This version: **August, 2004**
Original paper titled "Model-Free Impulse Responses" August, 2003

Estimation and Inference of Impulse Responses by Local Projections*

Abstract

This paper introduces methods to compute impulse responses without specification and estimation of the underlying multivariate dynamic system. The central idea consists in estimating local projections at each period of interest rather than extrapolating into increasingly distant horizons from a given model, as it is done with 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) joint or point-wise analytic inference is simple; 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. Monte Carlo evidence and an application to a simple, closed-economy, new-Keynesian model clarify these numerous advantages.

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

Òscar Jordà
Department of Economics
U.C. Davis
One Shields Ave.
Davis, CA 95616-8578
e-mail: ojorda@ucdavis.edu
URL: www.econ.ucdavis.edu/faculty/jorda

*This version of the paper has benefited from the suggestions of three anonymous referees and the editor. I thank Paolo Angelini, Colin Cameron, John Cochrane, Timothy Cogley, James Hamilton, Peter Hansen, Kevin Hoover, Monika Piazzesi, Simon Potter, Peter Robinson, Shinichi Sakata, Aaron Smith, Daniel Thornton, and seminar participants at the Federal Reserve Bank of Kansas City, Indiana University, U.C. Davis, and U.C. Santa Cruz for many useful suggestions. Massimiliano de Santis provided outstanding research assistance.

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, there is no specific reason to expect that the data are generated by a VAR. In fact, even assuming that a macroeconomy's variables are well characterized by a VAR, Zellner and Palm (1974), and Wallis (1977) show the macroeconomy's subset of variables practitioners can analyze at one time will follow a vector autoregressive-moving average (VARMA) model instead. From a different angle, Cooley and Dwyer (1998) show that the dynamics of basic real business cycle models often follow VARMA representations that are incompatible with VARs. These two observations often explain the relatively long lag lengths required in practice to properly calculate impulse responses with a VAR. Additionally, new second and higher order accurate solution techniques for nonlinear dynamic stochastic general equilibrium models (see, e.g. Kim et al., 2003) deliver equilibrium conditions that are polynomial (rather than linear) difference equations. VARs may indeed be a significantly misspecified representation of the data generating process (DGP).

Impulse responses (and variance decompositions) are important statistics in their own right: they provide the empirical regularities that substantiate theoretical models of the economy and are therefore a natural empirical objective. This paper introduces methods for computing impulse responses for a vector time series based on local projections (a term defined precisely in the next section) that do not require specification and estimation of the unknown true multivariate dynamic system itself.

The advantages of local projections are numerous: they can be estimated by simple least

squares; they provide appropriate inference (individual or joint) that does not require of asymptotic delta-method approximations nor of numerical techniques for its calculation; they are robust to misspecification of the DGP; and they easily accommodate experimentation with highly non-linear specifications that are often impractical or infeasible in a multivariate context. Since local projections can be estimated by univariate equation methods, they can be easily calculated with available standard regression packages and thus become a natural alternative to estimating impulse responses from VARs.

The key insight is that estimation of a model based on the sample, such as a VAR, represents a linear global approximation to the DGP ideal and is optimally designed for one-period ahead forecasting. Even when the model is misspecified, it may still produce reasonable one-period ahead forecasts (see Stock and Watson, 1999). However, an impulse response is a function of forecasts at increasingly distant horizons, and therefore misspecification errors are compounded with the forecast horizon. This paper suggests that it is preferable to use a collection of projections local to each forecast horizon instead, thus matching design and evaluation.

Local projections are based on sequential regressions of the endogenous variable shifted several steps ahead and therefore have many points of commonality with direct multi-step forecasting. The ideas behind direct forecasting (sometimes also called adaptive forecasting or dynamic estimation) go back to at least Cox (1961). Weiss (1991) establishes consistency and asymptotic normality of the direct forecasts under general conditions. The accuracy of direct forecasting has been evaluated in several papers. Tsay (1993) and Lin and Tsay (1996) show that direct forecasting performs very well even relative to models where cointegrating restrictions are properly incorporated. Bhansali (1996, 1997) and Ing (2003) show that direct forecasts outperform iterated forecasts for autoregressive models whose lag length is too short – a typical scenario when a VAR is used to approximate a VARMA model, for example. Bhansali (2002) provides a nice review on this recent literature.

Direct forecasting seeks an optimal multi-step forecast whereas the local projections proposed

here seek a consistent estimate of the corresponding impulse response coefficients. Obviously, these objectives are not disjoint in much the same way that they are not when estimating a VAR.

The paper contains ample Monte Carlo evidence illustrating the basic consistency and efficiency properties of local projections under ideal conditions and under several forms of linear and nonlinear misspecification, all relative to fixed parameter VARs and the more recent time-varying Bayesian VARs used in Cogley and Sargent (2001). As an illustration, I estimate impulse responses for a simple new-Keynesian model (see Galí, 1992, Fuhrer and Moore, 1995a, 1995b, and references in Taylor, 1999) based on cubic polynomial projections with threshold effects. The results are supportive of the view that the Fed faced a changing economic environment during the 1970s to mid 1980s (a view supported among others by Cogley and Sargent, 2001) rather than attributing the inflation-unemployment outcomes of the time to bad policy as DeLong (1997) and Romer and Romer (2002) have suggested.

2 Estimation and Inference

2.1 Estimation

Impulse responses are almost universally estimated from the Wold decomposition of a linear multivariate Markov model such as a VAR. However, this two-step procedure consisting on first estimating the model and then inverting its estimates to find the impulse responses, is only justified if the model coincides with the DGP. Furthermore, deriving correct impulse responses from cointegrated VARs can be extremely complicated (see Hansen, 2003). Instead, impulse responses can be defined without reference to the unknown DGP, even when its Wold decomposition does not exist (see Koop et al., 1996; and Potter, 2000). Specifically, an impulse response can be defined as the difference between two forecasts (see Hamilton, 1994; and Koop et al., 1996):

$$IR(t, s, \mathbf{d}_i) = E(\mathbf{y}_{t+s}|\mathbf{v}_t = \mathbf{d}_i; X_t) - E(\mathbf{y}_{t+s}|\mathbf{v}_t = \underline{\mathbf{0}}; X_t) \qquad s = 0, 1, 2, ... \tag{1}$$

where the operator $E(.|.)$ denotes the best, mean squared error predictor; $\mathbf{y}_t$ is an $n \times 1$ random vector; $X_t \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ...)'$; $\underline{0}$ is of dimension $n \times 1$; $\mathbf{v}_t$ is the $n \times 1$ vector of reduced-form disturbances; and D is an $n \times n$ matrix, whose columns $\mathbf{d}_i$ contain the relevant experimental shocks.

Time provides a natural arrangement of the dynamic causal linkages among the variables in $\mathbf{y}_t$, but does not identify its contemporaneous causal relations. The VAR literature has often relied on assuming a Wold-causal order for the elements of $\mathbf{y}_t$ to organize the triangular factorization of the reduced-form, residual variance-covariance matrix, $\Omega = PP'$. Such an identification mechanism, for example, is equivalent to defining the experimental matrix as $D = P^{-1}$ so that its i^{th} column, $\mathbf{d}_i$, then represents the "structural shock" to the i^{th} element in $\mathbf{y}_t$ (in the usual parlance of the VAR literature).

Statistical-based structural identification of contemporaneous causal links is so far elusive.[1] Further, a one-time shock to a given variable in the system may not be the only type of experiment of interest. For these reasons and to encompass broad designs, the remainder of the analysis is accommodates a generic choice of experiment D without loss of generality. Identification is an important issue but not one that is explored in this paper.

Expression (1) shows that the statistical objective in calculating impulse responses is to obtain the best, mean-squared, multi-step predictions. These can be calculated by recursively iterating on an estimated model optimized to characterize the dependence structure of successive observations, of which a VAR is an example. While this approach is optimal if indeed the postulated model correctly represents the DGP, better multi-step predictions can often be found with direct forecasting models that are reestimated for each forecast horizon. Therefore, consider projecting $\mathbf{y}_{t+s}$ onto the linear space generated by $(\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-p})'$, specifically

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

[1] Exceptions are the work by Granger and Swanson (1997), and Demiralp and Hoover (2003), for example.

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 denote the collection of h regressions in (2) as *local projections,* a term aptly evocative of non-parametric considerations.

According to definition (1), the impulse responses from the local-linear projections in (2) are

$$\widehat{IR}(t, s, \mathbf{d}_i) = \widehat{B}_1^s \mathbf{d}_i \qquad s = 0, 1, 2, ..., h \tag{3}$$

with the obvious normalization $B_1^0 = I$. An extensive literature (see Bhansali, 2002 and references therein) on the direct, multi-step forecasts implied by (2) establishes their consistency and asymptotic normality properties (see Weiss, 1991). However, here we are interested in establishing the consistency and distributional properties of the estimates $\widehat{B}_1^s$ – the impulse response coefficients. This is rather straightforward: as the next section shows, the residuals u_{t+s}^s in (2) are a moving average of the forecast errors from time t to $t+s$ and therefore uncorrelated with the regressors, which are dated $t-1$ to $t-p$.

A few final practical comments conclude this section. First, the maximum lag p (which can be determined by information criteria, for example) need not be common to each horizon s (to see this consider a $VMA(q)$ DGP, for example). Second, the lag length, and the dimension of the vector $\mathbf{y}_t$ will impose degree-of-freedom constraints on the maximum practical horizon h for very small samples. Third, consistency does not require that the sequence of h system regressions in (2) be estimated jointly – the impulse response for the j^{th} variable in $\mathbf{y}_t$ can be estimated by univariate regression of y_{jt} onto $X_t \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-p})'$.

Finally, local projections are also useful in computing the forecast-error variance decomposition. By definition, the error in forecasting $\mathbf{y}_{t+s}$ is given from expression (2) by

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

from which the unnormalized mean squared error (MSE_u) is

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

The choice experiment D renormalizes MSE_u into

$$MSE(E(\mathbf{y}_{t+s}|X_t)) = D^{-1}E(\mathbf{u}^s_{t+s}\mathbf{u}^{s\prime}_{t+s})D'^{-1} \qquad s = 0, 1, 2, ..., h \tag{4}$$

from which the traditional variance decompositions can be calculated by directly plugging in the sample-based equivalents from the projections in (2). For comparison, in traditional VARs the unnormalized MSE is

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

where the Ψ_i and $E(\mathbf{u}^0_t\mathbf{u}^{0\prime}_t)$ are derived 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.

2.2 Relation to VARs and Inference

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

$$\mathbf{y}_t = \boldsymbol{\mu} + \Pi' X_t + \mathbf{v}_t \tag{5}$$

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[2]

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

$$W_t \equiv \begin{bmatrix} \mathbf{y}_t - \boldsymbol{\mu} \\ \mathbf{y}_{t-1} - \boldsymbol{\mu} \\ \vdots \\ \mathbf{y}_{t-p+1} - \boldsymbol{\mu} \end{bmatrix}; F \equiv \begin{bmatrix} \Pi_1 & \Pi_2 & ... & \Pi_{p-1} & \Pi_p \\ I & 0 & ... & 0 & 0 \\ 0 & I & ... & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & I & 0 \end{bmatrix}; \boldsymbol{\nu}_t \equiv \begin{bmatrix} \mathbf{v}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{6}$$

and then realizing that according to (5) and (6),

$$W_t = FW_{t-1} + \boldsymbol{\nu}_t \tag{7}$$

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

$$\mathbf{y}_{t+s} - \boldsymbol{\mu} = \mathbf{v}_{t+s} + F_1^1\mathbf{v}_{t+s-1} + ... + F_1^s\mathbf{v}_t + F_1^{s+1}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + ... + F_p^{s+1}(\mathbf{y}_{t-p} - \boldsymbol{\mu}) \tag{8}$$

where F_i^s is the i^{th} 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 (5) is

$$\mathbf{y}_t = \boldsymbol{\gamma} + \mathbf{v}_t + F_1^1\mathbf{v}_{t-1} + F_1^2\mathbf{v}_{t-2} + ... + F_1^s\mathbf{v}_{t-s} + ... \tag{9}$$

and the impulse response function is given by

$$IR(t, s, \mathbf{d}_i) = F_1^s\mathbf{d}_i$$

Expression (8) establishes the relationship between the impulse responses calculated by local projection and with a VAR. Specifically, comparing expression (2), repeated here for convenience,

$$\mathbf{y}_{t+s} = \boldsymbol{\alpha}^s + B_1^{s+1}\mathbf{y}_{t-1} + B_2^{s+1}\mathbf{y}_{t-2} + ... + B_p^{s+1}\mathbf{y}_{t-p} + \mathbf{u}_{t+s}^s \qquad s = 0, 1, 2, ..., h \tag{10}$$

with expression (8) rearranged,

$$\mathbf{y}_{t+s} = (I - F_1^s - ... - F_p^s)\mu + F_1^{s+1}\mathbf{y}_{t-1} + ... + F_p^{s+1}\mathbf{y}_{t-p} + (\mathbf{v}_{t+s} + F_1^1\mathbf{v}_{t+s-1} + ... + F_1^s\mathbf{v}_t) \tag{11}$$

it is obvious that,

$$\begin{aligned} \boldsymbol{\alpha}^s &= (I - F_1^s - ... - F_p^s)\mu \\ B_1^{s+1} &= F_1^{s+1} \\ \mathbf{u}_{t+s}^s &= (\mathbf{v}_{t+s} + F_1^1\mathbf{v}_{t+s-1} + ... + F_1^s\mathbf{v}_t) \end{aligned} \tag{12}$$

Hence, when the DGP is the VAR in (5), expressions (10) and (11) establish the equivalence between impulse responses calculated by local projections and with this VAR. Expression (12) shows that the error terms of the local projection, $\mathbf{u}_{t+s}^s$, are a moving average of the forecast errors from time t to $t+s$, which knowledge can be used to improve efficiency.

Specifically, define $Y_t \equiv (\mathbf{y}_{t+1}, ..., \mathbf{y}_{t+h})$, $V_t \equiv (\mathbf{v}_{t+1}, ..., \mathbf{v}_{t+h})$, and $X_t \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-p})$, so that we can stack the h local projections in expression (10) and take advantage of the structure of the residuals implied by the VAR assumption and estimate the following system jointly

$$Y_t = X_t\Psi + V_t\Phi \tag{13}$$

where (ignoring the constant terms) the parameter matrices are constrained as follows

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

Defining $E(\mathbf{v}_t\mathbf{v}_t') = \Omega_v$, then $E(V_tV_t') = \Phi\left(I_h \otimes \Omega_v\right)\Phi' \equiv \Sigma$.

Maximum likelihood estimation of this system can then be accomplished by standard GLS formulas according to,

$$vec(\widehat{\Psi}) = \left[(I \otimes X)'\,\Sigma^{-1}\,(I \otimes X)\right]^{-1}(I \otimes X)'\,\Sigma^{-1}vec(Y) \tag{14}$$

The usual impulse responses would then be given by rows 1 through n and columns 1 through (nh) of $\widehat{\Psi}$ and standard errors are provided directly from the regression output. Further simplification is available due to the special structure of the variance-covariance matrix Σ, which allows GLS estimation of the system block by block.

In fact, ML estimation of (13) delivers asymptotically exact formulas for single and joint inference on the impulse response coefficients of the implied VAR rather than the usual pointwise, analytic, delta-method approximations (see Hamilton, 1994; Chapter 11), or simulation based estimates based on Monte Carlo or bootstrap replications.[3] In general the true DGP is unknown and so is the specific structure of Φ, hence the GLS restrictions described above are unavailable. This poses no difficulty, however, as the error terms u_{t+s}^s in expression (2) will follow some form of moving-average structure whose order is a function of the horizon s. Thus, impulse responses can be calculated by univariate regression methods with a heteroskedasticity and autocorrelation (HAC) robust estimator with little loss of efficiency. In principle, the efficiency of these estimators can be improved upon by recursively including the residuals of the stage $s-1$ local projection as

[3] Sims and Zha (1999) provide methods for joint inference in impulse responses but they involve complicated and rather computationally intensive Bayesian methods that most practitioners have not yet adopted.

regressors in the stage s local projection – an improvement whose details are reserved for another paper.

In practice the DGP is unknown and it is preferable to make as few assumptions as possible on its specification. Thus valid inference for local projection impulse responses can be obtained with HAC robust standard errors. For example, let $\widehat{\Sigma}_L$ be the estimated HAC, variance-covariance matrix of the coefficients $\widehat{B}_1^s$ in expression (2), then 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, unrestricted local projections experience small efficiency losses.

2.3 Comparison with Recent Impulse Response Estimators

Chang and Sakata (2002), Cochrane and Piazzesi (2002), and Thapar (2002) have recently introduced impulse response estimators that 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.[4] Chang and Sakata (2002) calculate $\widehat{v}_t$ with an autoregression, Cochrane and Piazzesi (2002) with forecast errors implied by financial prices, and Thapar (2002) with errors in surveys of forecasts. Hence, all of these methods can be seen as a truncated version of Barro's (1977, 1978) well known regressions.

The major selling point is that the error series $\widehat{v}_t$ is "fundamental" in some sense. The argument goes that because forecast errors are constructed from market-based (rather than econometric) expectations, all available information is appropriately incorporated and, in addition, one can dispense with the thorny issue of identification. These two features make these methods attractive.

However, there are some trade-offs to be considered. In general, it is perilous to disassociate the series of "shocks" from the model that generated them, specially in a multivariate context. The Wold decomposition theorem (see Brockwell and Davis, 1991) ensures that any covariance-

[4] The ensuing discussion is in the univariate context, hence the lowercase notation.

stationary process can be expressed as an infinite moving average of forecast errors 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.

Additionally, market-based expectations are available for a limited number of variables. Econometrically, except for Thapar (2002), the second stage regression includes moving-average terms involving information dated $t-1, t-2, ...$ which is problematic for consistency (to see this, substitute the Wold decomposition of y_t into the second stage regressions). Finally, it is difficult to produce correct inference as the second stage uses generated regressors, thus requiring bootstrap methods.

Impulse responses characterize the partial derivatives between different elements in $\mathbf{y}_t$ over time in the multi-dimensional process that relates $\mathbf{y}_t$ to its past. Thus, while small variations in the specification of this multi-dimensional process 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 critique of VARs 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), 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.

3 Flexible Local Projections

Linear models, such as VARs, bestow four restrictive properties to their implied impulse responses:[5] (1) *symmetry,* responses to positive and negative shocks are mirror images of each

[5] For a detailed discussion see Koop et al. (1996).

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 high-dimensional nonlinear functions of parameter estimates which complicate the calculation of standard errors and quickly compound misspecification errors. For example, there is no reason to expect that the output loss due to higher interest rates will be equivalent to the output gain when interest rates are lowered, nor that the output loss will simply double when interest rates double as well, nor that the same increase in interest rates will have the same effect on output whether we are coming out of a recession or just plunging into one.

Although local-linear projection methods dispense with the fourth of these problems, they are indeed linear and thus constrained by properties (1)-(3) just the same. In a traditional multivariate, model-based setting, investigation of nonlinearities is limited by at least three considerations: (1) the ability to jointly estimate a nonlinear system of equations with its inherent computational difficulties; (2) the complexity 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. Hence, it is natural to explore alternatives based on local projections.

Under mild assumptions, a non-linear time series process $\mathbf{y}_t$ can be expressed 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 approximated by 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),

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \Phi_i \mathbf{v}_{t-i} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \Phi_{ij} \mathbf{v}_{t-i}\mathbf{v}_{t-j} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \Phi_{ijk} \mathbf{v}_{t-i}\mathbf{v}_{t-j}\mathbf{v}_{t-k} + ... \tag{15}$$

which is a polynomial extension of the Wold decomposition in expression (9) with the constant omitted for simplicity. Similarly, consider extending the local projections in expression (2) with polynomial terms on $\mathbf{y}_{t-1}$:[6]

$$\begin{aligned} \mathbf{y}_{t+s} &= \boldsymbol{\alpha}^s + B_1^{s+1}\mathbf{y}_{t-1} + Q_1^{s+1}{\mathbf{y}_{t-1}}^2 + C_1^{s+1}{\mathbf{y}_{t-1}}^3 + \\ &\quad B_2^{s+1}\mathbf{y}_{t-2} + ... + B_p^{s+1}\mathbf{y}_{t-p} + \mathbf{u}_{t+s}^s \qquad s = 0, 1, 2, ..., h \end{aligned} \tag{16}$$

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 rather than as a requirement. It is readily apparent that the impulse response at time s now becomes,

$$\begin{aligned} IR(t, s, \mathbf{d}_i) &= \left\{ \widehat{B}_1^s (\mathbf{y}_{t-1} + \mathbf{d}_i) + \widehat{Q}_1^s (\mathbf{y}_{t-1} + \mathbf{d}_i)^2 + \widehat{C}_1^s (\mathbf{y}_{t-1} + \mathbf{d}_i)^3 \right\} - \\ &\quad \left\{ \widehat{B}_1^s \mathbf{y}_{t-1} + \widehat{Q}_1^s (\mathbf{y}_{t-1})^2 + \widehat{C}_1^s (\mathbf{y}_{t-1})^3 \right\} \\ &= \left\{ \widehat{B}_1^s \mathbf{d}_i + \widehat{Q}_1^s (2\mathbf{y}_{t-1}\mathbf{d}_i + \mathbf{d}_i^2) + \widehat{C}_1^s (3\mathbf{y}_{t-1}^2 \mathbf{d}_i + 3\mathbf{y}_{t-1}\mathbf{d}_i^2 + \mathbf{d}_i^3) \right\} \\ s &= 0, 1, 2, ..., h \end{aligned} \tag{17}$$

and with the obvious normalizations, $B_1^0 = I$, $Q_1^0 = 0_n$, and $C_1^0 = 0_n$. These nonlinear estimates can be easily calculated by least squares, equation by equation, with any conventional econometric software. When some of the terms Q_i^s and C_i^s are non-zero, the impulse response function will now vary according to the sign and with the size of the experimental shock defined by $\mathbf{d}_i$, thus dispensing with the symmetry and shape invariance restrictions. In addition, the impulse response depends on the local history $\mathbf{y}_{t-1}$ at which it is evaluated. In particular, impulse responses

[6] Since the impulse response coefficients involve the terms $\mathbf{y}_{t-1}$ only, it seems more parsimonious to restrict nonlinearities to these terms alone. In practice, if degrees of freedom are not a consideration, they can be extended to the remaining regressors although the gain of doing so is probably small.

comparable to local-linear or VAR-based impulse responses can be achieved by evaluation at the sample mean, i.e. $\mathbf{y}_{t-1} = \overline{\mathbf{y}}_{t-1}$. Different responses will be obtained if a different experimental value for $\mathbf{y}_{t-1}$ is chosen and one can consider a 3-D plot of the impulse response for a range of values for $\mathbf{y}_{t-1}$.[7] Finally, the 95% confidence interval for the cubic approximation in expression (16) can be easily calculated. Define the scaling $\boldsymbol{\lambda}_i \equiv (\mathbf{d}_i, \quad 2\mathbf{y}_{t-1}\mathbf{d}_i + \mathbf{d}_i^2, \quad 3\mathbf{y}_{t-1}^2\mathbf{d}_i + 3\mathbf{y}_{t-1}\mathbf{d}_i^2 + \mathbf{d}_i^3)'$ and denote $\widehat{\Sigma}_C$ the HAC, variance-covariance matrix of the coefficients $\widehat{B}_1^s, \widehat{Q}_1^s$, and $\widehat{C}_1^s$ in (16). Then, a 95% confidence interval for the impulse response at time s is approximately, $1.96 \pm \left(\boldsymbol{\lambda}_i' \widehat{\Sigma}_C \boldsymbol{\lambda}_i\right)$.

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

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

where $m^s(.)$ may include any parametric, semi-parametric and non-parametric approximation, and for which there is a rather extensive list of possible specifications to choose from. 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 polynomial projections with threshold effects.

4 Monte Carlo Evidence

This section contains two main simulations that evaluate the performance of local projections for impulse response estimation and inference. The first experiment is based on a standard monetary VAR that appears in Christiano, Eichenbaum and Evans (1996) and Evans and Marshall (1998), among many other papers. 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

[7] Potter (2000) contains a detailed and more formal discussion of alternative ways of defining and reporting nonlinear impulse responses in general.

(see Jordà and Salyer, 2003) to show that flexible local projections do a reasonable job at approximating the inherent nonlinearities of this model, and compares its performance relative to a Bayesian VAR with time-varying parameters and volatilities – a natural flexible alternative to conventional VARs.

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, and use the first 12 observations from the data to initialize all 500 runs. Information criteria based on the original data suggest the lag-length to be twelve when using Akaike's AIC and Hurvich and Tsai's[8] AIC_c, or two when using Schwartz's SIC. These choices are very consistent across the 500 simulated runs.[9]

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,[10] calculated as

[8] Hurvich and Tsai (1993) is a correction to AIC specifically designed for VARs and with superior properties to either AIC or SIC.

[9] 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.

[10] Responses to shocks in all the variables are available upon request and are not reported here in the interest of space.

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) 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[11] 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[12] corrected standard errors for the local projections. Table 1 reports these results for each variable in response to a shock in FF as well.

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 14^{th} 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, 12 lags, and horizons of 24 periods.

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

[12] The Newey-West lag correction is selected to increase with s, the horizon of the impulse response being considered.

4.2 Impulse Responses and Nonlinearities

The following Monte-Carlo experiment compares impulse responses calculated by local projections methods with a traditional and a flexibly parametrized VAR when the DGP is nonlinear. The specific DGP for this experiment is based on the SVAR-GARCH model in Jordà and Salyer (2003), which is a multivariate version of a traditional GARCH-M model. Here, I experiment with the following specification,

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = A \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{3t-1} \end{bmatrix} + Bh_{1t} + \begin{bmatrix} \sqrt{h_{1t}}\varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix}, \qquad \varepsilon_t \sim N(0, I_3) \tag{18}$$

$$h_{1t} = 0.5 + 0.3u_{1,t-1} + 0.5h_{1,t-1}; \qquad u_{1t} = \sqrt{h_{1t}}\varepsilon_{1t}$$

$$A = \begin{bmatrix} 0.5 & -0.25 & 0.25 \\ 0.75 & 0.25 & 0.25 \\ -0.25 & -0.25 & 0.75 \end{bmatrix}; \qquad B = \begin{bmatrix} -1.75 \\ -1.5 \\ 1.75 \end{bmatrix}$$

and a sample size of 300, replicated 500 times. This DGP is advantageous for several reasons. First, the SVAR-GARCH nests a linear VAR(1) and in fact impulses responses to shocks to ε_{2t} or ε_{3t}, or to "small" shocks to ε_{1t}, are equivalent in both models. Discrepancies arise with larger shocks to ε_{1t} since then there is a revision of their conditional variance (due to the GARCH term) that affects the conditional mean and makes the responses more nonlinear. Second, since I also specify a time-varying parameter/volatility VAR[13] (TVPVAR) à là Cogley and Sargent (2001) as a flexible approximator to the DGP, it is useful that the nonlinearity be of a smooth nature (say, relative to a model with structural breaks or switching-regimes). Notice that the DGP will have time-varying volatility with some effects on the conditional mean and one would expect that the

[13] I thank Tim Cogley for all his advice on the numerous intricacies of this model and Massimiliano de Santis for his invaluable assistance in estimating the model with his GAUSS code. For further details on the specification and estimation of the model the reader is referred to Cogley and Sargent (2001, 2003). The GAUSS code for estimating the time-varying parameter VAR can be obtained by e-mailing Massimiliano de Santis directly at: mdesantis@ucdavis.edu.

TVPVAR is well suited to capture these features.

As a foil to the cubic projection, the TVPVAR is estimated with Bayesian methods for each of the 500 Monte Carlo replications using the first 100 observations to calibrate the prior, leaving the remaining 200 for inference. The estimator is based on a Gibbs-sampler initialized with 2,000 draws and allowed to run for an additional 5,000 iterations to ensure convergence. This produces a history of 200 observations for each estimated parameter in the model. To calculate the impulse responses, I select the quintiles of the distribution of the residual for the first equation (the one with GARCH effects) which identify five dates from the last 200 observations in the sample (the ones with time-varying parameters). This allows the TVPVAR to tailor the impulse responses to different values of the conditional variance and to better capture any resulting nonlinearities.[14]

Calculating impulse responses for each of these five selected dates requires an additional Monte Carlo since the parameters of the model are varying over time stochastically. Hence, I generate 100 sequences of 1-8 step-ahead forecasts, conditional on the parameter values at each of the five dates previously selected and the driving processes estimated from the data. The average over these 100 histories is used to produce the impulse responses.[15]

Figure 3 displays the impulse responses from a shock to ε_{1t} of unit in size.[16] The thick solid lines in each graph represent the true impulse responses with and without GARCH effects (i.e. $B = \underline{0}$), the less variable of the two representing the latter case. Indistinguishable from the impulse responses generated when the GARCH effects are switched-off, both the linear VAR(1), and the linear projection responses are displayed by thin-dashed lines. Finally, the thick-dashed line with crosses displays the cubic projection responses, whereas the thick line with squares displays the TVPVAR responses, averaged over the five selected days (it will become clear momentarily why

[14] A provision in the code discards any Monte Carlo draws for which the stationarity condition for the distribution of the parameters is violated. If a draw is discarded, the Monte Carlo runs for an additional draw. In the end, 5-10% of the draws had to be replaced.

[15] The complete Monte-Carlo took nine days, two hous, and 17 minutes on a SUN Sunfire server with eight 900 Mhz processors and 16GB of RAM memory.

[16] Shocks to ε_{2t} and ε_{3t} would simply produce the usual linear VAR responses.

the averaging). Standard errors are omitted from the figure to improve clarity. Suffice it to say that conventional error bands are very narrow and clearly separate impulse responses from the DGP with and without GARCH effects, except at crossing points.

Several results deserve comment. The VAR(1), the linear projections, and the cubic projections evaluated at the sample mean (not reported), precisely capture the shape of the impulse responses from the DGP without GARCH effects, even though these are estimated from a sample generated without this constraint. The true impulse responses with GARCH effects are far more variable and to capture this feature, I consider cubic projections evaluated at five standard deviations away from the mean and I consider responses from the TVPVAR evaluated at the five dates selected previously. In the end, the TVPVAR responses displayed no variability for the first six to seven periods. After that, they fan out in different directions, much like the picture of a forecast confidence interval. Hence, to simplify the figure, I report the average over the five dates. As figure 3 clearly shows, the TVPVAR responses were unable to capture the nonlinearities in the model where as the cubic projections provided a much closer fit to the true impulse responses. Overall, local polynomial projections seem to afford better control over smooth nonlinearities since they nest linearity and their complexity can be easily controlled.

5 Application: Inflation-Output Trade-offs

Pioneering work by McCallum (1983) and Taylor (1993) inspired a remarkable amount of research on the efficacy, optimality, credibility, and robustness of interest rate rules for monetary policy. The performance of candidate policy rules is often evaluated in the context of a simple, new-Keynesian, 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 the dynamic properties of inflation, the output gap, and interest rates to provide a benchmark for competing theoretical models. The specific definitions of the variables I consider correspond to the definitions in Rudebusch and Svensson (1999) and are relatively standard for this literature: 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;[17] and i_t is the quarterly average of the federal funds rate in percent at an annual rate. The sample of quarterly data runs over the period 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. Similar studies, such as Galí (1992) and Fuhrer and Moore (1995a, b), use four lags for variables analyzed in the levels. This selection is confirmed by AIC_c, while AIC suggests six lags and SIC suggests two lags. Therefore, figure 5 displays the impulse responses based on a VAR(4), and local-linear and -cubic projections, all identified with a standard Cholesky decomposition[18] and the Wold-causal order y_t,π_t, and i_t.

The VAR(4) responses are depicted with a thick line, the solid line with crosses and the two accompanying dashed lines depict the responses from local-linear projections and the corresponding Newey-West corrected, two standard-error bands. The solid line with circles is the response from local-cubic projections evaluated at the sample mean. 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. Generally speaking, there is broad correspondence among the responses calculated by the different methods, with a few exceptions.

The cubic projection responses show that inflation is considerably more persistent to its own shocks than what is reflected by the responses calculated by either linear method. Perhaps not

[17] I thank an anonymous referee for spotting that I had used the *quantity* index in the previous version of this paper.

[18] This Cholesky ordering is consistent with the literature and facilitates replicability.

surprisingly, the associated interest rate response is also almost twice as aggressive (at 0.75% versus 0.4%), 12 quarters after impact. The responses of the system to interest rate shocks are perhaps more interesting from an economic point of view because they give us an idea of the relative output and inflation trade-offs in response to monetary policy. The VAR response of the output gap suggests a loss of output of 0.25% in response to a 0.8% increase in the fed funds rate, 12 quarters after impact. This loss is approximately half what linear and cubic projections predict (at around 0.5%). On the other hand, the response of inflation to this increase in the fed funds rate is mostly positive but not significantly different from zero. The linear projection is similar for the first seven quarters but is significantly negative thereafter, whereas the cubic projections show a more positive initial inflation response with a dramatic decline around quarter seven as well.

Based on this preliminary analysis, we now investigate for further nonlinearities in the impulse responses. 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 projections approximate smooth nonlinearities, they are less helpful in detecting the type of nonlinearity implicit in these examples. Therefore, I tested all linear projections[19] for evidence of threshold effects due to all four lags of y_t, π_t, and i_t using Hansen's (2000) test[20] . For example, a typical regression is,

$$\begin{aligned} z_t &= \boldsymbol{\rho}_L' X_{t-1} + \varepsilon_t^L \qquad \text{if } w_{t-j} \leq \delta \\ z_t &= \boldsymbol{\rho}_H' X_{t-1} + \varepsilon_t^H \qquad \text{if } w_{t-j} > \delta \end{aligned} \tag{19}$$

were z_t is respectively y_t, π_t, and i_t and w_{t-j} can be any of y_{t-j}, π_{t-j}, and i_{t-j}, $j \in \{1, 2, 3, 4\}$.

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

[20] The GAUSS routines to perform the test are available directly from Bruce Hansen's web site.

X_{t-1} collects lags 1 through four of the variables y_t, π_t, and i_t and $\boldsymbol{\rho}_k$, $k = L, H$ collects the coefficients and L stands 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.

Table 2 summarizes the results of these tests by reporting the estimated thresholds and p-values (in parenthesis) for all possible combinations of endogenous and threshold variables. Several results deserve comment. First, there are no "business-cycle" asymmetries associated to threshold effects in the output gap. Second, the null of linearity is rejected across equations for several lags of both inflation and the federal funds rate. Third, there is considerable correspondence between the estimated threshold values for the lags of a given variable across all equations. Fourth, the reported estimated threshold values correspond to the value that maximizes the likelihood. However, note that when the null of linearity is rejected, it is often rejected for an interval around this optimal value.

Despite the apparent complexity of these results, the overall message that emerges is straightforward: the estimated thresholds are dividing the data into the turbulent period of the 1970s to mid-1980s (I call it the "high-inflation" regime) and the rest of the sample (the "low-inflation" regime). Consequently, it is natural to consolidate these results by conducting two experiments. In the first experiment I allow for a threshold in the third lag of inflation at 4.75%. In the second, the threshold is in the third lag of the federal funds rate at 6% instead. Figure 4 displays these two thresholds in reference to the raw data to illustrate that the main difference is that the threshold in the federal funds rate extends the high-inflation regime up to the late 1980s.

Figure 6 compares the responses to a shock in the federal funds rate for these two experiments.[21] In particular, the left column displays the graphs corresponding to the inflation threshold, while the right column displays the graphs for the federal funds rate threshold. The solid thick line and the dashed lines are the cubic projections evaluated at the sample mean and the corresponding

[21] The responses to the remaining shocks are omitted for brevity but are available upon request.

Newey-West corrected, two standard-error bands. The solid line with crosses correspond to the low-inflation regime responses whereas the solid line with circles are the high-inflation regime responses. All experiments are normalized to a 0.8% shock in the federal funds rate to facilitate comparability (this is a one standard-error shock which is the one most often reported in standard econometric packages).

Figure 6 clearly shows that the nature of the inflation-output trade-offs varies quite substantially depending on regime but does not really depend on which variable is used as a threshold. Generally speaking, inflation and output are far more responsive to interest rates in the low-inflation regime than in the high-inflation regime, even though the federal funds rate responds somewhat more aggressively in the latter.

This empirical application is illustrative in several dimensions. Although the evidence is not definitive, these results support the view held by Cogley and Sargent (2001) among others, that the adverse inflation-unemployment outcomes of the 1970s where not the result of bad policy (as advocated by DeLong, 1997 and Romer and Romer, 2002) but rather the result of a changing economic environment. Perhaps one argument that could undermine these results would suggest that the Fed had become less credible during the 1970s although it seems clear that it had not become any less vigilant (this is specially evident in the responses depicted in the right column and last row of figure 6). The results in figure 6 also suggest that the "prize puzzle" (the common finding in the VAR literature that inflation actually *increases* in response to an increase in interest rates) does not characterize the current economic environment. In fact, the current regime is characterized by rather effective responses of the output gap and inflation to an increase in interest rates, an observation with important implications in the design of contemporary optimal policy responses that are not unduly contractionary. Finally, notice that while we have estimated flexible impulse responses that allow for threshold effects, the entire analysis was conducted by means of simple least squares regressions – an ostensible simplification relative to any multivariate alternative based on a flexible nonlinear model.

6 Conclusion

The first order Taylor series expansion of a function at a given point gives a reasonable approximation to the function in a neighbourhood of that point. However, the more nonlinear the function, the more the quality of the approximation deteriorates as we move further away from the original evaluation point. Similarly, a VAR linearly approximates the DGP to produce optimal, one-period ahead forecasts but impulse responses are functions of forecasts at ever-increasing horizons for which a VAR may provide a poor approximation.

This paper shows that impulse responses can be calculated by a sequence of projections of the endogenous variables shifted forward in time onto its lags. Hence, these projections are local to each forecast horizon and therefore more robust to misspecification of the unknown DGP. Local projections are therefore a natural and preferable alternative to VARs when the object of interest is to calculate impulse responses.

Inference for impulse responses from VARs is difficult because impulse response coefficients are high-dimensional nonlinear functions of estimated parameters. By contrast, local projections directly estimate impulse response coefficients so that standard errors from traditional HAC regression routines provide appropriate joint or point-wise inference.

The principles presented in this paper open a number of new avenues for research. The sequential nature of the local projections allow us to take advantage of the stage $s-1$ forecast errors to improve inference in the stage s projections. Preliminary Monte Carlo evidence not reported here shows significant gains in using this procedure, whose formal derivations are left for a different paper. The same sequential feature of local projections can be used to improve structural identification since any contemporaneous structure among the endogenous variables must remain as we shift time forward through each local projection.

Panel data applications in macroeconomics, where dynamics dominate cross-sectional considerations, are likely to become more prevalent. However, while the high dimensionality of VARs

make impulse response estimation prohibitive in this context, local projections offer a natural and simple alternative for estimation and inference of the dynamics of different treatment effects.

Recent sophisticated solution methods have opened the doors to increasingly complex nonlinear economic models. It is often impractical to calculate impulse response functions from multivariate nonlinear models (for reasons explained in section 3) or simply impossible for non-Gaussian data whose multivariate density is unknown, yet impulse responses can still be calculated relatively simply by local projection methods.

Finally, local projections methods can be used to formalize the estimation of deep parameters in rational expectations models in the manner proposed in Rotemberg and Woodford (1997) and used in several papers thereafter (Christiano, Eichenbaum and Evans, 2001, Amato and Laubach, 2003; and so on). The technique consists in conjecturing a solution path represented by an infinite moving average (MA) and then matching the deep parameters of the model to the MA coefficients with the method of undetermined coefficients. A minimum distance estimator between data-based impulse responses and the theoretically constrained MA coefficients thus produces estimates of these deep parameters. In work in progress, Sharon Kozicki and I use linear projections and optimal GMM-type weights to produce more efficient estimates and standard errors for a wide range of rational expectations models.

References

Amato, Jeffery D. and Thomas Laubach (2003) "Estimation and Control of An Optimization-Based Model with Sticky Prices and Wages," *Journal of Economic Dynamics and Control,* 27, 1181-1215.

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, Rajendra J. (1996) "Asymptotically Efficient Autoregressive Model Selection for Multistep Prediction," *Annals of the Institute of Statistical Mathematics,* 48, 577-602.

Bhansali, Rajendra J. (1997) "Direct Autoregressive Predictors for Multistep Prediction: Order Selection and Performance Relative to the Plug-in Predictors," *Statistica Sinica*, 7, 425-449.

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, 2^{nd} edition. Heidelberg, New York and Berlin: Springer-Verlag.

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.

Christiano, Lawrence J., Martin Eichenbaum and Charles L. Evans (2001) "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," NBER working paper 8403.

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.

Cogley, Timothy W. and Thomas J. Sargent (2001) "Evolving Post World War II U.S. Inflation Dynamics," **NBER Macroeconomics Annual 16**, 331-373.

Cogley, Timothy W. and Thomas J. Sargent (2003) "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S." *mimeo*, University of California, Davis.

Cooley, Thomas F. and Mark Dwyer (1998) "Business Cycle Analysis without much Theory: A Look at Structural VARs," *Journal of Econometrics,* 83, 1-2, 57-88.

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.

DeLong, J. Bradford (1997) "America's only Peacetime Inflation: the 1970's" in Christina Romer and David Romer (eds.), **Reducing Inflation. NBER Studies in Business Cycles**, v. 30.

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 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.

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

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

Hansen, Peter R. (2003) "Granger's Representation Theorem: A Closed Form Expression for I(1) Processes," *mimeo*, Stanford University.

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.

Ing, Ching-Kang (2003) "Multistep Prediction in Autoregressive Processes," *Econometric Theory,* 19, 254-279.

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," Federal Reserve Board, Finance and Economics Discussion Series, 2003-61.

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

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.

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.

Romer, Christina D. and David H. Romer (2002) "The Evolution of Economic Understanding and Postwar Stabilization Policy," Federal Reserve Bank of Kansas City, 2002 Jackson Hole Conference Volume.

Rotemberg, Julio J. and Michael Woodford (1997) "An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy." in Ben S. Bernanke and Julio J. Rotemberg (eds.), **NBER Macroeconomics Annual**. Cambridge: MIT Press, 297-346.

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.

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.

Stock, James H. and Mark W. Watson (1999) "A Comparison of Linear and Non-linear Univariate Models for Forecasting Macroeconomic Time Series," in Robert F. Engle and Halbert L. White (eds), **Cointegration, Causality and Forecasting: A Festschrift in Honor of Clive W. J. Granger**. Oxford: Oxford University Press.

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.*

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.

Wallis, Kenneth F. (1977) "Multiple Time Series and the Final Form of Econometric Models," *Econometrica,* 45, 1481-1497.

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

Zellner, Arnold and Franz Palm (1974) "Time Series Analysis and Simultaneous Equation Econometric Models," *Journal of Econometrics, 2, 17-54.*

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

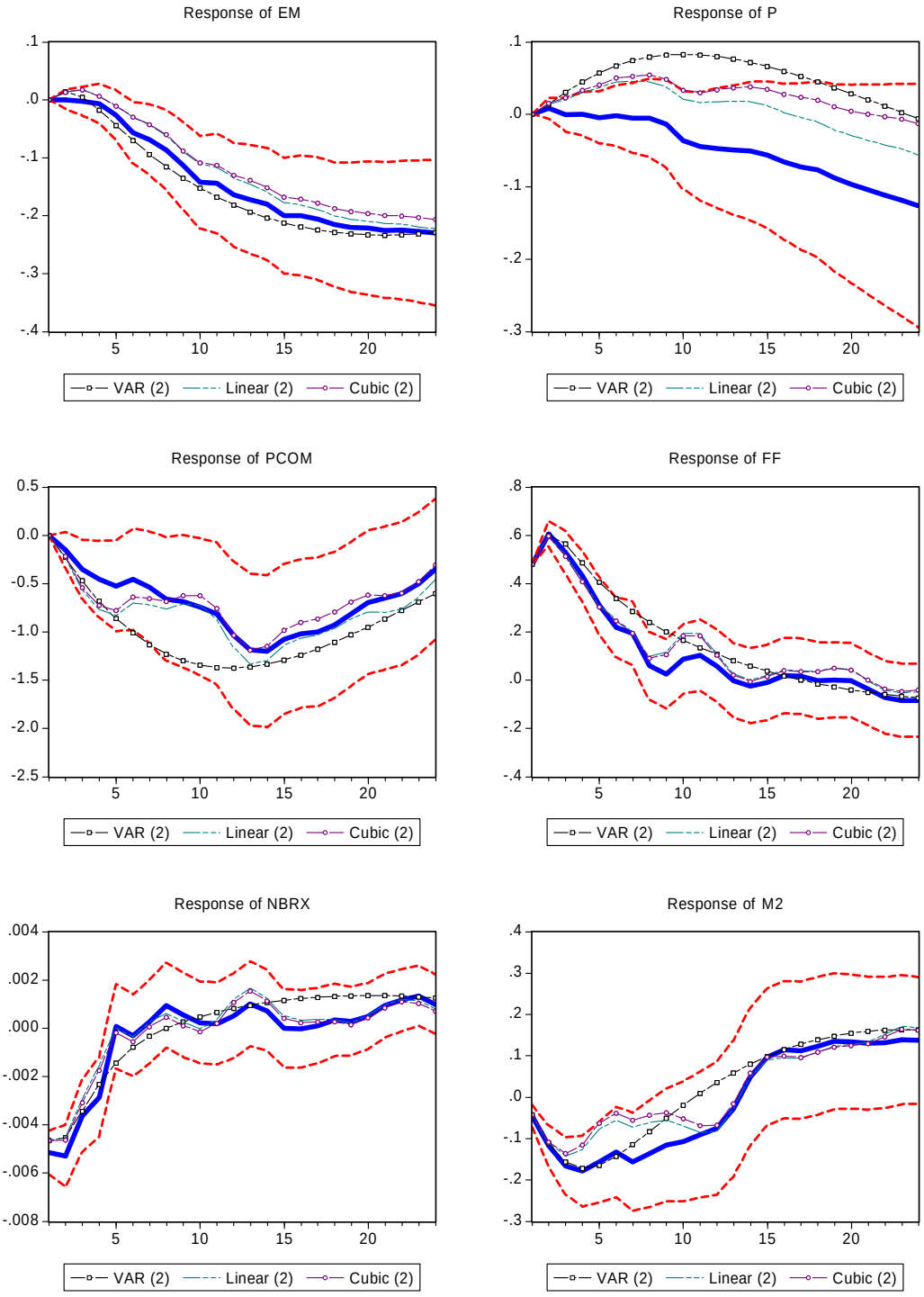
Notes: True-MC refers to the Monte Carlo (500 replications) standard errors for the impulse response coefficients due to a shock in *FF* in a VAR(12) with the variables *EM, P, PCOM, FF, NBRX, ΔM2*. Similarly, Newey-West (linear) refers to standard errors calculated from local-linear projections and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

Table 2 – Hansen's (2000) test for the presence of threshold effects. Threshold estimates and bootstrap p-values

Threshold	**Dependent Variable**		
Variable	Output Gap (y_t)	Inflation (π_t)	Fed Funds (i_t)
y_{t-1}	-0.85	-1.31	-0.09
	(0.74)	(0.62)	(0.24)
y_{t-2}	-1.97	-2.07	-0.85
	(0.73)	(0.33)	(0.33)
y_{t-3}	0.37	-1.42	-2.34
	(0.20)	(0.28)	(0.24)
y_{t-4}	-1.20	-1.25	-2.09
	(0.50)	(0.18)	(0.84)
π_{t-1}	***4.68***	4.00	4.93
	(0.03)	(0.39)	(0.10)
π_{t-2}	4.66	***4.54***	4.24
	(0.09)	***(0.02)***	(0.18)
π_{t-3}	3.91	***5.31***	***4.24***
	(0.30)	***(0.02)***	***(0.00)***
π_{t-4}	2.82	3.25	***3.98***
	(0.13)	(0.59)	***(0.04)***
i_{t-1}	5.94	6.52	***7.88***
	(0.53)	(0.46)	***(0.01)***
i_{t-2}	6.02	5.94	5.56
	(0.19)	(0.92)	(0.21)
i_{t-3}	***6.27***	8.16	***5.82***
	(0.04)	(0.95)	***(0.05)***
i_{t-4}	5.64	5.28	5.09
	(0.55)	(0.37)	(0.07)

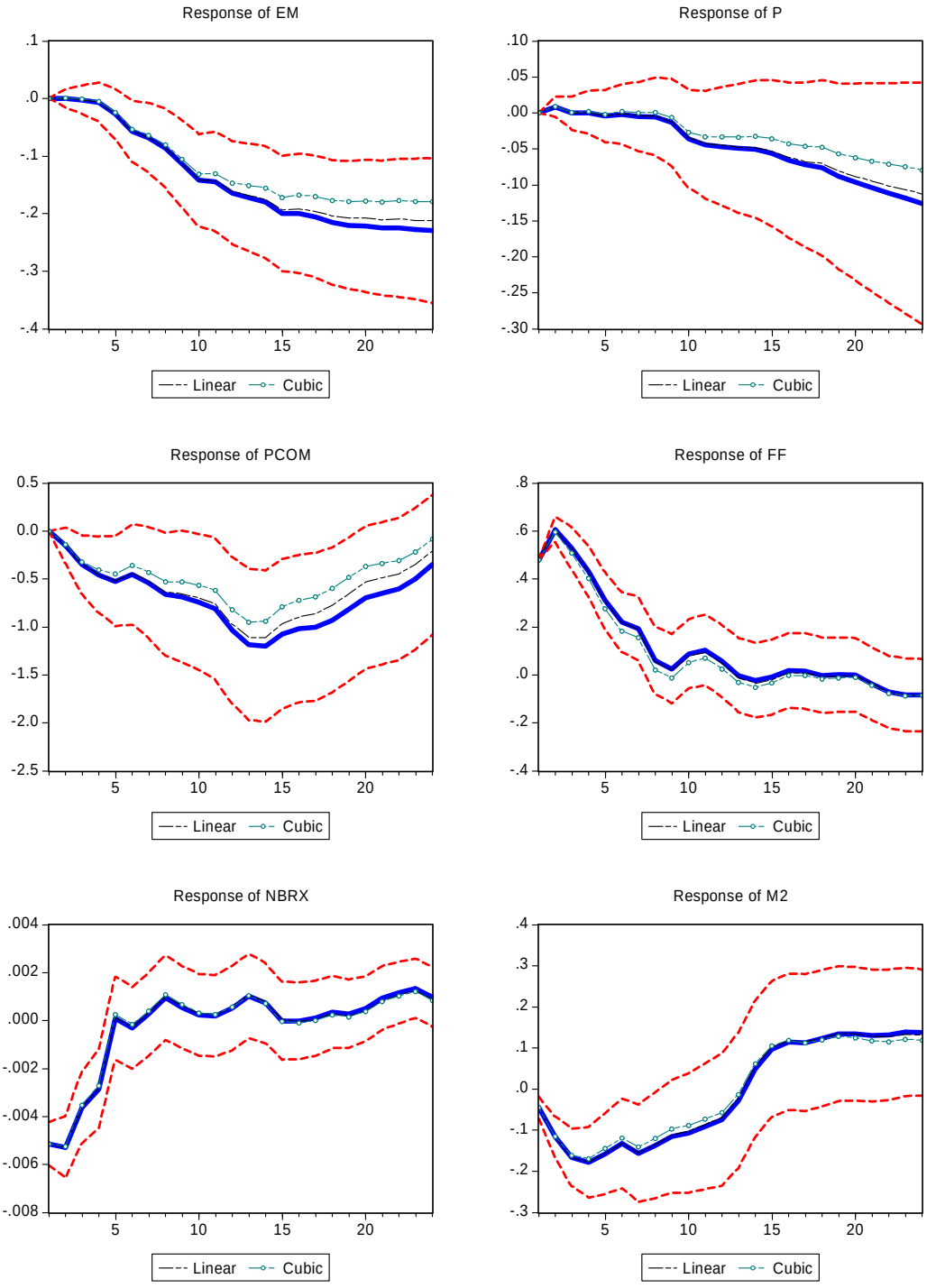
Notes: The equation for each dependent variable contains four lags of each of the dependent variables in the system. The test is an LM-type test for the null hypothesis that there is no threshold effect. Each cell contains the estimate of the optimal threshold value estimate and a bootstrap-based p-value (in parenthesis) calculated with 1,000 draws and a 20% trimming value of the sample to allow for sufficient degrees of freedom. The test corrects for left-over heteroskedasticity. The results on this table were calculated with the GAUSS code that Bruce Hansen makes available on his website based on his 2000 Econometrica paper. Entries in bold and italic signify evidence of a threshold at the conventional 95% confidence level.

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



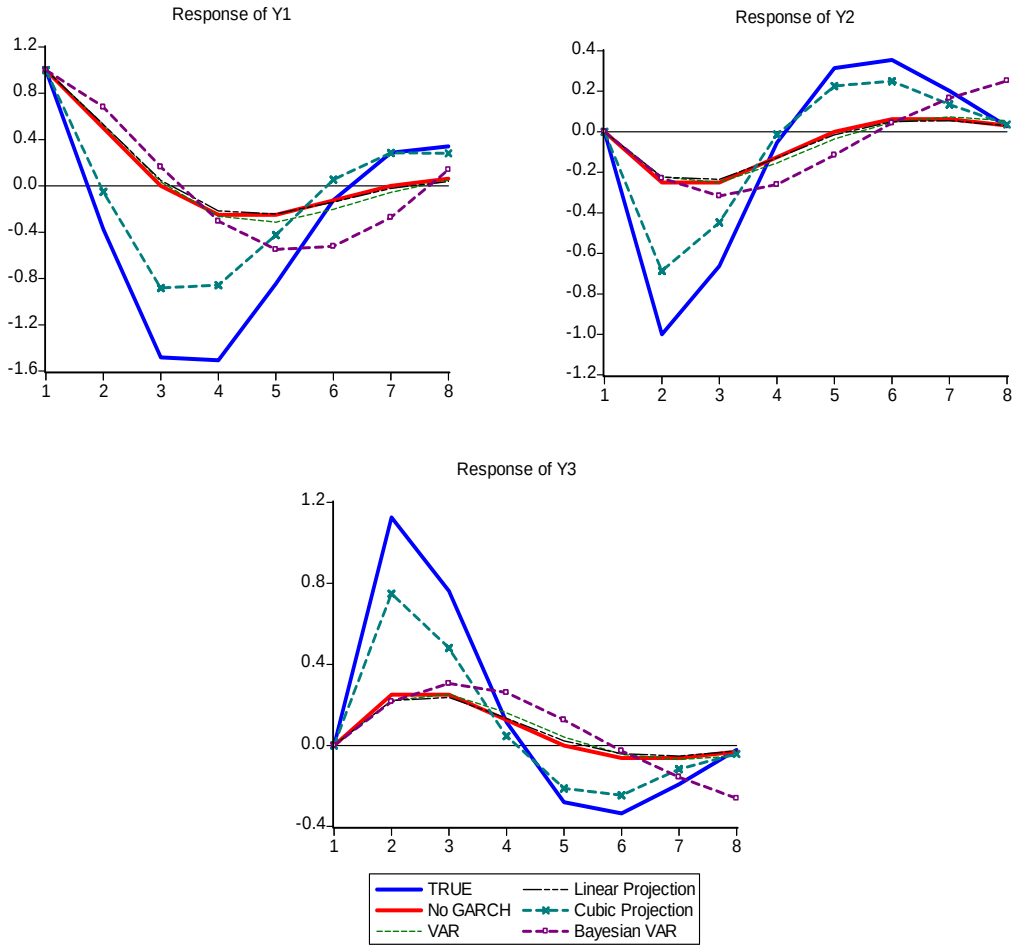
Evans and Marshall (1998) VAR(12) Monte Carlo Experiment. The thick line is the true impulse response based on a VAR(12). The thick-dashed lines are Monte Carlo 2-standard error bands. Three additional impulse responses are compared, based on estimates involving two lags only: (1) the response calculated 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



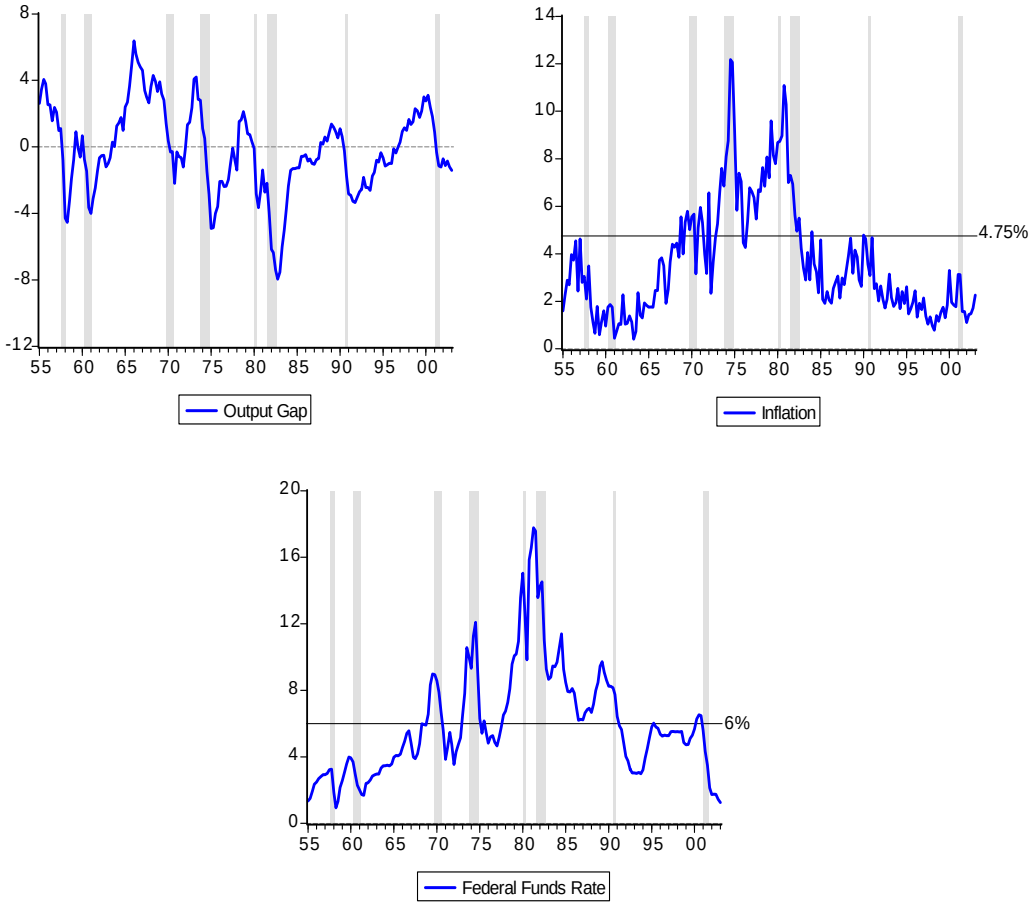
Evans and Marshall (1998) VAR(12) Monte Carlo Experiment. The thick line is the true impulse response based on a VAR(12). The thick-dashed lines are Monte Carlo, 2-standard error bands. Two additional impulse responses are compared: (1) the response calculated with a local-linear projection with 12 lags,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



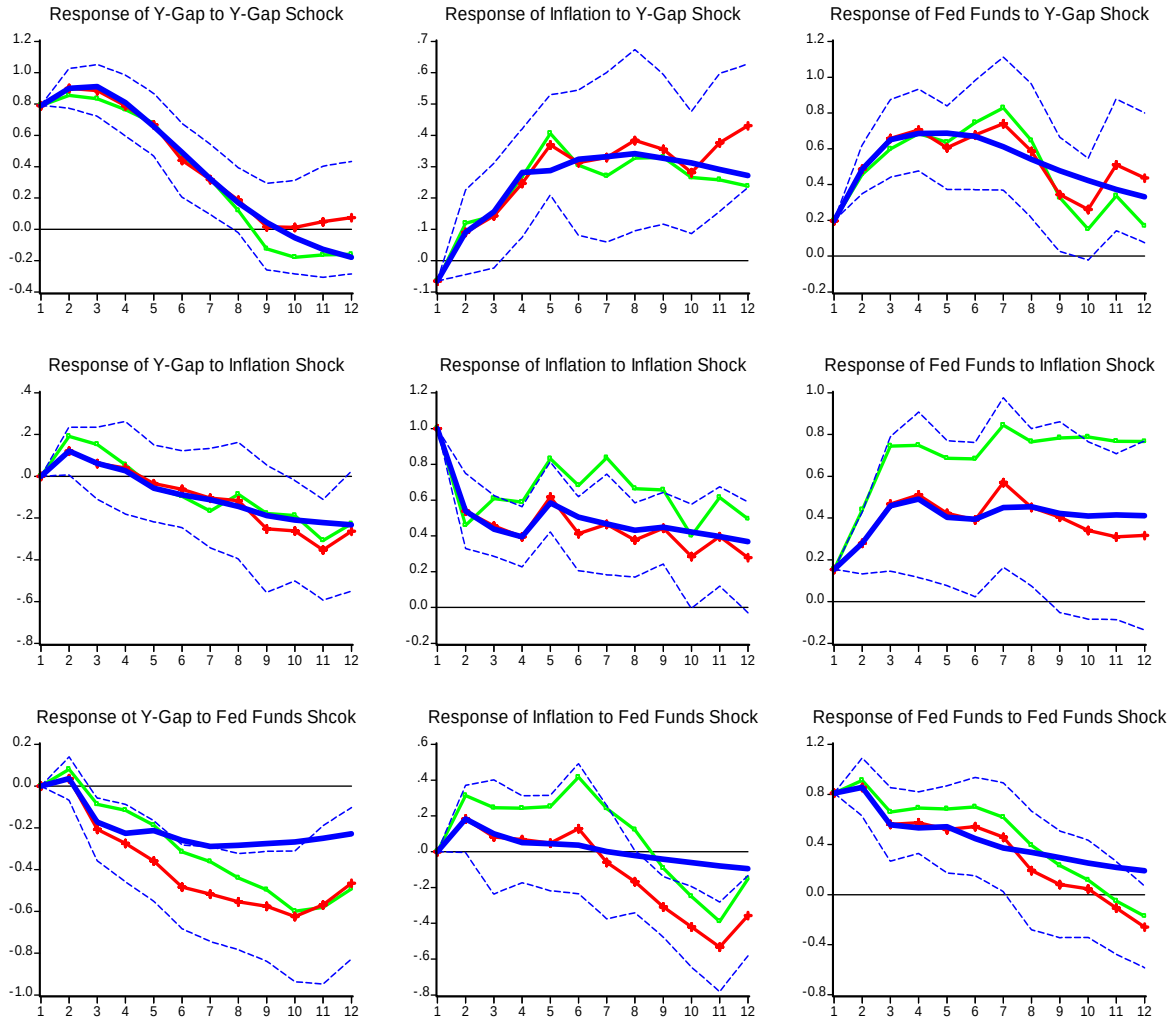
The thick-solid lines describe the true impulse response in the SVAR-GARCH model with and without GARCH effects (i.e. $B = 0_3$), the less variable referring to the latter. The thin-dashed lines are the responses from a VAR(1) and from local-linear projections. The thick-dashed line with crosses is the local-cubic projection whereas the thick-dashed line with squares is the impulse response from the Bayesian VAR.

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



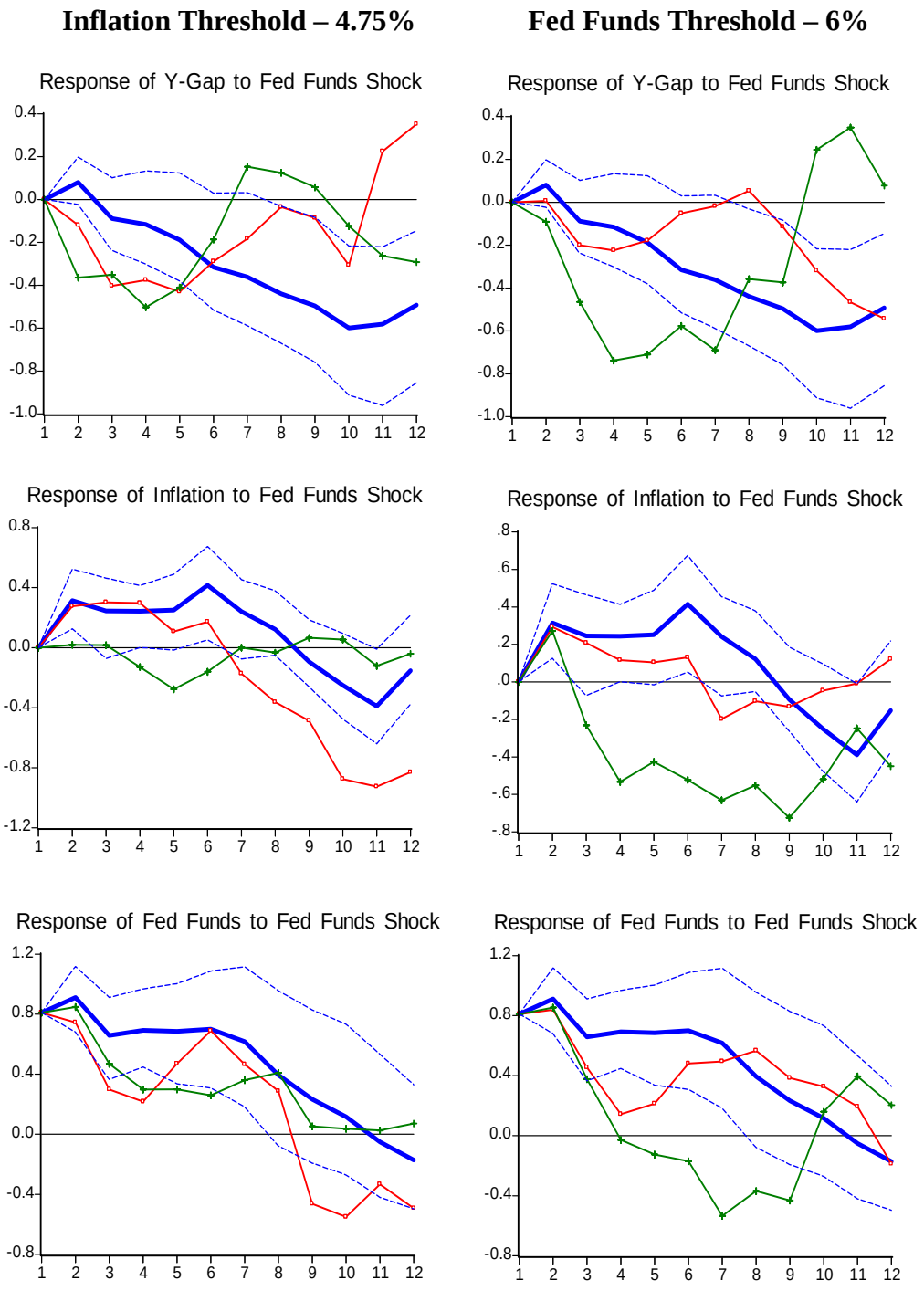
Notes: All variables in annual percentage rates. Shaded areas indicate NBER-dated recessions. Output gap is defined as the percentage difference between real GDP and potential GDP (Congressional Budget Office); Inflation is defined as the percentage change in the GDP, chain-weighted price index at annual rate; and the federal funds rate is the quarterly average of daily rates, in annual percentage rate. The solid horizontal lines display the thresholds detected by Hansen's (2000) test for Inflation and the federal funds rate.

Figure 5 – Impulse Responses for the New Keynesian Model based on a VAR, and linear and cubic projections.



Notes: The thick line is the response calculated from a VAR; the solid line with crosses is the response calculated by linear projection; the two dashed lines are 95% confidence level error bands for the individual coefficients of the linear projection response; and the solid line with circles is the response calculated by cubic projection evaluated at the sample mean. All responses calculated with four lags.

Figure 6 – Impulse Responses from New Keynesian Model. Cubic projections with threshold effects in inflation at 4.75% versus the federal funds rate at 6%



Notes: The thick line is the response calculated by cubic projection at the sample mean; the dashed lines are two standard error bands for the individual coefficients of the cubic response; the solid line with crosses is the response by cubic projection below the threshold; and the solid line with circles is the response by cubic projection above the threshold.