mathbf{u}^{0\prime}_t)$ are computed from the moving-average representation and the residual variance-covariance matrix of the estimated VAR. The quality of the variance decompositions will therefore depend on how well the Ψ_i are approximated by the VAR, and therefore depend more heavily on having the correct specification of the DGP, unlike expression (5).

2.2 Inference: Relation to VARs

A VAR specifies that the $n \times 1$ vector $\mathbf{y}_t$ depends linearly on $X_{t-1} \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-p})'$, through the expression

$$\mathbf{y}_t = \boldsymbol{\mu} + \Pi' X_{t-1} + \mathbf{v}_t \tag{6}$$

where $\mathbf{v}_t$ is an $i.i.d.$ vector of disturbances and $\Pi' \equiv [\ \Pi_1 \quad \Pi_2 \quad ... \quad \Pi_p]$. The VAR(1) companion form to this VAR can be expressed by defining[3]

$$W_t \equiv \begin{bmatrix} \mathbf{y}_t - \boldsymbol{\mu} \\ \mathbf{y}_{t-1} - \boldsymbol{\mu} \\ \vdots \\ \mathbf{y}_{t-p+1} - \boldsymbol{\mu} \end{bmatrix}; F \equiv \begin{bmatrix} \Pi_1 & \Pi_2 & ... & \Pi_{p-1} & \Pi_p \\ I & 0 & ... & 0 & 0 \\ 0 & I & ... & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & I & 0 \end{bmatrix}; \boldsymbol{\nu}_t \equiv \begin{bmatrix} \mathbf{v}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{7}$$

and then realizing that according to (6) and (7),

$$W_t = FW_{t-1} + \boldsymbol{\nu}_t \tag{8}$$

from which s-step ahead forecasts can be easily computed since

$$W_{t+s} = \boldsymbol{\nu}_{t+s} + F\boldsymbol{\nu}_{t+s-1} + ... + F^s\boldsymbol{\nu}_t + F^{s+1}W_{t-1}$$

and therefore

$$\begin{aligned} \mathbf{y}_{t+s} - \boldsymbol{\mu} = \mathbf{v}_{t+s} + F_1^1\mathbf{v}_{t+s-1} + ... + F_1^s\mathbf{v}_t + \\ F_1^{s+1}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + ... + F_p^{s+1}(\mathbf{y}_{t-p} - \boldsymbol{\mu}) \end{aligned} \tag{9}$$

where F_i^s is the ith upper, $n \times n$ block of the matrix F^s (i.e., F raised to the power s).

Assuming W_t is covariance-stationary (or in other words, that the eigenvalues of F lie inside the unit circle) the infinite vector moving-average representation of the original VAR in expression (6) is

$$\mathbf{y}_t = \boldsymbol{\gamma} + \mathbf{v}_t + F_1^1\mathbf{v}_{t-1} + F_1^2\mathbf{v}_{t-2} + ... + F_1^s\mathbf{v}_{t-s} + ... \tag{10}$$

[3] For a more detailed derivation of some of the expressions that follow the reader should consult Hamilton (1994), chapter 10.

and the impulse response function is given by

$$\left.\frac{\Delta \mathbf{y}_{t+s}}{\Delta \boldsymbol{\delta}_t}\right|_{\boldsymbol{\delta}_t=\mathbf{d}_i} = F_1^s \mathbf{d}_i$$

In practice, the coefficients of the impulse response function can be calculated with estimates of the VAR coefficients Π_i $i = 1, ..., p$ and the following recursion (see Hamilton, 1994)

$$\begin{aligned} F_1^1 &= \Pi_1 \\ F_1^2 &= \Pi_1 F_1^1 + \Pi_2 \\ &\vdots \\ F_1^s &= \Pi_1 F_1^{s-1} + \Pi_2 F_1^{s-2} + ... + \Pi_p F_1^{s-p} \end{aligned} \tag{11}$$

Expressions (9) and (11) are useful in establishing the relationship between VARs and local projections. Specifically, comparing expression (2), repeated here for convenience,

$$\mathbf{y}_{t+s} = \boldsymbol{\alpha}^s + B_1^{s+1}\mathbf{y}_{t-1} + B_2^{s+1}\mathbf{y}_{t-2} + ... + B_p^{s+1}\mathbf{y}_{t-p} + \mathbf{u}_{t+s}^s \qquad s = 0, 1, 2, ..., h \tag{12}$$

with expression (9) rearranged,

$$\mathbf{y}_{t+s} = (I - F_1^s - ... - F_p^s)\mu + F_1^{s+1}\mathbf{y}_{t-1} + ... + F_p^{s+1}\mathbf{y}_{t-p} + (\mathbf{v}_{t+s} + F_1^1\mathbf{v}_{t+s-1} + ... + F_1^s\mathbf{v}_t) \tag{13}$$

it is obvious that,

$$\begin{aligned} \boldsymbol{\alpha}^s &= (I - F_1^s - ... - F_p^s)\mu \\ B_1^{s+1} &= F_1^{s+1} \\ \mathbf{u}_{t+s}^s &= (\mathbf{v}_{t+s} + F_1^1\mathbf{v}_{t+s-1} + ... + F_1^s\mathbf{v}_t) \end{aligned} \tag{14}$$

Therefore, when the DGP for $\mathbf{y}_t$ is the VAR in expression (6), the local projections in expression (2) are equivalent to estimating the coefficients of the impulse response given by the sequence of regressions (13). The error terms $\mathbf{u}_{t+s}^s$will have a moving average form given by expression (14) involving the lags of the intervening residuals $\mathbf{v}_{t+s}$ up to time t, but which are otherwise uncorrelated with the regressors since these are dated $t-1, ..., t-p$. Proceeding with this comparison and momentarily ignoring the recursions in (11), consider calculating the impulse response coefficients from the VAR by estimating the following system instead. Let $Y_t \equiv (\mathbf{y}_{t+1}, ..., \mathbf{y}_{t+h})$, $V_t \equiv (\mathbf{v}_{t+1}, ..., \mathbf{v}_{t+h})$, and $X_t \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-p})$, then stack the VAR-implied expressions (9) to form the stacked-system

$$Y_t = X_t \Psi + V_t \Phi \tag{15}$$

where (ignoring the constant terms)

$$\Psi = \begin{bmatrix} F_1^1 & F_1^2 & ... & F_1^h \\ F_2^1 & F_2^2 & ... & F_2^h \\ \vdots & \vdots & ... & \vdots \\ F_p^1 & F_p^2 & ... & F_p^h \end{bmatrix}; \Phi = \begin{bmatrix} I_n & F_1^1 & ... & F_1^h \\ 0 & I_n & ... & F_1^{h-1} \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & I_n \end{bmatrix}$$

and given that $E(\mathbf{v}_t \mathbf{v}_t') = \Omega_v$, then $E(V_t V_t') = \Phi \left(I_h \otimes \Omega_v\right) \Phi' \equiv \Sigma$.

Maximum likelihood estimation of the system implied by the VAR expressions (9) in expression (15) could then be accomplished by standard GLS formulas according to,

$$vec(\widehat{\Psi}) = \left[(I \otimes X)' \Sigma^{-1} (I \otimes X)\right]^{-1} (I \otimes X)' \Sigma^{-1} vec(Y) \tag{16}$$

The usual impulse responses would then be given by rows 1 through n and columns 1 through (nh) of $\widehat{\Psi}$ and standard errors could be computed directly from the regression output rather than from delta method approximations or simulation methods based on Monte Carlo or bootstrap

replication. Further simplification would be achieved due to the special structure of the variance-covariance matrix Σ, which allows GLS estimation of the system block by block.

This disquisition not only illustrates a new method for computing standard errors from VARs (which while nice, is subsidiary to the main message of the paper) but it also shows that when the DGP is given by a VAR and the lag structure is properly specified, local projections give estimates of the impulse responses equivalent to those in the VAR. However, in general the true DGP is unknown so the specific structure of Φ will be unknown as well and we cannot use the GLS estimation strategy in expression (16). This poses no difficulty, however. The structure of Φ suggests that the error terms u_{t+s}^{s} in the expression for the local projections (2) will in general have some form of moving-average structure, whose order is a function of s, the horizon.

Therefore, a recommended strategy is to estimate linear projections by simple linear regression methods and to use heteroskedasticity and autocorrelation (HAC) robust standard errors. Thus, denoting by $\widehat{\Sigma}_L$ the estimated HAC, variance-covariance matrix of the coefficients $\widehat{B}_1^s$ in expression (2), a 95% confidence interval for each element of the impulse response at time s can be constructed approximately as $1.96 \pm \left(\mathbf{d}_i'\widehat{\Sigma}_L\mathbf{d}_i\right)$. Monte Carlo experiments in section 4 suggest, that even when the true underlying model is a VAR, there is virtually no loss in efficiency in proceeding this way. A final note is in order with regard to the practicality of the joint estimation implied in expression (16): the dimension of the system rapidly increases with the number of variables, lags, and horizons for which the impulse response is calculated. The practical implication for the counterpart linear projections based on expression (2) is that, unless the objective is to do cross-impulse response joint hypothesis tests (this point is discussed in more detailed in the next subsection), it is computationally more convenient to do block-by-block joint estimation only to the extent that the variance-covariance matrix of the $\widehat{B}_1^s$ is necessary for formal joint hypothesis tests. To highlight that the efficiency losses of single equation estimation are minor relative to joint estimation, the Monte Carlo experiments and empirical application proceed with single equation estimates.

2.3 Discussion

It is not difficult to grasp that impulse responses calculated by local projections are more robust to misspecification than VAR-based estimates: impulse responses characterize the slope or correlation between $y_{j,t+s}$ and $y_{i,t-1}$, conditional on the past and on the normalization of the marginal experiment – the "shock." While local projections estimate this sample moment directly from the data, VARs approximate it indirectly from their fit of the conditional model of y_t on its past. As a simple example, suppose the DGP is a VAR(2) incorrectly specified as a VAR(1), then according to the recursions in (11),

Impulse	VAR(1)	VAR(2)
F_1^1	$\widetilde{\Pi}_1$	Π_1
F_1^2	$\widetilde{\Pi}_1^2$	$\Pi_1^2 + \Pi_1\Pi_2$
F_1^3	$\widetilde{\Pi}_1^3$	$\Pi_1^3 + 2\Pi_1\Pi_2$
$\vdots$	$\vdots$	$\vdots$

(17)

where $\widetilde{\Pi}_1 = \Pi_1 + \Pi_2\Gamma_1\Gamma_0^{-1}$ and Γ_j is the jth autocovariance of y_t. Thus expression (17) demonstrates that a misspecified VAR produces biased estimates of the impulse response, the severity of which will naturally depend on the omitted terms (in this case Π_2) and on the persistence of the system (if the system is stationary, as $s \to \infty$, the impulse responses converge to zero so that the biases disappear in the long-run).

VARs mask another important problem affecting inference and which is highlighted in Sims and Zha (1999). Traditional, two standard-error bands for impulse responses reported in numerous empirical studies provide proper inference for point estimates of the impulse response's individual coefficients but are otherwise inappropriate for any type of joint hypothesis test. Because impulse responses are nonlinear functions of estimated coefficients (see expression (11)), it is difficult and cumbersome to calculate the variance-covariance matrix of the impulse response coefficients that would be necessary for such tests.

By contrast, the coefficients of impulse responses estimated by local projections are simply the coefficients in a standard regression and their variance-covariance matrix can be estimated as usual (provided a HAC robust estimator is used). Therefore, formal joint inference of coefficients, tests of coefficient restrictions and even tests of restrictions across impulse responses for different variables or shocks, is straight-forward. This is a significant advantage. However, in econometrics flexibility always comes at the price of efficiency and it is no different here, yet Monte Carlo evidence in the section 4 suggests efficiency losses are rather small. Furthermore, while VAR-based forecasts account for two sources of uncertainty, namely parameter estimation uncertainty and uncertainty about the shocks that will intervene in each period; local projections add another natural source: model specification uncertainty.

2.4 Comparison with other Impulse Response Estimators

A number of recent papers examine ways of estimating impulse responses alternative to VARs and it is worth comparing them to local projection methods. I consider three papers by Chang and Sakata (2002), Cochrane and Piazzesi (2002), and Thapar (2002). A common feature of these methods is that they proceed in two stages: in the first stage a forecast-error series, $\widehat{v}_t$, is created, which is then used in a second stage regression involving the original data y_t (for simplicity and without loss of generality, the ensuing discussion is in the univariate context, hence the lower case notation). Thus, in the first stage Chang and Sakata (2002) use an autoregression, Cochrane and Piazzesi (2002) forecast errors implied by financial prices, and Thapar (2002) errors in surveys of forecasts. The second stage regressions are respectively (with constants omitted for simplicity):

Chang and Sakata

$$y_{t+s} = \alpha_s \widehat{v}_t + \varepsilon_{t+s} \tag{CS}$$

Cochrane and Piazzesi

$$y_{t+s} - y_{t-1} = \alpha_s \widehat{v}_t + \varepsilon_{t+s} \tag{CP}$$

Thapar

$$y_{t+s} - E_t y_{t+s} = \alpha_s \widehat{v}_t + \varepsilon_{t+s} \tag{T}$$

for $s = 1, 2, ...$ These methods are reminiscent of the proposals in Barro (1977, 1978), whose second stage regression is instead,

$$y_t = \alpha_1 \widehat{v}_{t-1} + ... + \alpha_p \widehat{v}_{t-p} + \varepsilon_t$$

and therefore can be seen as a truncated but direct estimate of the infinite moving average representation of y_t. The appendix shows that except for Thapar's (2002) and Barro's (1977, 1978) proposals, the residuals of the second stage regression have moving average terms involving information dated $t-1$, $t-2, ...$ (in addition to moving average terms with information dated $t+s, ..., t+1$, which also appear in the local projection method). This observation and the fact that regressions with generated regressors make it difficult to incorporate the estimation uncertainty of the first stage, cause these authors to recommend bootstrap methods to compute appropriate standard errors.

The three methods just reviewed share in common the view that the error series $\widehat{v}_t$ is "fundamental" in some sense and for Cochrane and Piazzesi (2002) and Thapar (2002), this becomes a major selling point: because the forecast-errors are constructed from market-based (rather than econometrically-based) expectations, all available information is appropriately incorporated and in addition one can circumvent the issue of identification altogether. However, it is perilous to disassociate the series of "shocks" from the underlying mechanism that generated them, specially in a multivariate context. The Wold decomposition theorem (see Brockwell and Davis, 1991) ensures that any covariance-stationary process can be expressed as an infinite moving average of the

forecast errors (i.e., the impulse response form for $\mathbf{y}_t$) that are optimal in the mean-square sense. It does not guarantee however, that these "shocks" are structural in the sense of representing the residual series that describes the DGP. This correspondence holds true only if the DGP is linear and the linear forecasts that generate the forecast errors come from a correctly specified model.

The impulse response characterizes the partial derivatives that spell out the relative trade-offs between different elements in $\mathbf{y}_t$ over time in the multi-dimensional function that relates $\mathbf{y}_t$ to its past. Thus, while small variations in the specification of this multi-dimensional function may do little to alter the "slopes" that measure such trade-offs, they may well generate residual series that are relatively uncorrelated with each other. A similar point was raised by Sims (1998) in response to a paper by Rudebusch (1998).

This argument can be underscored by an additional observation, that while it is perfectly coherent to think of impulse responses in the context of a non-linear, non-Gaussian model for $\mathbf{y}_t$ (such as when the data are transition data[4]), there may not always be a natural series of "shocks" that can be manufactured for such a model. On the other hand, it is not conceptually difficult to see that one could obtain the impulse responses by computing the sequence of first-order marginal effects in models that seek to explain $\mathbf{y}_{t+s}$ as a function of information dated $t-1$, and beyond, just as the local projection does in expression (2).

3 Flexible Local Projections

Thus far the main apparent advantages of using local projections to estimate impulse responses appears circumscribed to robustness to misspecification of the lag-length and ease of computation of standard errors for joint inference (important attributes in their own right). However, because these projections are linear, they still restrict impulse responses to be symmetric, shape-invariant, and history independent. This section proposes generalizations of the local projection method that can account for these properties while still preserving the simplicity in the estimation and

[4] See Lancaster (1990).

the ability to compute appropriate standard errors.

In a traditional VAR, investigation of nonlinearities is limited by at least three considerations: (1) the ability to jointly estimate a nonlinear system of equations; (2) the difficulty in generating multiple-step ahead forecasts from a multivariate non-linear model (which, at a minimum, requires simulation methods since there are no closed forms available); and (3) the complication in computing appropriate standard errors for multiple step-ahead forecasts, and thus the impulse responses. However, with local projections the capacity to estimate impulse response coefficients directly from univariate regression output (such as is done in expression 4), basically eliminates these three drawbacks. Furthermore, since the impulse response coefficients are associated with the regressors $\mathbf{y}_{t-1}$ in expression (2), exploration of nonlinearities can be made parsimonious by concentrating on these terms alone.

A non-linear time series process $\mathbf{y}_t$ can be expressed, under mild assumptions, as a generic function of past values of a white noise process $\mathbf{v}_t$ in the form

$$\mathbf{y}_t = \Phi\left(\mathbf{v}_t, \mathbf{v}_{t-1}, \mathbf{v}_{t-2}, ...\right)$$

Assuming $\Phi(.)$ is sufficiently well behaved so that it can be expanded in a Taylor series expansion around some fixed point, say $\mathbf{0} = (0, 0, 0, ...)$, then the closest equivalent to the Wold representation in nonlinear time series is the Volterra series expansion (see Priestley, 1988),

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \Phi_i \mathbf{v}_{t-i} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \Phi_{ij} \mathbf{v}_{t-i}\mathbf{v}_{t-j} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \Phi_{ijk} \mathbf{v}_{t-i}\mathbf{v}_{t-j}\mathbf{v}_{t-k} + ... \tag{18}$$

which is a polynomial extension of the Wold decomposition in expression (10) with the constant omitted for simplicity. Therefore, it is natural to extend the local projections in expression (2) with polynomial terms that can approximate a wide class of smooth nonlinear functions in a similar way. For simplicity and as an example, consider including up to cubic terms as follows,

$$
\begin{aligned}
\mathbf{y}_{t+s} &= \boldsymbol{\alpha}^s + B_1^{s+1}\mathbf{y}_{t-1} + Q_1^{s+1}{\mathbf{y}_{t-1}}^2 + C_1^{s+1}{\mathbf{y}_{t-1}}^3 + \\
&\quad B_2^{s+1}\mathbf{y}_{t-2} + ... + B_p^{s+1}\mathbf{y}_{t-p} + \mathbf{u}_{t+s}^s \qquad s = 0, 1, 2, ..., h
\end{aligned} \tag{19}
$$

where I do not allow for cross-product terms so that ${\mathbf{y}_{t-1}}^2 = (y_{1,t-1}^2, y_{2,t-1}^2, ..., y_{n,t-1}^2)'$, as a matter of choice and parsimony. It is readily apparent that the impulse response at time s now becomes,

$$
\begin{aligned}
\left.\frac{\partial \widehat{\mathbf{y}_{t+s}}}{\partial \boldsymbol{\delta}_t}\right|_{\boldsymbol{\delta}_t = \mathbf{d}_i} &= \left\{\widehat{B}_1^s(\mathbf{y}_{t-1} + \mathbf{d}_i) + \widehat{Q}_1^s(\mathbf{y}_{t-1} + \mathbf{d}_i)^2 + \widehat{C}_1^s(\mathbf{y}_{t-1} + \mathbf{d}_i)^3\right\} - \\
&\quad \left\{\widehat{B}_1^s\mathbf{y}_{t-1} + \widehat{Q}_1^s(\mathbf{y}_{t-1})^2 + \widehat{C}_1^s(\mathbf{y}_{t-1})^3\right\} \\
&= \left\{\widehat{B}_1^s\mathbf{d}_i + \widehat{Q}_1^s(2\mathbf{y}_{t-1}\mathbf{d}_i + \mathbf{d}_i^2) + \widehat{C}_1^s(3\mathbf{y}_{t-1}^2\mathbf{d}_i + 3\mathbf{y}_{t-1}\mathbf{d}_i^2 + \mathbf{d}_i^3)\right\} \\
s &= 0, 1, 2, ..., h
\end{aligned} \tag{20}
$$

and with the obvious normalizations, $B_1^0 = I$, $Q_1^0 = 0_n$, and $C_1^0 = 0_n$. Several elements of this impulse response deserve comment. First, these nonlinear estimates can be easily calculated by least squares, equation by equation, with any conventional econometric software. Second, if some of the terms Q_i^s and C_i^s are non-zero, the impulse response function will vary according to the sign and with the size of the experimental shock defined by $\mathbf{d}_i$. Third, the impulse response depends on the local history $\mathbf{y}_{t-1}$ at which it is evaluated. In particular, impulse responses comparable to local-linear or VAR-based impulse responses can be achieved by evaluation at the sample mean, i.e. $\mathbf{y}_{t-1} = \overline{\mathbf{y}}_{t-1}$.

From a practical point of view, reporting impulse responses based on non-linear local projection methods requires some additional care: each horizon estimate is no longer a point but rather depends on the choice of $\mathbf{d}_i$ and, in this particular case, $\mathbf{y}_{t-1}$. One option is to commit to choices of $\mathbf{d}_i$ and $\mathbf{y}_{t-1}$ which are deemed relevant for the particular economic experiment of interest. Alternatively, one could consider reporting the expected value of the impulse response at each horizon, conditional on the distribution of $\mathbf{d}_i$ and $\mathbf{y}_{t-1}$. Finally, one could report three-dimensional

plots of the impulse response as a function of $\mathbf{y}_{t-1}$ for a given $\mathbf{d}_i$ of interest, for example. Potter (2000) contains a detailed and more formal discussion of alternative ways of defining the nonlinear impulse response.

Finally, notice that inference is still straight-forward. The 95% confidence interval for the cubic approximation in expression (19) can be calculated by defining the scaling $\boldsymbol{\lambda}_i \equiv (\mathbf{d}_i, \quad 2\mathbf{y}_{t-1}\mathbf{d}_i + \mathbf{d}_i^2, \quad 3\mathbf{y}_{t-1}^2\mathbf{d}_i + 3\mathbf{y}_{t-1}\mathbf{d}_i^2 + \mathbf{d}_i^3)'$, which depends on the local history of when the impulse response is evaluated through the terms in $\mathbf{y}_{t-1}$. Denoting $\widehat{\Sigma}_C$ the HAC, variance-covariance matrix of the coefficients $\widehat{B}_1^s, \widehat{Q}_1^s$, and $\widehat{C}_1^s$ in (19), a 95% confidence interval for the impulse response at time s is approximately, $1.96 \pm \left(\boldsymbol{\lambda}_i'\widehat{\Sigma}_C\boldsymbol{\lambda}_i\right)$.

Flexible local projections, such as the local-cubic projection in (19), offer several interesting possibilities. First, notice that there is no obvious multivariate specification of a primitive model whose implied impulse responses would have the structure given by (20). Second, the impulse responses are no longer symmetric – the quadratic terms are always positive irrespective of the sign of the shock. Third, the responses are no longer shape invariant since the quadratic and cubic terms are not invariant to the size of the shock. Fourth, the responses depend on the local history at which they are evaluated through the terms $\mathbf{y}_{t-1}$. Finally, these gains do not come at the cost of estimating wildly more complicated models (as would be necessary if we wanted to add flexibility to a VAR) – the impulse responses can still be estimated by least squares methods and, its error bands are easily computed.

Natural extensions to this example would consist in formulating a flexible specification for the terms $\mathbf{y}_{t-1}$ in expression (2), that is,

$$\mathbf{y}_{t+s} = m^s(\mathbf{y}_{t-1}; X_{t-2}) + \mathbf{u}_{t+s}^s \qquad s = 0, 1, 2, ..., h$$

where $m^s(.)$ is a flexible form and may include any parametric, semi-parametric and non-parametric approximation, such as Hamilton's (2001) parametric, flexible nonlinear model; flexible discrete-Fourier forms (see Granger and Hatanaka, 1964); artificial neural networks (see White, 1992);

wavelets (see Percival and Walden, 2000); or more generically, non-parametric methods (see Pagan and Ullah, 1999). In addition, since impulse responses can be calculated from univariate model estimates, the universe of regime-switching and non-linear time series models becomes readily available. See Granger and Teräsvirta (1993) for a review but to mention a few, these include Hamilton's (1989) switching-regimes model; Tong's (1983) threshold autoregressions (TAR); and so on. The specific choices will be dictated by the needs of each application, so an extensive review of the attributes of each alternative falls beyond the scope of this paper. Monte Carlo experiments in section 4 show some of the benefits of the local-cubic projection example just discussed, while the application in section 5 shows how to compute impulse responses based on local projections with a threshold model.

4 Monte Carlo Evidence

This section discusses two main simulations that evaluate the performance of local projections for impulse response estimation and inference. The first experiment is based on a conventional VAR that appears in Christiano, Eichenbaum and Evans (1996) and Evans and Marshall (1998), among others. The experiment illustrates that local projections deliver impulse responses that are robust to lag length misspecification, consistent, and only mildly inefficient relative to the responses from the true DGP. The second experiment simulates a SVAR-GARCH (see Jordà and Salyer, 2003) to show that flexible local projections do a reasonable job at approximating the inherent nonlinearities of this model.

4.1 Christiano, Eichenbaum and Evans (1996)

This Monte Carlo simulation is based on monthly data from January 1960 to February 2001 (494 observations). First I estimate a VAR of order 12 on the following variables: EM, log of non-agricultural payroll employment; P, log of personal consumption expenditures deflator (1996 = 100); $PCOM$, annual growth rate of the index of sensitive materials prices issued by the Conference Board; FF, federal funds rate; $NBRX$, ratio of nonborrowed reserves plus extended credit to total

reserves; and $\Delta M2$, annual growth rate of M2 stock. I then save the coefficient estimates from this VAR and simulate 500 series of 494 observations using multivariate normal residuals and the variance-covariance matrix from the estimation stage. To start the simulation, all 500 runs are initialized with the first 12 observations from the data. Information criteria based on the data suggest the lag-length to be twelve if using Akaike's AIC and Hurvich and Tsai's[5] AIC_c, or two if using Schwartz's SIC. These choices are very consistent across the 500 simulated runs.[6]

The first experiment compares the impulse responses that would result from fitting a VAR of order two (as SIC would suggest) with local-linear and -cubic projections of order two as well. Although a reduction from twelve to two lags may appear severe, this is a very mild form misspecification in practice. The results are displayed graphically in figure 1 rather than reporting tables of root mean-squared errors, which are less illuminating. Each panel in figure 1 displays the impulse response of a variable in the VAR due to a shock in the variable FF,[7] calculated as follows: the thick-solid line is the true VAR(12) impulse response with two standard-error bands displayed in thick-dashed lines (these are based on the Monte Carlo simulations of the true model). The responses based on a VAR(2) are displayed by the line with squares; the responses from the local linear approximation are displayed by the dashed line; and the responses from the cubic local approximation are displayed by the line with circles.

Several results deserve comment. The VAR(2) responses often fall within the two standard-error bands of the true response and have the same general shape. This supports the observation that the VAR(2) is only mildly misspecified. However, both the local-linear and -cubic projections are much more accurate at capturing detailed patterns of the true impulse response over time, even at medium- and long-horizons. In one case, the departure from the true impulse response was economically meaningful: the response of the variable P. The response based on the VAR(2)

[5] Hurvich and Tsai (1993) is a correction to AIC specifically designed for VARs.

[6] Although the true DGP contains 12 lags, the coefficients used in the Monte-Carlo are based on the estimated VAR and it is plausible that many of these coefficients are not significantly different from zero in practice.

[7] Responses to shocks in all the variables are available upon request. For the sake of brevity, the other figures are not enclosed in the paper. The omitted figures present results that are similar to the ones reported here.

is statistically different from the true response for the first 17 periods, and suggests that prices *increase* in response to an increase in the federal funds rate over 23 out of the 24 periods displayed. Many researchers have previously encountered this type of counterintuitive result and dubbed it the "price puzzle." Sims (1992) suggested this behavior is probably related to unresolved endogeneity issues and proposes including a materials price index, as it is done here with $PCOM$. In contrast, the local-linear projection is virtually within the true two standard error bands throughout the 24 periods depicted, and is strictly negative for the last 7 periods.

The second experiment shows that local projection methods are consistent under true specification by calculating impulse responses with up to 12 lags. The results are reported in figure 2, also for a shock to FF only. Thus, the thick line is the true impulse response, along with two standard error bands displayed in thick-dashed lines. The responses based on local linear projections are displayed with the dashed line and the responses based on local cubic projections are displayed by the line with circles. Generally speaking, the responses by either approximation literally lie on top of the true response[8] with occasional minor differences that disappeared with slightly bigger samples, not reported here.

The final set of experiments evaluates the standard error estimates of the impulse response coefficients (which are commonly used to display error bands around impulse responses). In order to stack the odds against local projection methods and because in practice we never know the true multivariate DGP describing the data, I consider standard errors calculated from univariate projections, equation by equation. Specifically, I generated 500 runs of the original series and then I fitted a VAR(12) and local-linear and -cubic projections with 12 lags as well. Then I computed Monte Carlo standard errors for the VAR(12) to give a measure of the true standard errors, and then calculated Newey-West[9] corrected standard errors for the local projections. Table 1 reports these results for each variable in response to a shock in FF as well.

[8] This is also true for the responses to all the remaining shocks that are not reported here but are available upon request.

[9] The Newey-West lag correction is selected to be equal to s, the horizon of the impulse response being considered.

In section 2 I argued that local projection estimates of impulse responses are less efficient than VAR-based estimates when the VAR is correctly specified and it is the true model. Table 1 confirms this statement but also shows that this loss of efficiency is not particularly big. The Newey-West corrected standard errors based on single equation estimates of the local linear projections are virtually identical to the Monte Carlo standard errors from the VAR, specially for the variables EM and P. The biggest discrepancy is for the variable $NBRX$ but this is because the VAR Monte Carlo standard errors actually *decline* as the horizon increases (specially after the 14th period). This anomaly, which is explained in Sims and Zha (1999), is not a feature of the local projection standard errors, which incorporate the additional uncertainty existing in long-horizon forecasts. Altogether, these results suggest that the efficiency losses are rather minor, even for a system that contains as many as six variables and 12 lags and for horizons of 24 periods.

4.2 Impulse Responses for a GARCH-SVAR

The following Monte-Carlo experiment gauges how well local projection estimates approximate the impulse responses from a nonlinear DGP relative to VAR-based estimates. In Jordà and Salyer (2003) we propose a multivariate version of the GARCH-M model that we use to determine the effects of monetary policy uncertainty on the term structure of interest rates. We call this model the GARCH-SVAR. Here, I experiment with the following specification,

$$\begin{aligned}
\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} &= A \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{3t-1} \end{bmatrix} + Bh_{1t} + \begin{bmatrix} \sqrt{h_{1t}}\varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix}, \qquad \varepsilon_t \sim N(0, I_3) \qquad (21) \\
h_{1t} &= 0.5 + 0.3u_{1,t-1} + 0.5h_{1,t-1}; \qquad u_{1t} = \sqrt{h_{1t}}\varepsilon_{1t} \\
A &= \begin{bmatrix} 0.5 & -0.25 & 0.25 \\ 0.75 & 0.25 & 0.25 \\ -0.25 & -0.25 & 0.75 \end{bmatrix}; \qquad B = \begin{bmatrix} -1.75 \\ -1.5 \\ 1.75 \end{bmatrix}
\end{aligned}$$

and a sample size of 300, replicated 500 times. Notice that the GARCH-SVAR in (21) behaves like a linear VAR most of the time (in fact, if the shock is to either ε_{2t} or ε_{3t}, it always behaves like a linear VAR). Only when the shock to ε_{1t} is of considerable magnitude there will be a revision in the conditional variance, and subsequently, in the conditional mean. Figure 3 displays the impulse responses from a shock to y_{1t} of unit size. The thick-solid line describes the true impulse response in the GARCH-SVAR. The solid line is the impulse response when the variance effects are set to zero (i.e. $B = 0_3$). The dashed line with stars is the impulse response from the linear projection and the dashed line with circles is the response from the local-cubic projections. Standard-error bands are omitted for clarity but suffice it to say that these are very narrow so that the impulse responses measured from the GARCH-SVAR with and without variance effects clearly remain statistically different from each other, except at crossing points or after the 8th period approximately.

It is important to comment first on the nature of the nonlinearity. When the variance effect is switched off, the impulse responses are more moderate and identical to those in a typical VAR. For example, y_1 responds by gradually returning to zero after the shock, barely crossing into the negative region. In contrast, there is an initial undershooting response of y_1 when the variance effect is allowed to kick-in (with similar under- and overshooting responses in y_2 and y_3), driving y_1 into strongly negative territory after the period of impact before returning to equilibrium after seven periods, approximately.

The first significant result is that the response without variance effects and the response estimated from local-linear projections, are virtually identical. During most of the sample, shocks remain small so there are no revisions in the conditional variance and the model behaves as if it were a typical VAR. Thus, to capture the nonlinearity, we can use the local-cubic projection estimates instead. When the responses estimated with this approximation are evaluated around the sample mean values of $\mathbf{y}_t$, as suggested in section 3, the impulse responses are identical to the responses calculated with a linear projection and therefore, are not displayed in the figure. Thus,

to enhance the nonlinearity and to match the true impulse response with variance effects, we evaluate the local-cubic projection at $\mathbf{y}_{t-1} = \overline{\mathbf{y}}_{t-1} + 5 \times (\widehat{\sigma}_{11}, \widehat{\sigma}_{22}, \widehat{\sigma}_{33})'$. This choice of experiment allows us to match relatively well the more extreme dynamics of the model, as figure 3 shows, and highlights the possibility (not explored here) of using significance tests on the quadratic and cubic terms of the local-cubic projections to test for nonlinearities in the responses implied by the data.

5 Application: Inflation-Output Trade-offs

Pioneering work by McCallum (1983) and Taylor (1993) inspired a remarkable amount of research on the efficacy, optimality, and robustness of interest rate rules for monetary policy. The performance of candidate policy rules is often evaluated in the context of a simple, closed-economy model that, at a minimum, can be summarized by three fundamental expressions: an IS equation, a Phillips relation, and the candidate policy rule itself. While models may differ on their degree of micro-foundation and forward-looking behavior (see Taylor's (1999) edited volume for examples) they share the need to reproduce the fundamental dynamic properties of actual economies with some degree of accuracy.

Consequently, it is natural to investigate these empirical dynamic properties for inflation, the output gap, and interest rates to provide a benchmark by which to compare the dynamic properties of competing theoretical models. The specific definitions of the variables I consider is the following: y_t is the percentage gap between real GDP and potential GDP (as measured by the Congressional Budget Office); π_t is quarterly inflation in the GDP, chain-weighted price index in percent at annual rate; and i_t is the quarterly average of the federal funds rate in percent at an annual rate. These variable definitions are those used for the version of the IS and Phillips relations in Rudebusch and Svensson (1999) and are relatively standard for this literature. The data for the analysis is quarterly for the sample 1955:I - 2003:I, and is displayed in figure 4.

A good starting point for the analysis is to calculate impulse responses with a VAR, and local-linear, and -cubic projections. The lag-length is determined by information criteria, allowing for a

maximum lag-length of eight. Studies with similar variables in Galí (1992) and Fuhrer and Moore (1995a, b) use four lags for variables analyzed in the levels. Such a selection is confirmed by AIC_c and AIC, both of which select a lag-length of three (SIC selected two lags). Figure 5 displays the impulse responses based on a VAR(3), local-linear and -cubic projections and identified with a standard Cholesky decomposition[10] and the Wold-causal order y_t, π_t, and i_t.

The VAR(3) responses are depicted with a dotted line, the short-dashed line and the two long-dashed lines depict the responses from local linear projections and the corresponding two standard-error, Newey-West corrected bands calculated as described in section 2. The solid line is the response from a local cubic projection.[11] Each row represents the responses of y_t, π_t, and i_t to orthogonalized shocks, starting with y_t, π_t, and then i_t, all measured in percentages. Several results stand out. Generally speaking, there is broad correspondence among the responses calculated by the different methods, with a few exceptions. The response of i_t to a shock in y_t calculated by local-cubic projection suggests a more strict (and statistically significant) tightening stance than the other methods, and similarly, the response of the output gap y_t to its own shock is statistically different from the linear projection response (albeit with the same general shape). However, this response corresponds closely to the output responses due to an aggregate supply shock found in Galí (1992), both with an initial increase of about 0.7% and peaking after four quarters at 1.1%.

Perhaps the most meaningful difference is that, while the VAR response of y_t to a shock in i_t suggests that the output loss after 12 quarters is approximately 0.3%, both local projection methods suggest the loss is twice as big, at a statistically (and economically) significant 0.65%. This difference exists despite the similarity among the time profiles for i_t calculated by any of the three methods considered. More generally, the VAR(3) responses have significantly smoother time profiles than responses from local projections. Further investigation revealed that when the

[10] I choose the Cholesly decomposition to identify the structural shocks since I make weak emphasis in the literal interpretation of the impulse responses and it can be easily replicated. However, this choice is consistent with traditional orderings in the VAR literature.

[11] The dot-dashed line is simply the zero line.

maximum possible lag length is increased to 12, AIC will select that length as the new optimum (although AIC_c and SIC remain at their previous levels). The responses from a VAR(12) lie almost on top of their local-projection counterparts, with the few exceptions we have already mentioned.[12] As an aside, this finding and the findings in the Monte Carlo experiments of section 4 suggest that the "price puzzle" is better addressed by specifying relatively long lags in the price equation rather than relying solely on inclusion of the a series for sensitive commodity prices, as is now conventional.

Based on this preliminary analysis, we are positioned to investigate further nonlinearities in the impulse responses. From the vast selection of flexible specifications available, one should select those that, within a general class, will more easily lend themselves to economic interpretation. In this case, it seems of considerable importance to determine whether the inflation-output gap trade-offs that the monetary authority faces vary with the business cycle, or during periods of high inflation, or when interest rates are close to the zero bound, for example. Although the polynomial terms in local projection approximate smooth nonlinearities, they are less helpful in detecting the type of nonlinearity implicit in these examples. Therefore, I tested all the first period local-linear projections[13] for evidence of threshold effects due to y_{t-1}, π_{t-1}, and i_{t-1} using Hansen's (2000) test[14] . For example, a typical regression is,

$$\begin{aligned} z_t &= \boldsymbol{\rho}_L' X_{t-1} + \varepsilon_t^L \qquad \text{if } w_{t-1} \leq \delta \\ z_t &= \boldsymbol{\rho}_H' X_{t-1} + \varepsilon_t^H \qquad \text{if } w_{t-1} > \delta \end{aligned} \tag{22}$$

were z_t is respectively y_t, π_t, and i_t and w_{t-1} can be any of y_{t-1}, π_{t-1}, and i_{t-1}. X_{t-1} collects lags 1 through p of the variables y_t, π_t, and i_t and $\boldsymbol{\rho}_i$, $i = L, H$ collects the coefficients and L stands

[12] The figure displaying these responses is available upon request.

[13] I used the local linear projections for the test for parsimony although the final analysis is based on cubic projections.

[14] The GAUSS routines to perform the test are available directly from Bruce Hansen's web site. I owe a debt of gratitude for having this code publicly available.

for "low" and H stands for "high." The test is an F-type test that sequentially searches for the optimal threshold δ and adjusts the corresponding distribution via 1,000 bootstrap replications.

The tests for the nine possible combinations of dependent variables and threshold variables are summarized in table 2. Only one combination shows a significant departure from the null of linearity: the response of interest rates with a threshold due to y_{t-1}. Figure 6 displays the value of Hansen's test for a range of possible values for the threshold δ. The minimum is achieved for $\delta = -0.0766\%$, and is very close to the canonical value $\delta = 0\%$, which also lies above the 95% critical region. This finding suggests that the responses of interest rates depend on whether the economy is currently above or below potential.

Further investigation revealed that this two-state, interest rate response is significant for the response to an interest rate shock only.[15] Consequently, I investigate for threshold effects in the responses to all three variables in the system due to a shock in i_t, where the threshold effect is determined by lagged deviations of output from potential. Figure 7 displays these responses as follows: the solid line depicts responses calculated by cubic local projection and correspond to those displayed in figure 5. The accompanying long-dashed lines are two standard-error bands, Newey-West corrected and based on the cubic projection as described in section 3. The dotted line shows the response when the output gap is negative, and the green-dashed line when the output gap is positive. I have omitted the responses to shocks in y_t and π_t since these are identical to those in figure 5.

Several results deserve comment. When the economy is below potential, there is essentially no response to the interest rate shock (of size 0.8% on impact) during the first two years and only a slight decline thereafter (up to 0.2% in year three). By contrast, when the economy is above potential, the initial output decline peaks four quarters after impact with a loss of approximately 0.5%, returning to zero at the end of the third year. Part of this behavior is explained by the time profiles of interest rates themselves. In particular, the interest rate response when output is above

[15] The figure showing this result is available upon request.

potential is high (relative to when output is below potential) for the first four quarters but then declines quickly and remains at a zero level for quarters six and beyond. This more aggressive monetary policy stance results in an immediate fall in inflation, dropping by 0.5% in quarter three. However, as interest rates quickly come down to counteract the loss of output, inflation takes off, increasing by 0.5% in quarter seven.

Notice that, when the responses are allowed to vary according to whether output is above or below potential, they often fall outside the two standard error bands estimated for the single regime, local-cubic projection alternative. These differences offer a markedly different picture regarding the costs of raising interest rates in terms of output loss and inflation. They suggest that the output loss of controlling inflation when output is below potential is significantly lower than when output is above potential. It is to be expected that if such considerations where incorporated in the design of an optimal monetary policy response, they would suggest policy rules that differ substantially from the recommendations routinely expressed in the literature. Naturally, such considerations deserve a more detailed investigation than is germane to the focus of the paper and serve to illustrate the potential benefits of flexible local projections in practice.

6 Conclusion

This paper shows how to calculate impulse response functions for a vector time series without estimating a specific dynamic, multivariate model. Instead, I propose estimating the sequence of s least squares regressions,

$$\mathbf{y}_{t+s} = \boldsymbol{\alpha}^s + B_1^{s+1}\mathbf{y}_{t-1} + B_2^{s+1}\mathbf{y}_{t-2} + ... + B_p^{s+1}\mathbf{y}_{t-p} + \mathbf{u}_{t+s}^s \qquad s = 0, 1, 2, ..., h$$

from which the impulse response at time s is given by

$$\left.\frac{\partial \widehat{\mathbf{y}_{t+s}}}{\partial \boldsymbol{\delta}_t}\right|_{\boldsymbol{\delta}_t = \mathbf{d}_i} = \widehat{B}_1^s \mathbf{d}_i \qquad s = 0, 1, 2, ..., h$$

and whose standard error can be calculated as the HAC-robust standard error of the regression coefficient estimates $\widehat{B}_1^s$ with readily available regression routines in most econometrics software packages. These methods provide a natural alternative to estimating impulse response functions based on VARs.

The advantages of estimating impulse responses with these local projections include robustness to misspecification that Monte Carlo evidence shows not come with significant efficiency loses. In fact, because the variance-covariance matrix of the impulse response coefficients coincides with the variance-covariance matrix of regression coefficient estimates, joint hypothesis tests can be performed in a traditional fashion and with little complication. This is a feature seldom explored in the literature despite the warnings in Sims and Zha (1999) and can probably be explained by the inherent enormous computational difficulties of existing methods based on VARs.

Additional improvements in inference can be obtained with local projection methods. As section 2.2 shows, the error terms of the local projections contain moving average terms that are a function of the forecast errors for the periods intervening between $t+s$ and t, which are unobserved in principle. However, the sequential nature of the calculations in (2) provides a natural estimate of this forecast error and suggests that including the error terms $\widehat{u}_{t+s-1}^{s-1}$ as regressors in the local projection (2) at time $t+s$ will improve inference. This idea is similar to that in direct multi-period forecasting (see Bhansali, 2002) where the forecasts $\widehat{\mathbf{y}}_{t+s|t-1}$ are included as regressors in the prediction regressions for $\mathbf{y}_{t+s+1}$. Preliminary Monte Carlo evidence shows remarkable reductions in the impulse response standard errors and thus, the formal derivation of these results is left for a different paper.

The demands of increasingly complex nonlinear economic models whose second (or sometimes higher) order solutions[16] deliver equilibrium conditions in the form of polynomial, stochastic difference equations require impulse response estimators that can accommodate such nonlinearities.

[16] These solutions techniques have been advanced by the pioneering work of Collard and Juillard (2001), Kim et al. (2003), and Schmitt-Grohé and Uribe (2004).

While it is a daunting task to do this for a multivariate model, it can be easily accomplished by local projection methods. The empirical example shows how higher order polynomial terms and threshold effects can be jointly incorporated and appropriate inference reported – features that do not have obvious counterpart multivariate specifications.[17]

There are several useful applications of local projection methods worth remarking. First, local projections can be used to investigate the dynamic features of non-Gaussian models for which a multivariate extension is not readily available. Examples of such models include Engle and Russell's (1998) autoregressive conditional duration model, Hamilton and Jordà's (2002) autoregressive conditional hazard model, and numerous count-data specifications (see Cameron and Trivedi, 1998). Here the approach would consist in estimating a sequence of univariate models where the dependent variable is evaluated at time $t+1, t+2$, ..., $t+s$. Similarly, panel data models offer obvious opportunities for local projections. Relatively short samples in the time dimension and high-dimensionality make multivariate time-series specifications impractical for panel-data. However, local projections can deliver estimates of the dynamic impact of treatment effects in an economical and feasible manner, an issue that is largely ignored in this literature.

7 Appendix

Suppose y_t has the following Wold decomposition

$$y_t = \sum_{i=0}^{\infty} \psi_i v_{t-i}$$

Notice that

$$y_{t+s} = \psi_s v_t + \sum_{i=1}^{s-1} \psi_i v_{t+s-i} + \sum_{i=1}^{\infty} \psi_{s+i} v_{t-i}$$

[17] Tsay (1998) and Krolzig (1997) expand the threshold model and the Markov-switching model to a multivariate context respectively but do not account for polynomial terms nor discuss impulse response calculation and inference.

Comparing this expression with Chang and Sakata's (2002) second stage regression, repeated here for convenience

$$y_{t+s} = \alpha_s \widehat{v}_t + \varepsilon_{t+s} \tag{CS}$$

it is clear that then that $\alpha_s = \psi_s$ and the error term ε_{t+s} collects the moving average terms

$$\varepsilon_{t+s} = \sum_{i=1}^{s-1} \psi_i v_{t+s-i} + \sum_{i=1}^{\infty} \psi_{s+i} v_{t-i}$$

where it seems obvious that the last term in the previous expression could be inverted and one could avoid the autocorrelation of the residuals for information dated $t-1, ...$ by including both lags of y_t and lags of $\widehat{v}_t$ instead. Similarly, one can show that the second stage regression in Cochrane and Piazzesi (2002), reproduced here for convenience

$$y_{t+s} - y_{t-1} = \alpha_s \widehat{v}_t + \varepsilon_{t+s} \tag{CP}$$

can be rewritten as

$$y_{t+s} - y_{t-1} = \psi_s v_t + \sum_{i=1}^{s-1} \psi_i v_{t+s-i} + \sum_{i=1}^{\infty} \left(\psi_{s+i} - \psi_i\right) v_{t-i}$$

where the last term also involves information dated $t-1, ...$ Finally, Thapar's (2002) second stage regression

$$y_{t+s} - E_t y_{t+s} = \alpha_s \widehat{v}_t + \varepsilon_{t+s} \tag{T}$$

can be expressed as

$$y_{t+s} - E_t y_{t+s} = \psi_s v_t + \sum_{i=1}^{s-1} \psi_i v_{t+s-i}$$

and therefore does not contain a moving-average component dated $t-1, ...$

References

Barro, Robert J. (1977) "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review,* March, 67(2), 101-115.

Barro, Robert J. (1978) "Unanticipated Money, Output, and the Price Level in the United States," *Journal of Political Economy*, August, 86(4), 549-580.

Bhansali, R. J. (2002) "Multi-Step Forecasting," in **A Companion to Economic Forecasting.** Michael P. Clements and David F. Hendry, eds. Oxford: Blackwell Publishers.

Brockwell, Peter J. and Richard A. Davis (1991) **Time Series: Theory and Methods.** Springer Series in Statistics, 2nd edition. Heidelberg, New York and Berlin: Springer-Verlag.

Cameron, A. Colin and Pravin Trivedi (1998) **Regression Analysis of Count Data.** Econometric Society Monographs, 30. Cambridge: Cambridge University Press.

Chang, Pao-Li and Shinichi Sakata (2002) "A Misspecification-Robust Impulse Response Estimator," University of Michigan, *mimeo.*

Christiano, Lawrence J., Martin Eichenbaum and Charles L. Evans (1996) "Identification and the Effects of Monetary Policy Shocks," in **Financial Factors in Economic Stabilization and Growth.** Mario I. Blejer, Zvi Eckstein, Zvi Hercowitz, and Leonardo Leiderman (eds.). Cambridge: Cambridge University Press, 36-74.

Clements, Michael P. and David F. Hendry (1998) **Forecasting Economic Time Series**. Cambridge, U.K.: Cambridge University Press.

Cochrane, John H. and Monika Piazzesi (2002) "The Fed and Interest Rates – A High Frequency Identification," *American Economic Review, Papers and Proceedings,* May, 92(2), 90-95.

Collard, Fabrice and Michel Juillard (2001) "A Higher-Order Taylor Expansion Approach to Simulation of Stochastic Forward-Looking Models with an Application to a Nonlinear Phillips Curve Model," *Computational Economics,* 17(2-3), 125-139.

Cox, David R. (1961) "Prediction by Exponentially Weighted Moving Averages and Related Methods," *Journal of the Royal Statistical Society*, Series B, 23, 414-422.

Demiralp, Selva and Kevin D. Hoover (2003) "Searching for the Causal Structure of a Vector Autoregression," U.C. Davis Working Paper 03-03.

Engle, Robert F. and Jeffrey R. Russell (1998) "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, September, 66(5), 1127-1162.

Evans, Charles L. and David A. Marshall (1998) "Monetary Policy and the Term Structure of Nominal Interest Rates: Evidence and Theory," *Carnegie-Rochester Conference Series on Public Policy*, 49(0), 53-111.

Fuhrer, Jeffrey C. and George R. Moore (1995a) "Inflation Persistence," *Quarterly Journal of Economics*, February, 127-159.

Fuhrer, Jeffrey C. and George R. Moore (1995b) "Monetary Policy Trade-offs and the Correlation between Nominal Interest Rates and Real Output," *American Economic Review*, March, 219-239.

Galí, Jordi (1992) "How Well Does the IS-LM Model fit Postwar U.S. Data?" *Quarterly Journal of Economics*, May, 709-738.

Granger, Clive W. J. and Michio Hatanaka (1964) **Spectral Analysis for Economic Time Series.** Princeton, NJ: Princeton University Press.

Granger, Clive W. J. and Norman R. Swanson (1997) "Impulse Response Functions Based on a Causal Approach to Residual Orthogonalization in Vector Autoregressions," *Journal of the American Statistical Association,* March, 92(437), 357-367.

Granger, Clive W. J. and Timo Teräsvirta (1993) **Modelling Nonlinear Economic Relationships.** Oxford: Oxford University Press.

Hamilton, James D. (1989) "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica,* 57, 357-384.

Hamilton, James D. (1994) **Time Series Analysis**. Princeton, New Jersey: Princeton University Press.

Hamilton, James D. (2001) "A Parametric Approach to Flexible Nonlinear Inference," *Econometrica,* 69, 537-573.

Hamilton, James D. and Òscar Jordà (2002) "A Model for the Federal Funds Rate," *Journal of Political Economy,* October, 110(5), 1135-1167.

Hansen, Bruce E. (2000) "Sample Splitting and Threshold Estimation," *Econometrica,* v.68, n. 3, 575-604.

Hurvich, Clifford M. and Chih-Ling Tsai (1993) "A Corrected Akaike Information Criterion for Vector Autoregressive Model Selection," *Journal of Time Series Analysis,* v.14, n. 3, 271-279.

Jordà, Òscar and Kevin D. Salyer (*2003*) "The Response of Term Rates to Monetary Policy Uncertainty," *Review of Economic Dynamics,* October, 6(4), 941-962.

Kim, Jinill, Sunghyun Kim, Ernst Schaumburg, and Christopher A. Sims (2003) "Calculating and Using Second Order Accurate Solutions of Discrete Time Dynamic Equilibrium Models," Princeton University, *mimeo.*

Koop Gary, M. Hashem Pesaran, and Simon M. Potter (1996) "Impulse Response Analysis in Nonlinear Multivariate Models," *Journal of Econometrics,* v. 74, 119-147.

Krolzig, Hans-Martin (1997) **Markov-switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis**. Lecture Notes in Economics and Mathematical Systems, v. 454. Heidelberg and New York: Springer.

Lancaster, Tony (1990) **The Econometric Analysis of Transition Data.** Econometric Society Monographs, 17. Cambridge: Cambridge University Press.

Lin, Jin-Lung and Ruey S. Tsay (1996) "Co-Integration Constraint and Forecasting: An Empirical Examination," *Journal of Applied Econometrics*, v. 11, n. 5, 519-538.

McCallum, Bennett T. (1983) "Robustness Properties of a Rule for Monetary Policy," *Carnegie-Rochester Conference Series on Economic Policy*, 29, 173-203.

Pagan, Adrian and Aman Ullah (1999) **Nonparametric Econometrics.** Cambridge, U.K.: Cambridge University Press.

Percival, Donald B. and Andrew T. Walden (2000) **Wavelet Methods for Time Series Analysis.** Cambridge, U.K.: Cambridge University Press.

Potter, Simon M. (2000) "Nonlinear Impulse Response Functions," *Journal of Economic Dynamics and Control*, September, 24(10), 1425-1446.

Priestley, M. B. (1988) **Non-linear and Non-stationary Time Series Analysis**, London: Academic Press.

Rudebusch, Glenn D. (1998) "Do Measures of Monetary Policy in a VAR Make Sense?" *International Economic Review*, 39(4), 907-931.

Rudebusch, Glenn D. and Lars E. O. Svensson (1999) "Policy Rule for Inflation Targeting," in **Monetary Policy Rules.** John B. Taylor (ed.). NBER Conference Report. Chicago: University of Chicago Press, 203-246.

Schmitt-Grohé, Stephanie and Martin Uribe (2004) "Solving Dynamic General Equilibrium Models Using a Second Order Approximation to the Policy Function," *Journal of Economic Dynamics and Control*, v. 28, 755-775.

Sims, Christopher A. (1980) "Macroeconomics and Reality," *Econometrica*, 48(6), 1-48.

Sims, Christopher A. (1992) "Interpreting the Macroeconomic Time Series Facts: The Effects of Monetary Policy," *European Economic Review*, 36(10), 975-1000.

Sims, Christopher A. (1998) "Do Measures of Monetary Policy in a VAR Make Sense?, A Reply" *International Economic Review*, 39(4), 943-48.

Sims, Christopher A. and Tao Zha (1999) "Error Bands for Impulse Responses," *Econometrica,* v. 67, n. 5, 1113-1156.

Taylor, John B. (1993) "Discretion versus Policy Rules in Practice," *Carnegie-Rochester Conference Series on Public Policy*, 39, 195-214.

Taylor, John B. (1999) **Monetary Policy Rules.** NBER Conference Report. Chicago: University of Chicago Press.

Thapar, Aditi (2002) "Using Private Forecasts to Estimate the Effects of Monetary Policy," New York University, *mimeo.*

Tong, Howell (1983) **Threshold Models in Nonlinear Time Series Analysis**. Lecture Notes in Statistics, 21. Berlin: Springer.

Tsay, Ruey S. (1993) "Comment: Adaptive Forecasting," *Journal of Business and Economic Statistics,* v. 11, n.2, 140-144.

Tsay, Ruey S. (1998) "Testing and Modelling Multivariate Threshold Models," *Journal of the American Statistical Association*, 93(443), 1188-1202.

Weiss, Andrew A. (1991) "Multi-step Estimation and Forecasting in Dynamic Models," *Journal of Econometrics*, April-May, 48(1-2), 135-149.

White, Halbert (ed.) (1992) **Artificial Neural Networks: Approximation and Learning Theory.** Oxford: Basil Blackwell.

Table 1 – Standard Errors for Impulse Responses

	EM			P			PCOM		
s	True-MC	Newey-West (Linear)	Newey-West (Cubic)	True-MC	Newey-West (Linear)	Newey-West (Cubic)	True-MC	Newey-West (Linear)	Newey-West (Cubic)
1	0.000	0.007	0.008	0.000	0.007	0.007	0.000	0.089	0.096
2	0.008	0.011	0.012	0.007	0.010	0.011	0.094	0.146	0.161
3	0.013	0.015	0.016	0.012	0.014	0.015	0.155	0.191	0.212
4	0.018	0.019	0.021	0.015	0.017	0.018	0.202	0.224	0.250
5	0.022	0.023	0.025	0.018	0.020	0.022	0.240	0.255	0.284
6	0.027	0.026	0.030	0.021	0.023	0.025	0.267	0.279	0.311
7	0.031	0.030	0.033	0.025	0.026	0.029	0.296	0.301	0.335
8	0.035	0.033	0.037	0.028	0.029	0.032	0.325	0.322	0.357
9	0.038	0.036	0.040	0.031	0.032	0.035	0.350	0.340	0.376
10	0.041	0.039	0.043	0.035	0.035	0.039	0.361	0.356	0.392
11	0.044	0.042	0.046	0.038	0.038	0.042	0.377	0.371	0.407
12	0.046	0.044	0.048	0.042	0.042	0.045	0.390	0.380	0.416
13	0.048	0.046	0.050	0.046	0.045	0.049	0.402	0.385	0.423
14	0.050	0.048	0.053	0.049	0.048	0.052	0.402	0.389	0.427
15	0.051	0.050	0.055	0.052	0.052	0.056	0.399	0.392	0.430
16	0.053	0.052	0.057	0.055	0.055	0.059	0.393	0.394	0.434
17	0.054	0.054	0.058	0.059	0.058	0.063	0.393	0.396	0.437
18	0.055	0.055	0.060	0.062	0.062	0.066	0.386	0.399	0.441
19	0.057	0.057	0.061	0.066	0.065	0.070	0.381	0.402	0.444
20	0.059	0.058	0.062	0.070	0.068	0.073	0.380	0.405	0.448
21	0.060	0.059	0.064	0.074	0.071	0.076	0.378	0.409	0.453
22	0.061	0.061	0.065	0.078	0.075	0.080	0.377	0.415	0.462
23	0.063	0.062	0.066	0.082	0.078	0.083	0.377	0.423	0.472
24	0.064	0.063	0.068	0.086	0.081	0.086	0.371	0.431	0.484

Notes: True-MC refers to the Monte Carlo (500 replications) standard errors for the impulse response coefficients due to a shock in *FF* in a VAR(12) with the variables *EM*, *P*, *PCOM*, *FF*, *NBRX*, *ΔM2*. Similarly, Newey-West (linear) refers to standard errors calculated from local-linear projections and their Newey-West corrected standard errors, while Newey-West (cubic) refers to the local-cubic projections instead.

Table 1 (contd.) – Standard Errors for Impulse Responses

	FF			NBRX			ΔM2		
s	True-MC	Newey-West (Linear)	Newey-West (Cubic)	True-MC	Newey-West (Linear)	Newey-West (Cubic)	True-MC	Newey-West (Linear)	Newey-West (Cubic)
1	0.000	0.022	0.024	0.0005	0.0005	0.0005	0.014	0.012	0.014
2	0.027	0.036	0.041	0.0007	0.0006	0.0007	0.025	0.023	0.026
3	0.044	0.046	0.052	0.0008	0.0007	0.0008	0.035	0.032	0.035
4	0.054	0.053	0.060	0.0008	0.0008	0.0009	0.044	0.039	0.043
5	0.061	0.058	0.065	0.0009	0.0008	0.0009	0.050	0.045	0.050
6	0.064	0.062	0.069	0.0009	0.0008	0.0009	0.056	0.050	0.056
7	0.067	0.064	0.072	0.0009	0.0008	0.0009	0.061	0.056	0.062
8	0.072	0.066	0.074	0.0009	0.0008	0.0009	0.066	0.060	0.067
9	0.073	0.067	0.075	0.0009	0.0009	0.0010	0.070	0.064	0.072
10	0.074	0.069	0.077	0.0009	0.0009	0.0010	0.074	0.069	0.076
11	0.075	0.072	0.080	0.0009	0.0009	0.0010	0.078	0.073	0.081
12	0.077	0.075	0.083	0.0009	0.0009	0.0010	0.082	0.077	0.085
13	0.079	0.078	0.087	0.0009	0.0009	0.0010	0.084	0.080	0.088
14	0.079	0.080	0.089	0.0009	0.0009	0.0010	0.085	0.082	0.090
15	0.080	0.082	0.090	0.0008	0.0009	0.0010	0.084	0.084	0.092
16	0.080	0.083	0.091	0.0008	0.0009	0.0010	0.085	0.085	0.093
17	0.081	0.084	0.092	0.0008	0.0009	0.0010	0.085	0.086	0.094
18	0.081	0.084	0.093	0.0008	0.0009	0.0010	0.085	0.087	0.095
19	0.079	0.085	0.093	0.0007	0.0009	0.0010	0.084	0.088	0.096
20	0.079	0.086	0.093	0.0007	0.0009	0.0010	0.083	0.088	0.096
21	0.077	0.086	0.094	0.0007	0.0009	0.0010	0.082	0.088	0.096
22	0.077	0.087	0.094	0.0007	0.0009	0.0010	0.081	0.088	0.096
23	0.077	0.087	0.095	0.0006	0.0009	0.0010	0.080	0.088	0.096
24	0.077	0.087	0.095	0.0006	0.0009	0.0010	0.078	0.088	0.096

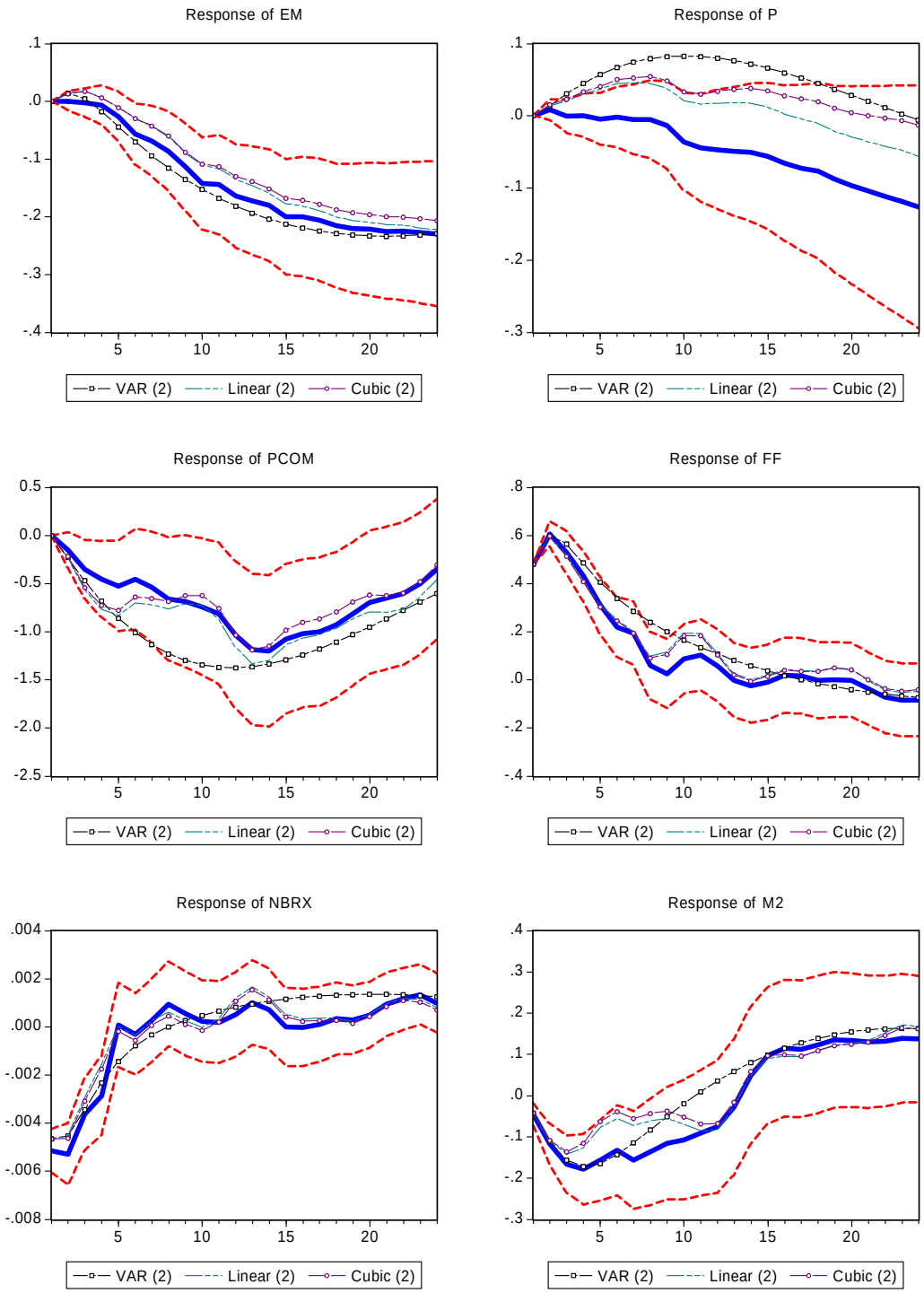
Notes: True-MC refers to the Monte Carlo (500 replications) standard errors for the impulse response coefficients due to a shock in *FF* in a VAR(12) with the variables *EM, P, PCOM, FF, NBRX, ΔM2*. Similarly, Newey-West (linear) refers to standard errors calculated from local-linear projections and their Newey-West corrected standard errors, while Newey-West (cubic) refers to the local-cubic projections instead.

Table 2 – Hansen's (2000) Test for Threshold Effects – p-values

	Dependent Variable		
Threshold Variable	y_t	π_t	i_t
y_{t-1}	0.852	0.850	0.028
π_{t-1}	0.954	0.964	0.738
i_{t-1}	0.335	0.349	0.264

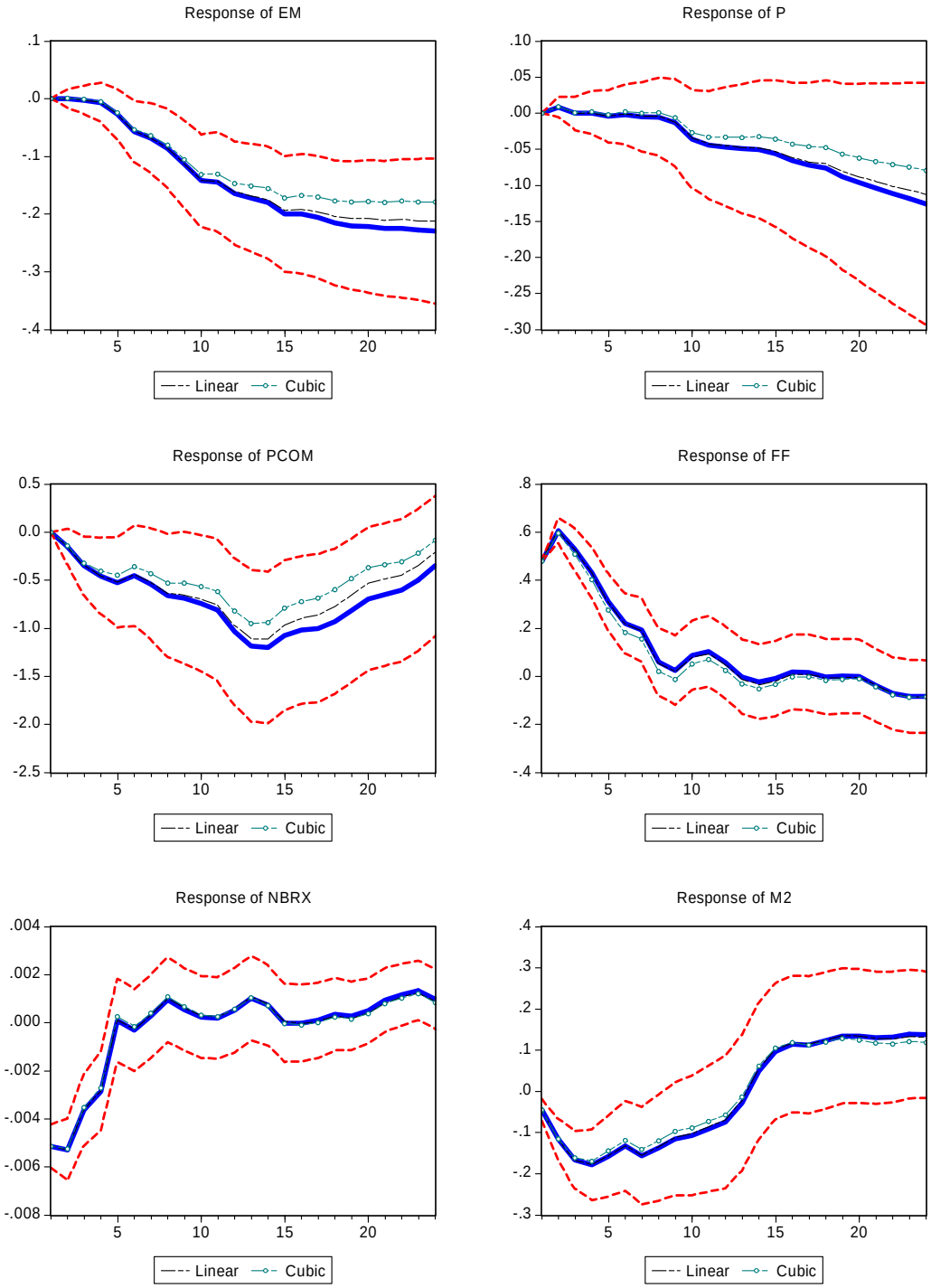
Notes: The test is of the null of linearity against the alternative of threshold effects. The values reported are p-values of the F-type test calculated from 1,000 bootstrap replications.

Figure 1 – Impulse Responses to a Shock in *FF*. Lag Length: 2



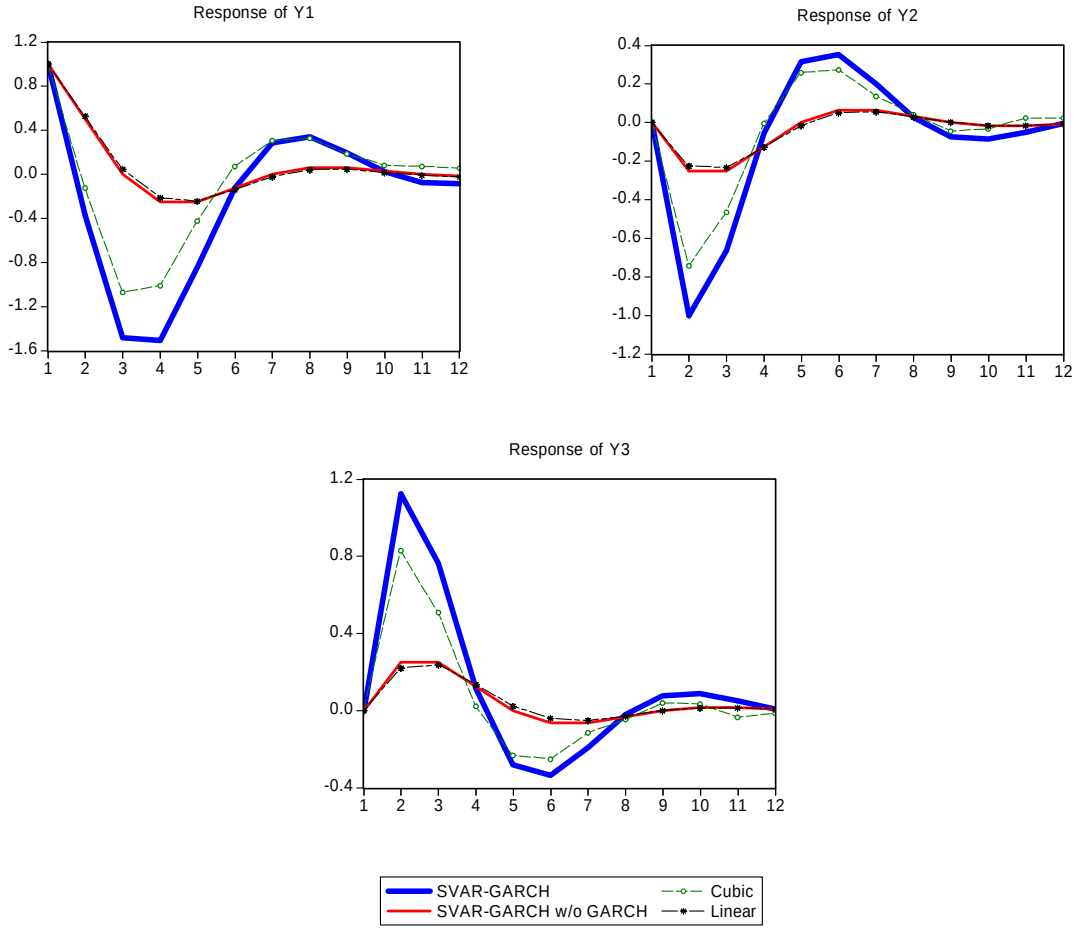
Evans and Marshall (1998) VAR(12) Monte Carlo Experiment. The thick line is the true impulse response based on a VAR(12). The thick-dashed lines are Monte Carlo 2-standard error bands. Three additional impulse responses are compared, based on estimates involving two lags only: (1) the response calculated by fitting a VAR(2) instead, depicted by the line with squares; (2) the response calculated with a local-linear projection, depicted by the dashed line; and (3) the response calculated with a local-cubic projection, depicted by the line with circles. 500 replications.

Figure 2 – Impulse Responses to a Shock in FF. Lag Length: 12



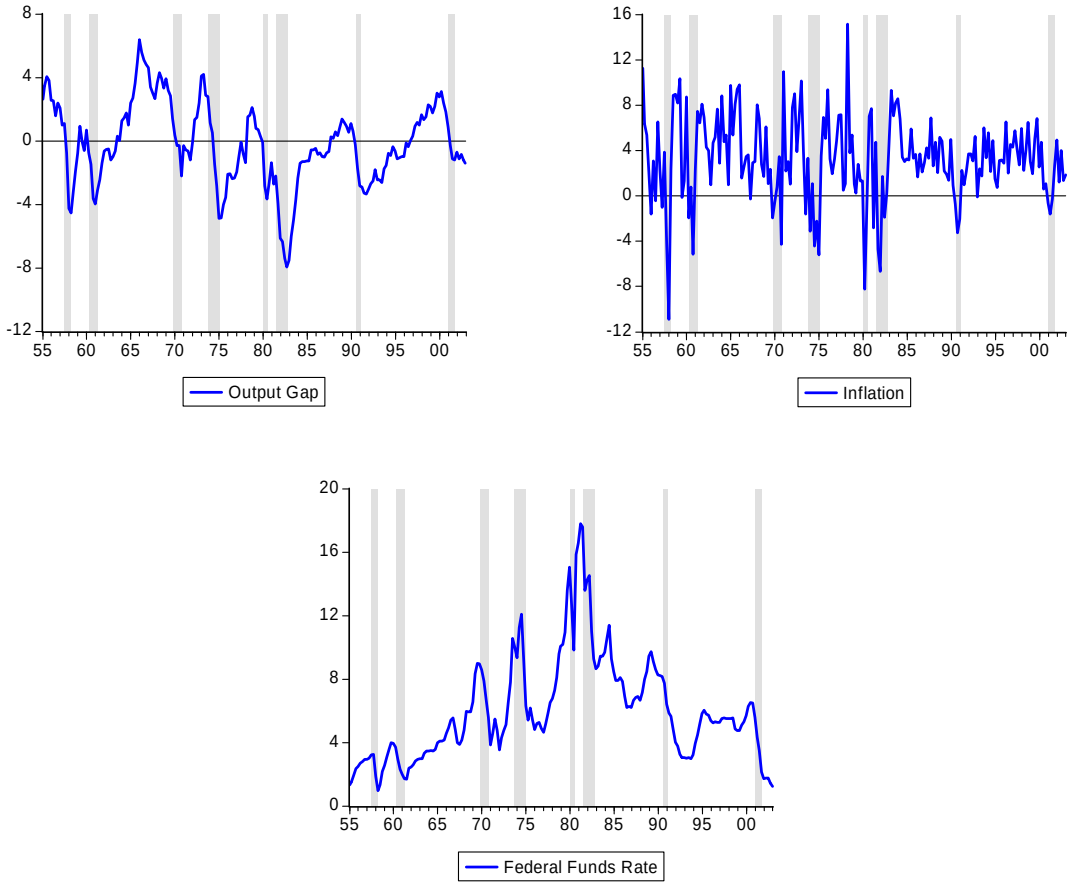
Evans and Marshall (1998) VAR(12) Monte Carlo Experiment. The thick line is the true impulse response based on a VAR(12). The thick-dashed lines are Monte Carlo, 2-standard error bands. Two additional impulse responses are compared: (1) the response calculated with a local-linear projection with 12 lags, depicted by the dashed line; and (3) the response calculated with a local-cubic projection, depicted by line with circles. 500 replications.

Figure 3 – Impulse Responses to a Shock in Y1 from a SVAR-GARCH



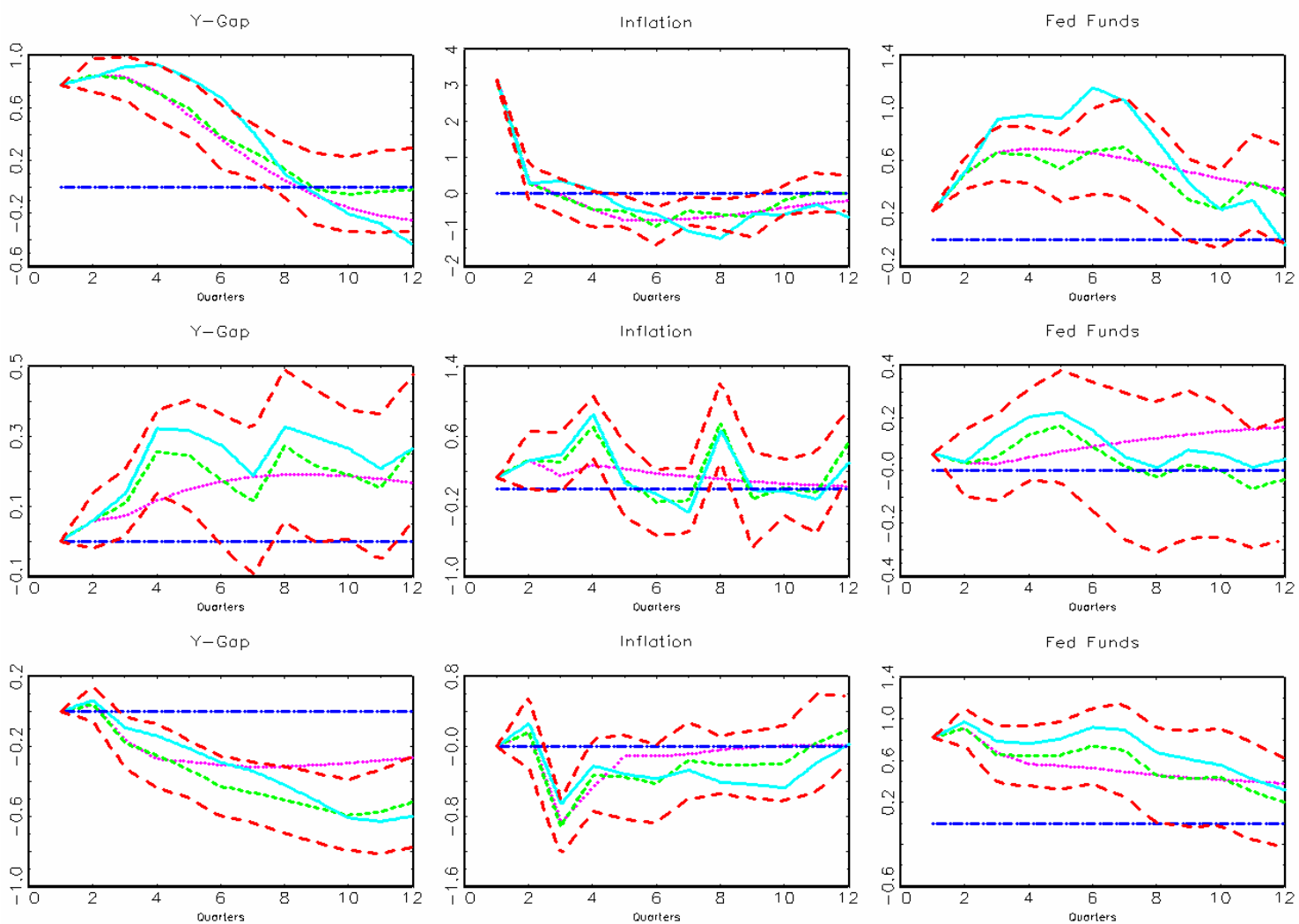
The thick-solid line describes the true impulse response in the VAR-GARCH model. The solid line is the impulse response when the variance effects are set to zero (i.e. $B = 0_3$). The dashed line with stars is the local-linear projection to the impulse response. The dashed line with squares is the local-cubic projection to the impulse response.

Figure 4 – Time Series Plots of the Output Gap, Inflation, and the Federal Funds Rate



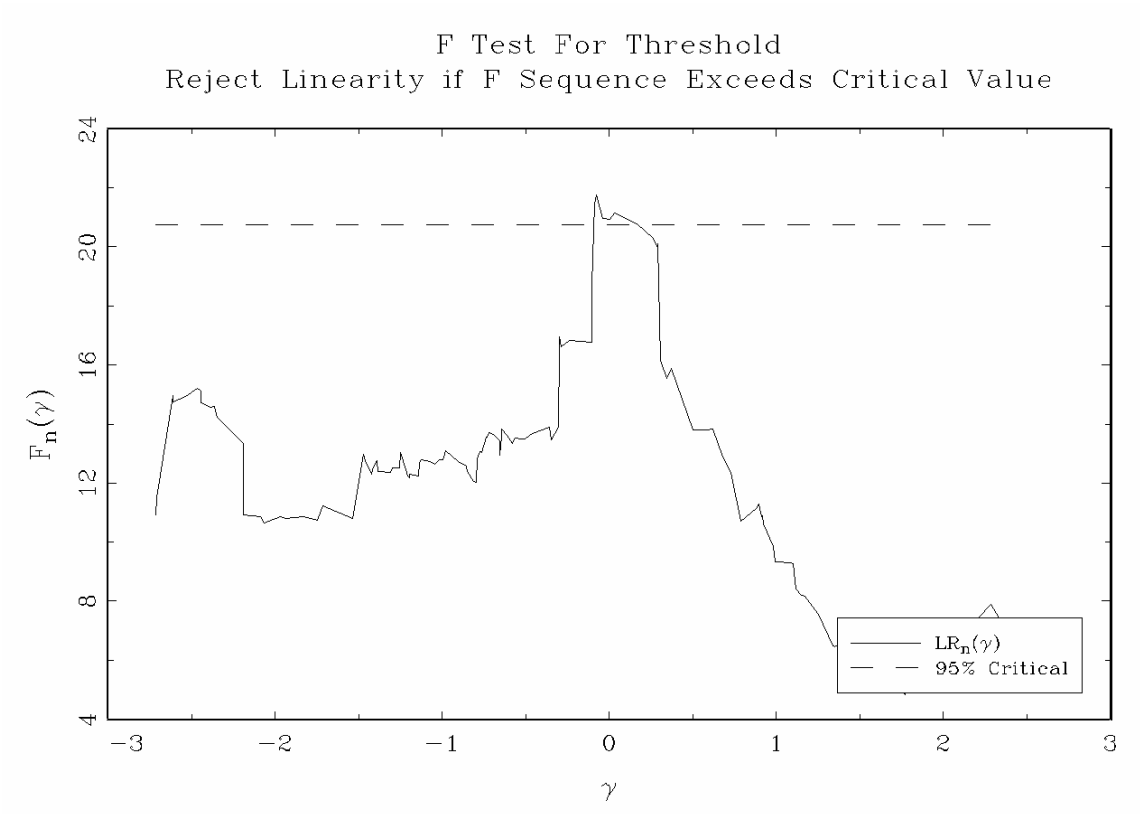
Notes: All variables in annual percentage rates. Shaded areas indicate NBER-dated recessions. Output gap is defined as the percentage difference between real GDP and potential GDP (Congressional Budget Office); Inflation is defined as the percentage change in the GDP, chain-weighted price index at annual rate; and the federal funds rate is the quarterly average of daily rates, in annual percentage rate.

Figure 5 – Impulse Responses Calculated from: a VAR, a Local-Linear and a Local-Cubic Projections



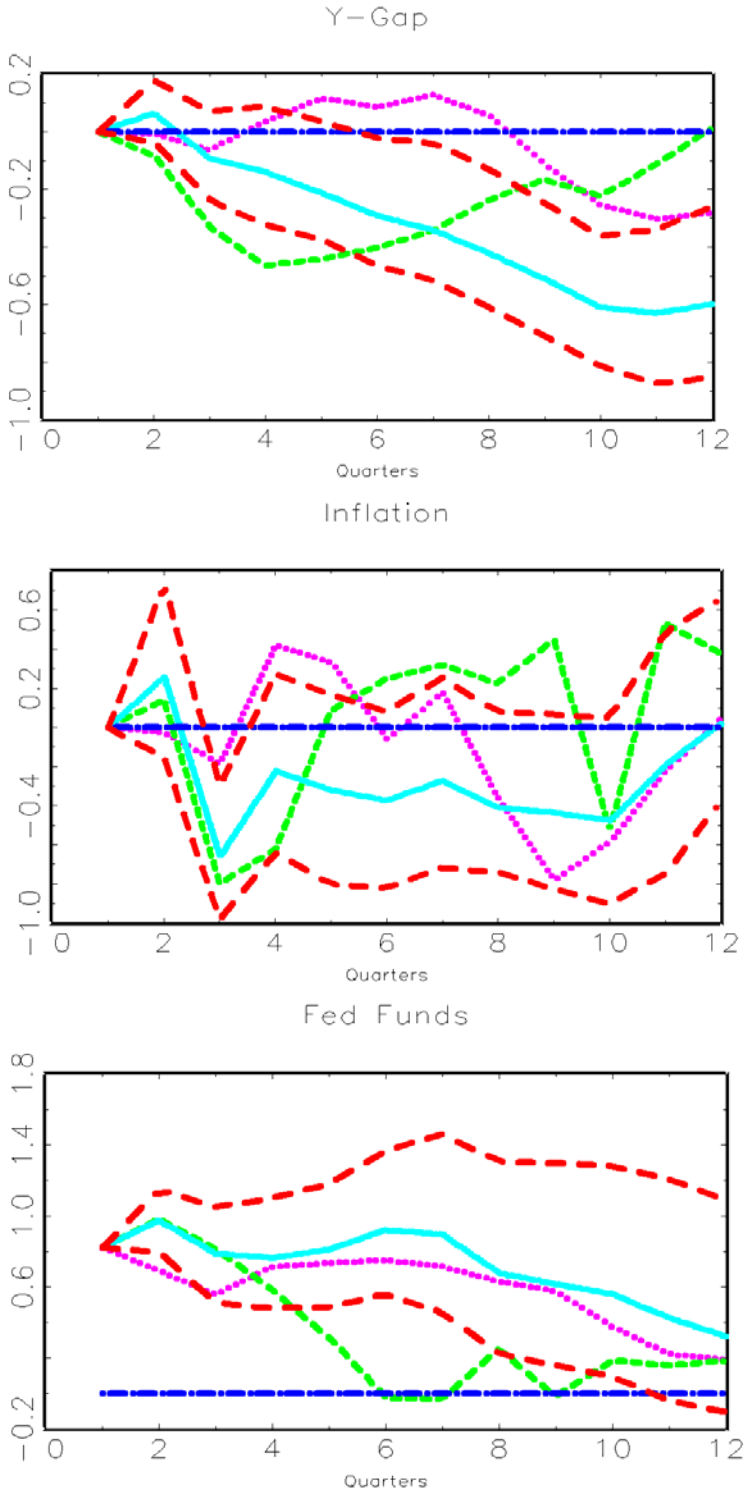
Notes: the dotted line is the VAR(3) response, the short-dashed line is the IRF based on local linear projection, the long-dashed lines are the corresponding Newey-West corrected 2 S.E. bands for the linear projection. The solid line is the IRF based on cubic projection. The dark dot-dashed line is the zero line. All responses in percentages.

Figure 6 - Sequential Test for a Threshold in y_{t-1} for the i_t Equation



Notes: Test of the null hypothesis of linearity against the alternative of a threshold. The sequential test displayed is based on Hansen (2000) and is obtained from GAUSS code available from his website. The threshold is estimated at -0.0765%. The output gap has a mean of -0.189% and a standard error of 2.584%. The p-value of the test is 0.028.

Figure 7 – Impulse Responses with Threshold Effects. Shock to i_t



Notes: the solid line is the IRF from local cubic projection, the long-dashed lines are 2 S.E. bands. The dotted line is the IRF when output gap is negative, the small-dashed line is IRF when output is positive. The dot-dashed line is the zero line. All responses in percentages.