

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.

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

*I thank Colin Cameron, Jim Hamilton, Kevin Hoover, Aaron Smith and the participants of the Econometrics brown-bag at U.C. Davis for many useful discussions.

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). VARs afford a natural decomposition of the economy into systematic responses and random sources of variation. Since then, researchers in macroeconomics often compute dynamic multipliers of interest (such as impulse responses and forecast error variance decompositions) by specifying a VAR as well. However, often times the main object of interest is not the VAR itself as much as the implied impulse responses, which are routinely employed as a foil to the dynamic features of a particular theoretical macroeconomic model. This method of computing impulse responses imposes a number of constraints that are seldom recognized, in particular¹ : (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 conditional history beyond the experimental shock; 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. Thus, a typical monetary VAR predicts interest rates will drop in response to a deflationary shock, even if current interest rates are already at the zero bound, for example.

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 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

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

misspecification; standard error calculation is direct; 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 a large variety 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 – non-linear forecasts beyond one-period ahead often 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, which suggests that a local approximation at each time horizon would be more desirable. The methods proposed here are inspired by the ideas discussed in Cox (1961) and Tsay (1993). For example, Lin and Tsay (1996) denominate these methods “adaptive forecasting,” and use local linear projections in cointegrated VARs to evaluate the forecasting advantages of imposing cointegrating restrictions. Other authors, such as Clements and Hendry (1998) and references therein, refer to these methods as “dynamic estimation” and provide conditions under which local projections improve forecasting performance.

An advantage of calculating impulse responses by local projections is that the forecasting accuracy increases relative to a wide class of model misspecification, as the forecast horizon increases. 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 ad-

vantage 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.

2 Model-Free Estimation of Impulse Responses

2.1 Motivation

Sims (1980) introduced the seminal ideas behind the definition of an impulse response function in the context of a linear, multivariate, Markov model – a VAR. However, when more general, nonlinear, alternative specifications are considered, this definition is by no means universal (e.g. Potter, 2000 for a useful discussion). The definition that I adopt in this paper is that found in Hamilton (1994) and Koop et al. (1996) and is given by

$$\left. \frac{\Delta \mathbf{y}_{t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t} = \varepsilon_i^e} = E(\mathbf{y}_{t+s} | \varepsilon_{i,t} = \varepsilon_i^e; X_{t-1}) - E(\mathbf{y}_{t+s} | \varepsilon_{i,t} = 0; X_{t-1}) \quad s = 1, 2, \dots, h \quad (1)$$

where \mathbf{y}_t is an $n \times 1$ random vector; $X_{t-1} \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots, \mathbf{y}_{t-p})'$; and $\varepsilon_{i,t} = \varepsilon_i^e$ defines an experimental disturbance associated with the i^{th} variable a time t that consists on perturbing \mathbf{y}_{t-1} by an amount ε_i^e , which is of dimension $n \times 1$. The operator $E(\cdot)$ denotes the best, mean squared error predictor.

When the process \mathbf{y}_t is a mean zero, linear process of the form

$$\mathbf{y}_t = \sum_{j=0}^{\infty} \Psi_j \mathbf{u}_{t-j} \quad \{\mathbf{u}_t\} \sim IID(0, \Omega_u) \quad (2)$$

with $\sum_{j=0}^{\infty} \|\Psi_j\| < \infty$ (i.e. absolute summability) then the best mean square error predictor $E(\mathbf{y}_{t+s}|X_{t-1})$ and the best linear predictor, say $E_L(\mathbf{y}_{t+s}|X_{t-1})$, are identical (see e.g. Brockwell and Davis, 1991). Expression (2) will be immediately recognized as an informal expression of the multivariate Wold decomposition theorem. In this case, the \mathbf{u}_t are the forecast errors $\mathbf{u}_t = \mathbf{y}_t - E_L(\mathbf{y}_t|X_{t-1})$ and it follows immediately that,

$$\left. \frac{\Delta \mathbf{y}_{t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t} = \varepsilon_i^e} = \Psi_s \varepsilon_i^e \quad (3)$$

which can be easily calculated by inverting a VAR(p) into its infinite moving average representation² (see Hamilton, 1994) – the most popular method of computing impulse responses in practice.

ε_i^e describes the experimental design of interest. A meaningful economic experiment requires that the contemporaneous “structural” correlations in \mathbf{y}_t be taken into account. As an example, a common method of finding a sensible experiment ε_i^e is to hypothesize a Wold-causal order of the elements in \mathbf{y}_t and then to apply a triangular factorization to the original residual variance-covariance matrix³, so that $E(\mathbf{u}_t \mathbf{u}_t') = \Omega_u = PP'$ and $\varepsilon_i^e = p_i^{-1}$, with p_i^{-1} indicating the i^{th} column of P^{-1} . In general, and because this method does not uniquely identify the contemporaneous structure in \mathbf{y}_t , the choice of experiment ε_i^e will depend on each researcher’s identification strategies⁴. The remainder of the paper thus proceeds by taking ε_i^e as a pre-selected experiment of interest.

Expression (1) is a natural definition for an impulse response. It asks the question: What is

² I deliberately maintain the notation ε_t and \mathbf{u}_t separate because, although the terms coincide in this example, they do not in general.

³ This particular choice of normalization not only ensures orthogonality among the innovations but also normalizes the size of the experimental disturbance to be of unit variance.

⁴ To my knowledge, Swanson and Granger 1997, and Demiralp and Hoover (2003) provide the only statistical-based methods to identify the true structure.

the best guess about the time profile of \mathbf{y}_t into the future when the system is perturbed from its current state, X_{t-1} , by an amount ε_i^e ? To find an answer to this question, it is common to estimate a VAR from which to construct the forecast difference in expression (1), which coincides with (3) if the data is well represented by it.

In this paper I propose a different strategy altogether. Estimating a specific model (be it a VAR or a more complicated alternative) to compute the impulse responses is akin to using a Taylor series expansion around one point to extrapolate the function at an increasingly distant range of values: the risk that the approximation will be poor (and even misleading), increases with the distance from the initial evaluation point. Similarly, a VAR is optimized to produce the best, linear, one-step ahead forecasts. If the data do not conform to this DGP (even by simple misspecification of the lag length) we may get reasonable one-step ahead forecasts but the quality of the forecasts at increasingly distant horizons, will decline steadily.

2.2 Local Projections for Impulse Responses: Estimation

All statistical models are approximations, hence, it seems more sensible that to calculate impulse responses, we construct approximations at each horizon $s = 1, 2, \dots, h$ so that parameter estimation is directly linked to the object of data analysis (see Granger, 1993). In expression (1), this object consists on the best, mean squared error projection for \mathbf{y}_{t+s} , given information up to time t . A natural way to accomplish this objective is by finding the local-linear, orthogonal projection of \mathbf{y}_{t+s} on to the space generated by $X_t \equiv (y_t, \dots, y_{t-p+1})'$ with the least squares regressions,

$$\mathbf{y}_{t+s} = B_0^{(s)} + B_1^{(s)}\mathbf{y}_t + B_2^{(s)}\mathbf{y}_{t-1} + \dots + B_p^{(s)}\mathbf{y}_{t-p+1} + \mathbf{u}_{t+s}^{(s)} \quad s = 1, 2, \dots, h \quad (4)$$

where the $B_i^{(s)}$ denote the matrix of coefficients for lag $(i - 1)$ and the regression for the s^{th} horizon, so that the impulse response is simply

$$\left. \frac{\Delta \mathbf{y}_{t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t} = \varepsilon_i^e} = \widehat{B}_1^{(s)} \varepsilon_i^e \quad s = 1, 2, \dots, h \quad (5)$$

The parameters $\widehat{B}_1^{(s)}$ are consistently estimated from simple least squares since, although the residuals $\mathbf{u}_{t+s}^{(s)}$ will not be white noise in general, they will be otherwise uncorrelated with information dated t and beyond. The unknown form of the dependence is determined by the specific DGP and can only be determined for specific cases. In the next section, I show the form that this dependence takes when the DGP is a VAR(p). Expression (4) will be quickly recognized as the “adaptive forecasts” in Lin and Tsay (1996) or the “dynamic forecasts” in Clements and Hendry (1998).

Although expression (4) describes a system of n equations, they can be estimated equation by equation as univariate regressions. For example, the response of $y_{j,t+s}$ to the experiment $\varepsilon_i^e = (\varepsilon_{i,1}^e, \varepsilon_{i,2}^e, \dots, \varepsilon_{i,j}^e, \dots, \varepsilon_{i,n}^e)'$ can be estimated by least squares on

$$\begin{aligned} y_{j,t+s} &= \beta_{0j}^{(s)} + \beta_{11}^{(s)} y_{1,t} + \beta_{12}^{(s)} y_{2,t} + \dots + \beta_{1j}^{(s)} y_{j,t} + \dots + \beta_{1n}^{(s)} y_{n,t} + \\ &\quad B_2^{(s)} \mathbf{y}_{t-1} + \dots + B_p^{(s)} \mathbf{y}_{t-p+1} + u_{j,t+s}^{(s)} \quad s = 1, 2, \dots, h \end{aligned} \quad (6)$$

from which the estimated coefficients can be combined with ε_i^e to produce,

$$\left. \frac{\Delta \widehat{y}_{j,t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t} = \varepsilon_i^e} = \widehat{\beta}_{11}^{(s)} \varepsilon_{i,1}^e + \dots + \widehat{\beta}_{1j}^{(s)} \varepsilon_{i,j}^e + \dots + \widehat{\beta}_{1n}^{(s)} \varepsilon_{i,n}^e \quad s = 1, 2, \dots, h.$$

Estimating impulse responses by linear projection methods means that we lose the initial $s + p$ observations. In situations where degrees of freedom are at a premium, it may be convenient to restrict the dynamics in expression (6) to contain only lags of the dependent variable, specifically,

$$\begin{aligned} y_{j,t+s} &= \beta_{0j}^{(s)} + \beta_{11}^{(s)} y_{1,t} + \beta_{12}^{(s)} y_{2,t} + \dots + \beta_{1j}^{(s)} y_{j,t} + \dots + \beta_{1n}^{(s)} y_{n,t} + \\ &\quad \beta_{2j}^{(s)} y_{j,t-1} + \dots + \beta_{pj}^{(s)} y_{j,t-p+1} + v_{j,t+s}^{(s)} \quad s = 1, 2, \dots, h. \end{aligned} \quad (7)$$

Correspondingly,

$$\left. \frac{\Delta \tilde{y}_{j,t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t}=\varepsilon_i^e} = \tilde{\beta}_1^{(s)} \varepsilon_{i,1}^e + \dots + \tilde{\beta}_j^{(s)} \varepsilon_{i,j}^e + \dots + \tilde{\beta}_n^{(s)} \varepsilon_{i,n}^e \quad s = 1, 2, \dots, h.$$

However, since (7) omits the lags of the remaining variables in the system, the consistency of the $\tilde{\beta}_i^{(s)}$ is no longer guaranteed and its practical usefulness can only be elucidated by Monte Carlo experimentation. Section 5 reports these experiments.

The local projections described in expressions (4)-(7) are also useful in calculating other dynamic multipliers of interest, such as variance decompositions of the forecast error variances. Thus, the contribution of the i^{th} orthogonalized innovation to the mean squared error (MSE) of the s -period ahead forecast is (according to Hamilton, 1994):

$$MSE(E(\mathbf{y}_{t+s}|\varepsilon_{i,t} = \varepsilon_i^e; X_{t-1})) = \sum_{i=1}^n \left[\begin{aligned} &(\varepsilon_i^e)^{-1} (\varepsilon_i^e)^{-1'} + B_1^{(1)} (\varepsilon_i^e)^{-1} (\varepsilon_i^e)^{-1'} B_1^{(1)'} + \dots \\ &+ B_1^{(s-1)} (\varepsilon_i^e)^{-1} (\varepsilon_i^e)^{-1'} B_1^{(s-1)'} \end{aligned} \right]$$

where the $B_1^{(i)}$ for $i = 1, \dots, s-1$ can be replaced with the estimates $\hat{B}_1^{(i)}$ for $i = 1, \dots, s-1$ in expression (4), for example.

Expression (4) was used in Lin and Tsay (1996) to compute s -step ahead forecasts of cointegrated VARs and found to perform well in six out of seven data sets analyzed for one- to ten-steps ahead forecasts. In fact, this method achieved a 60% reduction in root mean squared error in short-term forecasts for U.S. interest rates, which are commonly used in monetary VARs. However, Phillips (1998) shows that it is advisable to impose the cointegrating restrictions (rather than using the unrestricted VAR in the levels) to avoid inconsistencies in the long-horizon values of the impulse responses. Cointegration is a restriction that can be easily accommodated when calculating impulse responses by local projection methods. Let $\mathbf{z}_t = A' \mathbf{y}_t$ denote the cointegrating vectors, then expression (4) can be appropriately re-specified as

$$\Delta \mathbf{y}_{t+s} = C_0^{(s)} + C_c^{(s)} \mathbf{z}_t + C_1^{(s)} \Delta \mathbf{y}_t + \quad (8)$$

$$C_2^{(s)} \Delta \mathbf{y}_{t-1} + \dots + C_p^{(s)} \Delta \mathbf{y}_{t-p+1} + \mathbf{u}_{t+s}^{(s)} \quad s = 1, 2, \dots, h$$

Calculating impulse responses by local projection methods is rather straight-forward: in its most basic implementation, it only requires simple least squares regression. This suggests that, unlike when impulse responses are calculated from pre-specified models, we can easily increase the quality of the local projections by using more flexible methods. Furthermore, since the value and the properties of the impulse response at time s are contained in the terms associated with \mathbf{y}_t only (that is, the $\widehat{B}_1^{(s)}$ in expression (4)), a parsimonious way of improving the local projection is to concentrate the flexibility on those terms only. The universe of options for making the approximation more flexible is limited only by the imagination of each practitioner. Hence, to keep the exposition intentionally uncomplicated, consider enriching the quality of the local approximation in expression (4) with a cubic polynomial as follows,

$$\begin{aligned} \mathbf{y}_{t+s} = & B_0^{(s)} + B_1^{(s)} \mathbf{y}_t + Q_1^{(s)} \mathbf{y}_t^2 + C_1^{(s)} \mathbf{y}_t^3 + \\ & B_2^{(s)} \mathbf{y}_{t-1} + \dots + B_p^{(s)} \mathbf{y}_{t-p+1} + \mathbf{u}_{t+s}^{(s)} \quad s = 1, 2, \dots, h \end{aligned} \quad (9)$$

To simplify the exposition, I do not allow for cross-product terms so that $\mathbf{y}_t^2 = (\mathbf{y}_{1t}^2, \mathbf{y}_{2t}^2, \dots, \mathbf{y}_{nt}^2)'$, for example. It is readily apparent that the impulse response at time s now becomes,

$$\begin{aligned} \left. \frac{\partial \widehat{\mathbf{y}}_{t+s}}{\partial \varepsilon_{i,t}} \right|_{\varepsilon_{i,t} = \varepsilon_i^e} &= \left\{ \widehat{B}_1^{(s)} (\mathbf{y}_{t-1} + \varepsilon_i^e) + \widehat{Q}_1^{(s)} (\mathbf{y}_{t-1} + \varepsilon_i^e)^2 + \widehat{C}_1^{(s)} (\mathbf{y}_{t-1} + \varepsilon_i^e)^3 \right\} - \\ &\quad \left\{ \widehat{B}_1^{(s)} (\mathbf{y}_{t-1}) + \widehat{Q}_1^{(s)} (\mathbf{y}_{t-1})^2 + \widehat{C}_1^{(s)} (\mathbf{y}_{t-1})^3 \right\} \\ &= \left\{ \widehat{B}_1^{(s)} (\varepsilon_i^e) + \widehat{Q}_1^{(s)} (2\mathbf{y}_{t-1}\varepsilon_i^e + \varepsilon_i^{e^2}) + \widehat{C}_1^{(s)} (3\mathbf{y}_{t-1}^2\varepsilon_i^e + 3\mathbf{y}_{t-1}\varepsilon_i^{e^2} + \varepsilon_i^{e^3}) \right\} \end{aligned} \quad (10)$$

It is important to notice that, unlike impulse responses estimated with local-linear projections, the responses in expression (10) depend on where the impulse response is evaluated through the term

y_{t-1} . This is a feature that is shared by any nonlinear, flexible projection method. Despite the obvious gains in flexibility, it is not particularly difficult to estimate the responses in expression (10), which can still be done with a simple OLS routine. Furthermore, when the evaluation point is set to the sample mean, i.e. $\mathbf{y}_{t-1} = \bar{\mathbf{y}}_{t-1}$, then the impulse responses are evaluated at the same point as the local-linear projection in (4) and those from a traditional VAR. Numerically, the values of all three methods are identical for large samples when the true model is a VAR. Section 5 below illustrates with Monte Carlo experimentation some of the advantages of the local-cubic projection (9) in the context of a nonlinear model and the relevance of the evaluation point.

Estimating impulse responses by local approximation suggests that any parametric, semi-parametric and non-parametric approximation technique can be used. For example, rather than using a polynomial approximation to the conditional mean, as is done in expression (9), we could have used Hamilton's (2001) parametric, flexible nonlinear model, a flexible discrete-Fourier form (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, because the impulse responses can be calculated on the basis of 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, 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.

3 Relation to VAR-based Impulse Responses

This section establishes the correspondence between traditional VAR-based impulse response function analysis and impulse responses calculated from local projections. In a VAR, the $n \times 1$ vector \mathbf{y}_t depends linearly on its p -lags $X_{t-1} \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots, \mathbf{y}_{t-p})'$, through the expression

$$\mathbf{y}_t = \boldsymbol{\mu} + \Pi' X_{t-1} + \boldsymbol{\varepsilon}_t \quad (11)$$

where $\boldsymbol{\varepsilon}_t$ is an *i.i.d.* vector of disturbances and $\Pi' \equiv [\Pi_1 \ \Pi_2 \ \dots \ \Pi_p]$. Consider now the VAR(1) companion form to this VAR(p) by defining

$$W_t \equiv \begin{bmatrix} \mathbf{y}_t - \boldsymbol{\mu} \\ \mathbf{y}_{t-1} - \boldsymbol{\mu} \\ \vdots \\ \mathbf{y}_{t-p+1} - \boldsymbol{\mu} \end{bmatrix}; F \equiv \begin{bmatrix} \Pi_1 & \Pi_2 & \dots & \Pi_{p-1} & \Pi_p \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix}; \mathbf{v}_t \equiv \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (12)$$

and then realizing that according to (11) and (12),

$$W_t = FW_{t-1} + \mathbf{v}_t \quad (13)$$

Expression (13) simplifies the calculation of s -step ahead forecasts considerably since $E(W_{t+s}|W_t) = F^s W_t$, and therefore,

$$E(\mathbf{y}_{t+s}|X_t) = \boldsymbol{\mu} + F_{11}^{(s)}(\mathbf{y}_t - \boldsymbol{\mu}) + F_{12}^{(s)}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \dots + F_{1p}^{(s)}(\mathbf{y}_{t-p+1} - \boldsymbol{\mu}) \quad (14)$$

where $F_{11}^{(s)}$ denotes the upper left block of F^s , which is the matrix F raised to the s^{th} power. $F_{11}^{(s)}$ indicates rows 1 through n and columns 1 through n of the $(np \times np)$ matrix F^s , $F_{12}^{(s)}$ indicates rows 1 through n and columns $(n+1)$ through $2n$ of F^s , and so on. From the definition (1) and expression (14), it is immediately apparent that

$$\left. \frac{\Delta \mathbf{y}_{t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t} = \varepsilon_i^e} = F_{11}^{(s)} \varepsilon_i^e \quad s = 1, 2, \dots, h. \quad (15)$$

In fact, under the maintained assumption that \mathbf{y}_t is a p^{th} order VAR, then

$$\mathbf{y}_{t+s} - \boldsymbol{\mu} = \boldsymbol{\varepsilon}_{t+s} + F_{11}^{(1)} \boldsymbol{\varepsilon}_{t+s-1} + \dots + F_{11}^{(s-1)} \boldsymbol{\varepsilon}_{t+1} + F_{11}^{(s)} (\mathbf{y}_t - \boldsymbol{\mu}) + \dots + F_{1p}^{(s)} (\mathbf{y}_{t-p+1} - \boldsymbol{\mu}) \quad (16)$$

where the $F_{11}^{(s)}$ for $s = 1, 2, \dots, h$ clearly correspond to the infinite moving average representation of the VAR(p), as long as the eigenvalues of F lie inside the unit circle so that $F^s \rightarrow 0$ as $s \rightarrow \infty$. Expression (15) corresponds to expression (5) and expression (16) naturally suggests that the terms $F_{11}^{(s)}$ could be estimated directly from the following regressions,

$$\mathbf{y}_{t+s} = \boldsymbol{\mu} + F_{11}^{(s)} (\mathbf{y}_t - \boldsymbol{\mu}) + \dots + F_{1p}^{(s)} (\mathbf{y}_{t-p+1} - \boldsymbol{\mu}) + \mathbf{u}_{t+s} \quad s = 1, 2, \dots, h \quad (17)$$

which correspond to the expression for the local-linear projection (4). The error terms in (17) will be autocorrelated since,

$$\mathbf{u}_{t+s} = \boldsymbol{\varepsilon}_{t+s} + F_{11}^{(1)} \boldsymbol{\varepsilon}_{t+s-1} + \dots + F_{11}^{(s-1)} \boldsymbol{\varepsilon}_{t+1}. \quad (18)$$

thus suggesting that in calculating standard errors for the coefficients $\hat{B}_1^{(s)}$ in (4), heteroskedasticity and autocorrelation consistent (HAC) residual variance estimates are advised. This aspect is explored in more detail in the next section.

The more direct and customary way of calculating the terms $F_{11}^{(s)}$ for $s = 1, 2, \dots, h$ is based on the VAR(p) estimates of Π and the following recursion (see Hamilton, 1994):

$$\begin{aligned} F_{11}^{(1)} &= \Pi_1 \\ F_{11}^{(2)} &= \Pi_1 F_{11}^{(1)} + \Pi_2 \\ &\vdots \\ F_{11}^{(s)} &= \Pi_1 F_{11}^{(s-1)} + \Pi_2 F_{11}^{(s-2)} + \dots + \Pi_p F_{11}^{(s-p)} \end{aligned} \quad (19)$$

The assumption that the primitive econometric model generating the impulse responses is the VAR given by (11) has several important implications. Equation (15) makes clear that at every horizon s , the impulse response is linear in ε_i^e . Consequently, for all s , impulse responses will be *symmetric* to sign changes in ε_i^e , that is,

$$\left. \frac{\Delta \mathbf{y}_{t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t}=\varepsilon_i^e} = F_{11}^{(s)} \varepsilon_i^e = - \left. \frac{\Delta \mathbf{y}_{t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t}=-\varepsilon_i^e} = - \left(F_{11}^{(s)} (-\varepsilon_i^e) \right) \quad (20)$$

In addition, the magnitude of the shock does not change the general shape of the response since,

$$\left. \frac{\Delta \mathbf{y}_{t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t}=\varepsilon_i^e} = F_{11}^{(s)} \varepsilon_i^e = \frac{1}{k} \left. \frac{\Delta \mathbf{y}_{t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t}=k\varepsilon_i^e} = \frac{1}{k} F_{11}^{(s)} (k\varepsilon_i^e) \quad (21)$$

Furthermore, expression (15) clearly demonstrates that the local history prior to the date of experimentation is irrelevant since

$$\left. \frac{\Delta \mathbf{y}_{t+s}}{\Delta \varepsilon_{i,t}} \right|_{\varepsilon_{i,t}=\varepsilon_i^e; X_{t-1}} = \left. \frac{\Delta \mathbf{y}_{r+s}}{\Delta \varepsilon_{i,r}} \right|_{\varepsilon_{i,r}=\varepsilon_i^e; X_{r-1}} = F_{11}^{(s)} \varepsilon_i^e \quad \text{for all } t \text{ and } r. \quad (22)$$

Finally, expression (19) shows that the estimated impulse response functions will be nonlinear, high-dimensional functions of the VAR estimated coefficient matrices, Π_j , $j = 1, 2, \dots, p$ so that the $(n \times n)$ matrix $F_{11}^{(s)}$ will be a function of kn^2 coefficients for $k = \min(s, p)$. As an illustration, any $F_{11}^{(s)}$ for $s \geq 12$ calculated on a six variable, twelfth order VAR, will be a function of 432 coefficients!

As the time horizon of the impulse response increases, the assumption that the VAR is correctly specified and properly describes the data, is increasingly critical. The recursive formulas in expression (19) show that misspecifications are compounded steadily with the impulse response horizon. As a simple example, consider a VAR(2) which is incorrectly specified as a VAR(1), the impulse responses calculated under both scenarios are:

Impulse	VAR(1)	VAR(2)
$F_{11}^{(1)}$	$\tilde{\Pi}_1$	Π_1
$F_{11}^{(2)}$	$\tilde{\Pi}_1^2$	$\Pi_1^2 + \Pi_2$
$F_{11}^{(3)}$	$\tilde{\Pi}_1^3$	$\Pi_1^2 + \Pi_2\Pi_1 + \Pi_2$
$F_{11}^{(4)}$	$\tilde{\Pi}_1^4$	$\Pi_1^3 + 2\Pi_2\Pi_1^2 + \Pi_2\Pi_1 + \Pi_2$
\vdots	\vdots	\vdots

(23)

where $\tilde{\Pi}_1 = \Pi_1 + \Pi_2\Gamma_1\Gamma_0^{-1}$ and Γ_j is the j^{th} autocovariance of y_t . Depending on Π_2 , the bias in the impulse responses can be quite substantial when we move beyond the first few horizons. If the system is stationary, so that $F_{11}^{(\infty)} \rightarrow 0$, then the bias will disappear at very long horizons since all variables will return to their unconditional mean values. Thus, the persistence of the data will be an important factor affecting the severity of possible misspecification, aside from more general forms of misspecification due to nonlinearities, for example.

Impulse response functions estimated by local projections have several advantages. First, the local projection for a particular class (linear, cubic, or more complex specifications) is the best mean squared error predictor at each horizon s . By contrast, a multivariate dynamic specification is optimized for one-period ahead forecasting only. If the specification of this system is correct, then its impulse responses will be consistently and more efficiently estimated than those from local projections. As the recursion in expression (19) makes clear in the case of a VAR, impulse responses calculated from a given model impose numerous, cross-horizon restrictions. These restrictions provide efficiency gains relative to the local projection counterparts. However, when the model is misspecified, such cross-horizon restrictions are invalid and deliver inconsistent impulse responses, specially when the data are persistent or non-stationary.

Flexible local projections, such as the local-cubic projection in (9), 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 (10). 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, as we will see shortly, its error bands are easily computed

4 Local Projections for Impulse Responses: Inference

Although impulse responses can be easily calculated from an estimated VAR, their standard-error bands are not. Analytical computation of these bands requires an approximation based on the delta method. However, as expression (19) shows, the mapping between the Π_j and the $F_{11}^{(j)}$ becomes increasingly nonlinear as the horizon j increases and hence, the quality of the delta method approximation deteriorates steadily. Sims and Zha (1999) discuss another undesirable feature of traditional error band computation: when $F^s \rightarrow 0$ as $s \rightarrow \infty$ (so that the VAR is stationary), the error bands will shrink as a function of the smallest eigenvalue of F , call it λ , at a rate λ^{-s} . This suggests that uncertainty dissipates as we move further into the future, a counterintuitive result. Alternative, non-analytic methods for error band computation include numerically intensive Monte Carlo and bootstrap methods and the Bayesian methods developed in Sims and Zha (1999).

The equivalence between the blocks in (25) and expression (4) suggests that standard errors should be calculated with heteroskedasticity and autocorrelation consistent (HAC) methods. When the DGP is a VAR, we have seen that the error terms $\mathbf{u}_{t+s}^{(s)}$ in the local projection regressions have moving average components whose order is a function of s , the horizon of the impulse response under consideration. In general, although we will not know whether or not the DGP is a VAR, it is reasonable to suspect that the error terms $\mathbf{u}_{t+s}^{(s)}$ in expression (4) will have some

form of dynamic dependence that is contingent on the intervening periods s . Thus, it is advisable to use a HAC, variance-covariance estimator of the residual variance, available in most standard regression packages. Whether we are using local linear projections based on expressions (4)-(8) or more flexible approximations, such as the cubic projection of expression (9), notice that there is a one-to-one correspondence between the error bands and the standard errors of the coefficients in these expressions.

Specifically, let $\widehat{\Sigma}_L$ denote the estimated HAC, variance-covariance matrix of the coefficients $\widehat{B}_1^{(s)}$ in expression (4), for example. Then a 95% confidence interval for the impulse response at time s can be constructed approximately as $1.96 \pm \left(\varepsilon_i^{e'} \widehat{\Sigma}_L \varepsilon_i^e \right)$. Similarly, the 95% confidence interval for the cubic approximation in expression (9) can be calculated by defining the scaling $\gamma_i^e \equiv (\varepsilon^e, \quad 2\mathbf{y}_{t-1}\varepsilon_i^e + \varepsilon_i^{e^2}, \quad 3\mathbf{y}_{t-1}^2\varepsilon_i^e + 3\mathbf{y}_{t-1}\varepsilon_i^{e^2} + \varepsilon_i^{e^3})'$, 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 $\widehat{B}_{11}^{(s)}$, $\widehat{Q}_{11}^{(s)}$, and $\widehat{C}_{11}^{(s)}$ in (9), a 95% confidence interval for the impulse response at time s is approximately, $1.96 \pm \left(\gamma_i^{e'} \widehat{\Sigma}_C \gamma_i^e \right)$. Monte Carlo experiments reported below show that the efficiency losses associated with estimating error bands with Newey-West estimates on expression (4) relative to Monte Carlo error bands from the true VAR(p) are rather minor. Calculation of error bands for impulse responses has long been a controversial issue (see Kilian, 1998 and Sims and Zha, 1999) for several reasons but perhaps most notably because VAR-based impulse responses are nonlinear functions of estimated coefficients. By contrast, local projections are direct estimates of the impulse response coefficients, and therefore, less susceptible to these considerations.

4.1 Relation to VAR-based Impulse Responses: Efficiency

As before, assume here that the DGP is the VAR(p) in expression (11). The derivations in section 3 show that,

$$\begin{aligned} \mathbf{y}_{t+s} &= \boldsymbol{\mu} + F_{11}^{(s)}(\mathbf{y}_t - \boldsymbol{\mu}) + \dots + F_{1p}^{(s)}(\mathbf{y}_{t-p+1} - \boldsymbol{\mu}) + \\ \boldsymbol{\varepsilon}_{t+s} &+ F_{11}^{(1)}\boldsymbol{\varepsilon}_{t+s-1} + \dots + F_{11}^{(s-1)}\boldsymbol{\varepsilon}_{t+1} \quad s = 1, 2, \dots, h \end{aligned} \quad (24)$$

Thus, consider estimating impulse responses for $s = 1, 2, \dots, h$ periods and that instead of using the usual recursions in (19), we estimate the following system of equations. Let $Y_t \equiv (\mathbf{y}_{t+1}, \dots, \mathbf{y}_{t+h})$, $\epsilon_t \equiv (\boldsymbol{\varepsilon}_{t+1}, \dots, \boldsymbol{\varepsilon}_{t+h})$, and $X_t \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots, \mathbf{y}_{t-p})$, then

$$Y_t = X_t \Psi + \epsilon_t \Phi \quad (25)$$

where

$$\Psi = \begin{bmatrix} F_{11}^{(1)} & F_{11}^{(2)} & \dots & F_{11}^{(h)} \\ F_{12}^{(1)} & F_{12}^{(2)} & \dots & F_{12}^{(h)} \\ \vdots & \vdots & \dots & \vdots \\ F_{1p}^{(1)} & F_{1p}^{(2)} & \dots & F_{1p}^{(h)} \end{bmatrix}; \Phi = \begin{bmatrix} I_n & F_{11}^{(1)} & \dots & F_{11}^{(h)} \\ 0 & I_n & \dots & F_{11}^{(h-1)} \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & I_n \end{bmatrix}$$

and given that $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') = \Omega_\varepsilon$, then $E(\epsilon_t \epsilon_t') = \Phi (I_h \otimes \Omega_\varepsilon) \Phi' \equiv \Sigma$. Thus, maximum likelihood estimation of the system in expression (25) can be accomplished by standard GLS formulas according to,

$$vec(\hat{\Psi}) = [(I \otimes X)' \Sigma^{-1} (I \otimes X)]^{-1} (I \otimes X)' \Sigma^{-1} vec(Y) \quad (26)$$

The usual impulse responses are given by rows 1 through n and columns 1 through (nh) of $\hat{\Psi}$ and standard errors could be computed directly from the regression output. The special structure of the variance-covariance matrix Σ allows this system to be estimated by GLS, block by block. Expression (25) is equivalent to stacking the local linear projections (4) for periods $s = 1, \dots, h$ and by using knowledge that the DGP is a VAR to construct the specific structure of the variance covariance matrix. In general, the true DGP is unknown, and although local projections can be

estimated as a system by S.U.R. methods, the Monte Carlo experiments in the next section suggest there are few advantages to doing so. In the interest of clarity, I will report my experiments from univariate estimates of local projections aimed at highlighting simplicity and accuracy rather than sophistication.

5 Monte Carlo Evidence

This section discusses three 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 is based on Phillips (1998) and shows that local projections can reduce the inconsistency of impulse responses at long horizons when there is cointegration. The third experiment simulates a SVAR-GARCH (see Jordà and Salyer, in press) to show that local projections do a reasonable job at approximating the inherent nonlinearities of this model.

5.1 Evans and Marshall (1998)

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’s⁵ 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 in figure 1. Each panel in figure 1 displays the impulse response of a variable in the VAR to a shock in the variable FF ,⁷ calculated as follows: the blue, thick-solid line is the true VAR(12) impulse response with two standard-error bands displayed in red, 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 black line with squares; the responses from the local linear approximation are displayed by the green, dashed line; and the responses from the cubic local approximation are displayed by the purple line with circles.

Several results deserve comment. The VAR(2) responses often fall within the two standard-error bands for the true response and have the same general shape. 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 rate⁸ over 23 out of the 24 periods displayed. In contrast, the linear local projection is virtually within the true two standard error bands throughout the 24 periods depicted, and is strictly negative for the last 7 periods.

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.

⁸ 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 $PCOM$.

The response of P therefore offers a good opportunity to experiment with the alternative local linear specification proposed in expression (7). This response contains only one lag of all the variables (instead of the two lags we have used so far) but allows for 12 lags of P , thus totaling 18 regressors (the 12th local linear projection requires 73 regressors instead). Figure 2 displays this response and compares it to the true impulse responses and responses from a VAR(2) and a VAR(6). Figure 2 shows that, while even a VAR(6) produces an impulse response that becomes negative only after 20 periods, the response based on (7) is quite close to the true response throughout. This suggests that reasonably accurate estimates of $B_1^{(s)}$ can be obtained with the dynamics of the variable being shocked alone, rather than by including lags for all the variables.

The third experiment shows that model-free methods are consistent under true specification by calculating impulse responses with local projections and 12 lags. The results are reported in figure 3, also for a shock to FF only. Thus, the blue, thick line is the true impulse response, along with two standard error bands displayed in red, thick-dashed lines. The responses based on local linear projections are displayed with the black, dashed line and the responses based on local cubic projections are displayed by the green line with circles. Generally speaking, the responses by either approximation literally lie on top of the true response⁹ with occasional minor differences that disappeared with slightly bigger samples, not reported here.

The final set of experiments evaluates error bands computed with model-free methods. In practice, we do not know the true multivariate DGP underlying the data, so we will typically choose to apply model-free methods, equation by equation. Therefore, consider the following experiments. 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 each method (the Monte Carlo standard errors for the VAR(12) give a measure of the true standard errors), and additionally calculated Newey-West¹⁰ corrected standard errors for the

⁹ This is also true for the responses to all the remaining shocks that are not reported here but are available upon request.

¹⁰ The Newey-West lag correction is selected to be equal to s , the horizon of the impulse response being considered.

local projections. Therefore, figure 4 displays the VAR(12) Monte Carlo standard error to a shock in FF with a blue, thick line, with a red, thick-dashed line for the local linear projection, and with a green, dashed line with circles for the local cubic projection. The Newey-West corrected standard errors for the local linear projection are displayed with a black line with squares and by a green line with stars for the local cubic projection.

In section 4, I argued that model-free estimates of impulse responses are less efficient than VAR based estimates when the VAR is correctly specified and it is the true model. The graphs in figure 4 confirm this statement but also show 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 (for example, notice the panels for the responses in 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 counterintuitive result, explained by Sims and Zha (1999), was already mentioned in section 4 and differs markedly from the behavior of Newey-West corrected and Monte Carlo standard errors for the local projections.

5.2 Phillips (1998)

Phillips (1998) showed that when there is cointegration, impulse responses based on an unrestricted VAR estimated in the levels are no longer asymptotically normal and can be biased at very long horizons when the sample size is small. To illustrate this phenomenon empirically, Phillips (1998) designed the following cointegrated VAR,

$$y_t = Ay_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, I_3) \quad (27)$$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 1 & -1 & 0 \end{bmatrix}$$

so that the system has one unit root and two cointegrating vectors. The original Monte Carlo is done for 1,000 replications of series of length 112 so that the last 12 observations are saved for an out-of-sample forecast evaluation. Here, I simulate 1000 series of 100, 200, and 400 observations in length. For each series-length, I then estimate impulse responses from a VAR(1), local linear, and cubic projections. The results are directly comparable with figure 4, panels (a), (c), and (d) in Phillips (1998). Figure 5 reports the responses of the variables in the VAR to a shock in the first variable, labeled y_1 . The blue, thick-solid line displays the true, theoretical impulse response. The red, dashed line is the response based on the unrestricted VAR(1), the green line with circles is the linear projection with one lag, and the black line with stars is the cubic projection instead.

The graphs corresponding to a sample size of 100 replicate the results in Phillips (1998), and are similar even with those for a series of 200 observations in length. Thus, estimates from the local linear projections quickly converge to the unrestricted VAR estimates, although there is a slight advantage to using local cubic projections¹¹ : with a sample size of 200, there is no appreciable bias for approximately the first 10 periods, and for the response of y_3 the bias remains close to zero for 25 periods when the sample size increases to 400. These results confirm Phillips' recommendation to estimate the vector error correction form rather than the unrestricted VAR, and suggest local projections be specified as in expression (8).

5.3 Impulse Responses for a GARCH-SVAR

The final Monte-Carlo experiment gauges how well model-free estimates approximate the impulse responses from a nonlinear DGP relative to estimates with a VAR. In Jordà and Salyer (in press) 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,

¹¹ Additional lags did not improve the quality of these projections.

$$\begin{aligned}
\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} &= A \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{3t-1} \end{bmatrix} + B h_{1t} + \begin{bmatrix} \sqrt{h_{1t}} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix}, \quad \varepsilon_t \sim N(0, I_3) \quad (28) \\
h_{1t} &= 0.5 + 0.3u_{1,t-1} + 0.5h_{1,t-1}; \quad u_{1t} = \sqrt{h_{1t}} \varepsilon_{1t} \\
A &= \begin{bmatrix} 0.5 & -0.25 & 0.25 \\ 0.75 & 0.25 & 0.25 \\ -0.25 & -0.25 & 0.75 \end{bmatrix}; \quad B = \begin{bmatrix} -1.75 \\ -1.5 \\ 1.75 \end{bmatrix}
\end{aligned}$$

and a sample size of 300, replicated 500 times. Notice that the GARCH-SVAR in (28) 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 6 displays the impulse responses from a shock to y_{1t} of unit size. The blue, thick-solid line describes the true impulse response in the GARCH-SVAR. The red, solid line is the impulse response when the variance effects are set to zero (i.e. $B = 0_3$). The black, dashed line with stars is the impulse response from the linear projection and the green, dashed line with circles is the response from the cubic projection. 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 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 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 cubic projection estimates instead. When the responses estimated with this approximation are evaluated around the sample mean values of \mathbf{y}_t , 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 cubic projection at $\mathbf{y}_{t-1} = \bar{\mathbf{y}}_{t-1} + 5 \times (\hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\sigma}_{33})'$. This choice allows us to match relatively well the more extreme dynamics of the model. As figure 6 shows, the match is relatively good 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 data.

6 Application: Inflation-Output Trade-offs

Pioneering work by McCallum (1983) and Taylor (1993) has inspired a substantial amount of research that investigates the efficacy, optimality, and robustness of interest rate rules for monetary policy. Thus, 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. The models differ on their degree of micro-foundation and forward-looking behavior: Rotemberg and Woodford (1999) and Rudebusch and Svensson (1999) are but two examples representing the spectrum of choices. However, a unifying thread uniting this research is the need to produce models capable of reproducing the fundamental dynamic properties of actual economies with some degree of accuracy.

Consequently, it is natural to investigate these dynamic properties empirically for variables such as inflation, the output gap, and interest rates so as to provide a benchmark by which to

compare the dynamic properties of competing theoretical models. Here, I specifically consider the following variables: 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). The data for the analysis is quarterly for the sample 1955:I - 2003:I, and is displayed in figure 7.

These data appear to be stationary, an impression confirmed by standard augmented Dickey-Fuller tests. The p-value of the null hypothesis of a unit root is rejected at a 0.56%, 0%, and 12.15% levels for y_t , π_t , and i_t , respectively. Because the computation of impulse responses imposes no constraints on i_t and because standard economic arguments suggest interest rates should be treated as stationary, it seems the rejection of the unit root for i_t at a level higher than 5% is not a significant issue.

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). Using a standard Cholesky decomposition¹² based on the Wold-causal order y_t, π_t , and i_t , figure 8 displays these impulse responses.

The VAR(3) responses are depicted in a pink-dotted line, the green-dashed line and the two red-dashed lines depict the responses from local linear projections and the corresponding two standard-error, Newey-West corrected bands calculated as in section 4. The light blue-solid line is

¹² I choose the Cholesky 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.

the response from a local cubic projection¹³. 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. Thus, 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. Similarly, the response of the output gap y_t to its own shock is statistically different (albeit with the same general shape) but corresponds closely to the output responses 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 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 mentioned¹⁴.

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 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,

¹³ The dark blue-dashed line is simply the zero line.

¹⁴ The figure displaying these responses is available upon request.

or when interest rates are close to the zero bound, for example. Therefore, I tested all the first period local-linear projections¹⁵ for evidence of threshold effects due to y_{t-1}, π_{t-1} , and i_{t-1} using Hansen’s (2000) test¹⁶. For example, a typical regression is,

$$\begin{aligned} z_t &= \Psi'_L X_{t-1} + \varepsilon_t^L & \text{if } w_{t-1} \leq \delta \\ z_t &= \Psi'_H X_{t-1} + \varepsilon_t^H & \text{if } w_{t-1} > \delta \end{aligned} \tag{29}$$

where 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 $\Psi_i, i = L, H$ 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 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 1. Only one combination reports a significant departure from the null of linearity: the response of interest rates with a threshold due to y_{t-1} . Figure 9 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 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.

Table 1 - Hansen’s (2000) test of the null of linearity against the alternative of a threshold (p-values)

¹⁵ I used the local linear projections for the test for parsimony although the final analysis is based on cubic projections.

¹⁶ 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.

Threshold variable	Dependent 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

* significant at a 95% confidence level.

Further investigation revealed that this two-state, interest rate response is significant¹⁷ for the response to an interest rate shock only. Consequently, I allow all three responses¹⁸ to a shock in i_t to vary according to whether the current output gap is positive or negative, but restrict the remaining responses to be constant for threshold effects. Thus, figure 10 displays the responses calculated by local cubic projection along with two standard-error bands and allows for state-dependent responses when the shock is to i_t . The light blue-solid line depicts responses calculated by cubic local projection and correspond to those displayed in figure 8. The accompanying red-dashed lines are two standard-error bands, Newey-West corrected and based on the cubic projection as described in section 4. The last row of figure 9 displays the responses to a shock in i_t and shows with a pink-dotted line the response when the output gap is negative, and with a green-dashed line when the output gap is positive.

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 potential is high, relative to when

¹⁷ The figure showing this result is available upon request.

¹⁸ To be internally consistent, one should allow the responses of y_t and π_t to a shock in i_t to vary as well even if the threshold effect is detected only for i_t since different time profiles for interest rate responses would generate different time profiles for y_t and π_t , even in a constant parameter, linear model.

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. Surely, such differences must be important when evaluating the appropriate response of the monetary authority and illustrate the benefits of having simple but more flexible ways of calculating impulse responses.

7 Conclusion

This paper shows how to calculate impulse response functions for a vector of time series without estimating a specific, dynamic, multivariate model. Instead, I propose estimating a sequence of simple univariate equations by standard regression techniques to obtain robust estimates of the impulse response and its standard-error bands. These methods provide a natural alternative to estimating impulse response functions based on VARs.

Impulse responses calculated by local projections have desirable properties. Monte Carlo evidence showed that they are more robust to misspecification, and that standard-error calculation is simple and direct. Although these methods can be used instead of traditional VARs, their flexibility makes them appealing for a much wider variety of situations. Thus, the empirical application in the paper shows that there is little effort involved in calculating flexible, non-linear impulse response functions that allow for threshold effects. Although there exist multivariate models that allow for similar features (Tsay, 1998 extends the threshold model to the multivariate context, and Krolzig, 1997 introduces a vector version of the markov-switching model), these are restricted in practice by the complexity that their estimation and study requires.

Estimation of impulse responses by local projection methods can be extended in a number of interesting ways. The tone of the paper was deliberately against over-sophistication but it is clear that the complexity of the local approximations can be adapted as circumstances require with the nonlinear and nonparametric techniques mentioned above. Another context where local projections are likely to prove useful is in estimating the dynamic effects of policy interventions in a panel-data context. The relatively short time-dimension samples do not allow estimation of even the simplest of multivariate models. However, estimation of impulse responses with local projections can be easily done by sequential estimation with traditional panel-data techniques.

References

- 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.
- 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.
- 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.
- 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. (1993) "On the Limitations of Comparing Mean Squared Forecast Errors: Comment," *Journal of Forecasting*, 12, 651-652.
- Granger, Clive W. J. and Michio Hatanaka (1964) **Spectral Analysis for Economic Time Series**. Princeton, NJ: Princeton University Press.

- 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.
- 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 (*in press*) "The Response of Term Rates to Monetary Policy Uncertainty," *Review of Economic Dynamics*.
- Kilian, Lutz (1998) "Small Sample Confidence Intervals for Impulse Response Functions," *Review of Economics and Statistics*, 218-230.
- 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.
- 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.
- Phillips, Peter C. B. (1998) "Impulse Response and Forecast Error Variance Asymptotics in Nonstationary VARs," *Journal of Econometrics*, 83, 21-56.
- Rotemberg, Julio J. and Michael Woodford (1999) "Interest Rate Rules in an Estimated Sticky Price Model," in **Monetary Policy Rules**. John B. Taylor (ed.). NBER Conference Report. Chicago: University of Chicago Press, 57-119.
- 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. and Tao Zha (1999) "Error Bands for Impulse Responses," *Econometrica*, v. 67, n. 5, 1113-1156.

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

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

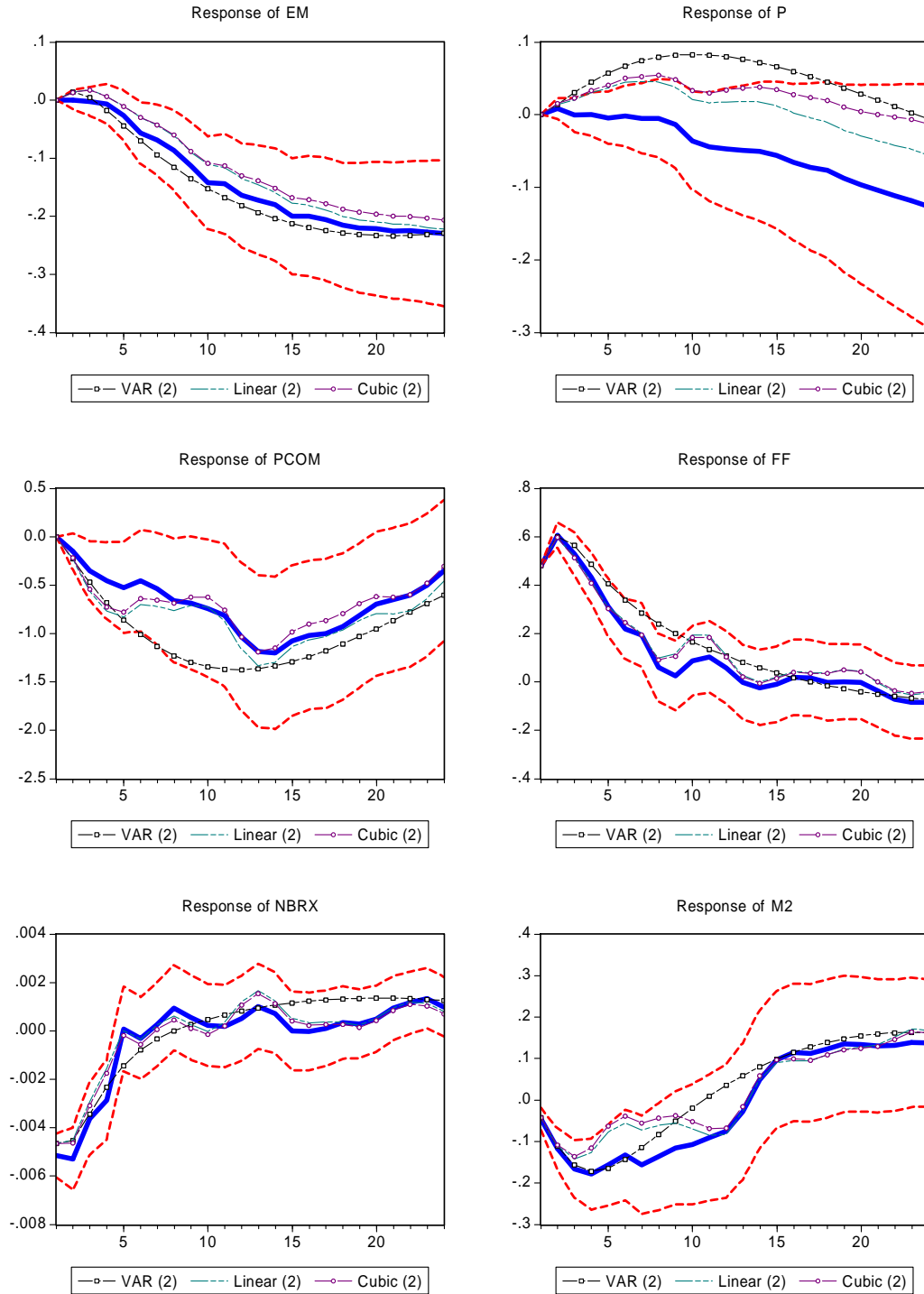
Tong, Howell (1983) **Threshold Models in Nonlinear Time Series Analysis**. Lecture Notes in Statistics, 21. Berlin: Springer.

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.

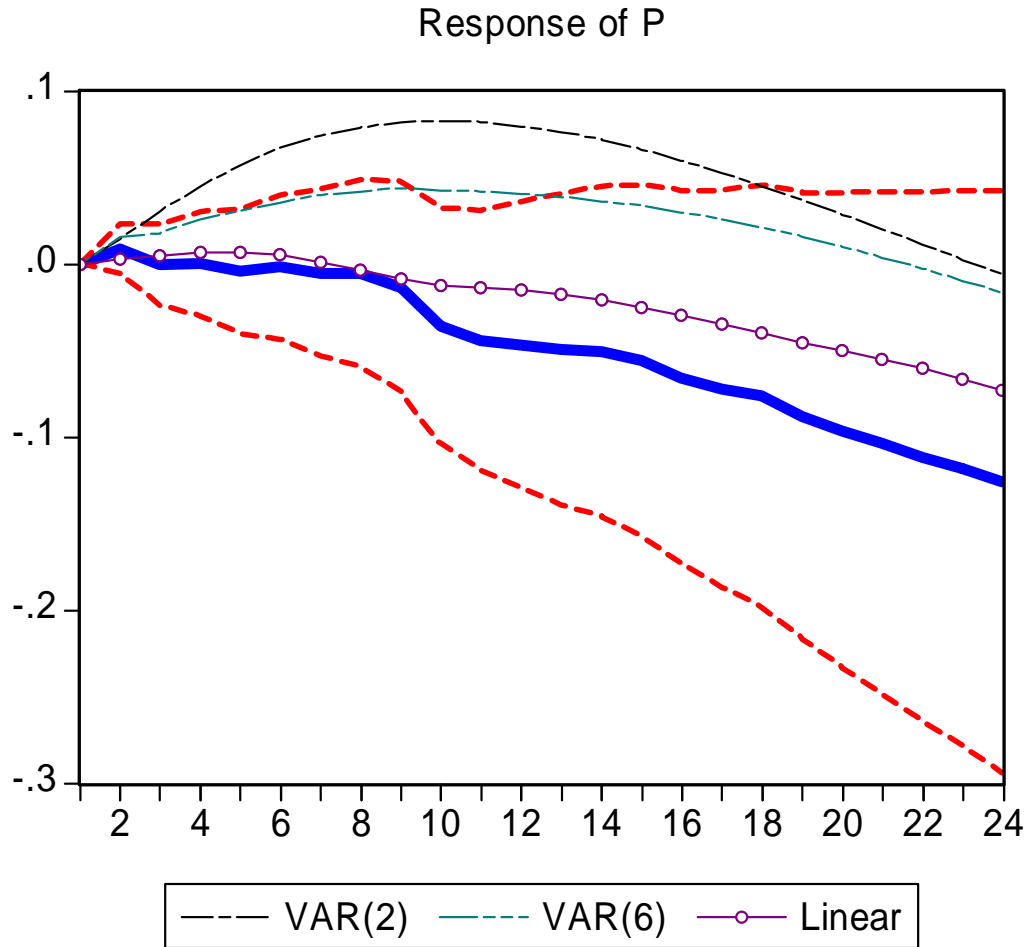
White, Halbert (ed.) (1992) **Artificial Neural Networks: Approximation and Learning Theory**. Oxford: Basil Blackwell.

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



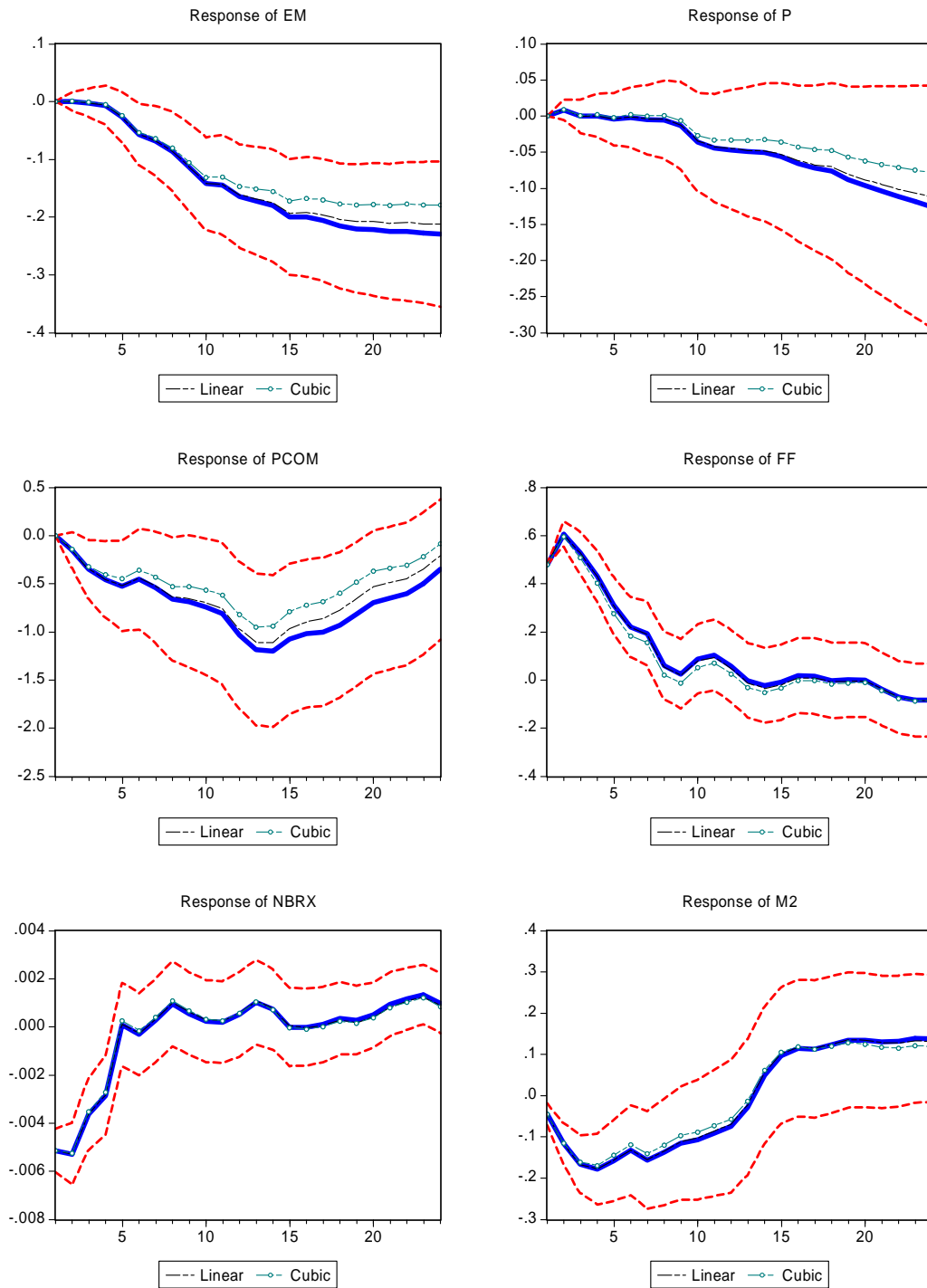
Evans and Marshall (1998) VAR(12) Monte Carlo Experiment. The thick, blue line is the true impulse response based on a VAR(12). The thick-dashed, red lines are Monte Carlo 2-standard error bands. Three additional impulse responses are compared, based on estimates involving two lags only: (1) the response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

Figure 2 – Response of P to a shock in FF: An Alternative Linear Projection



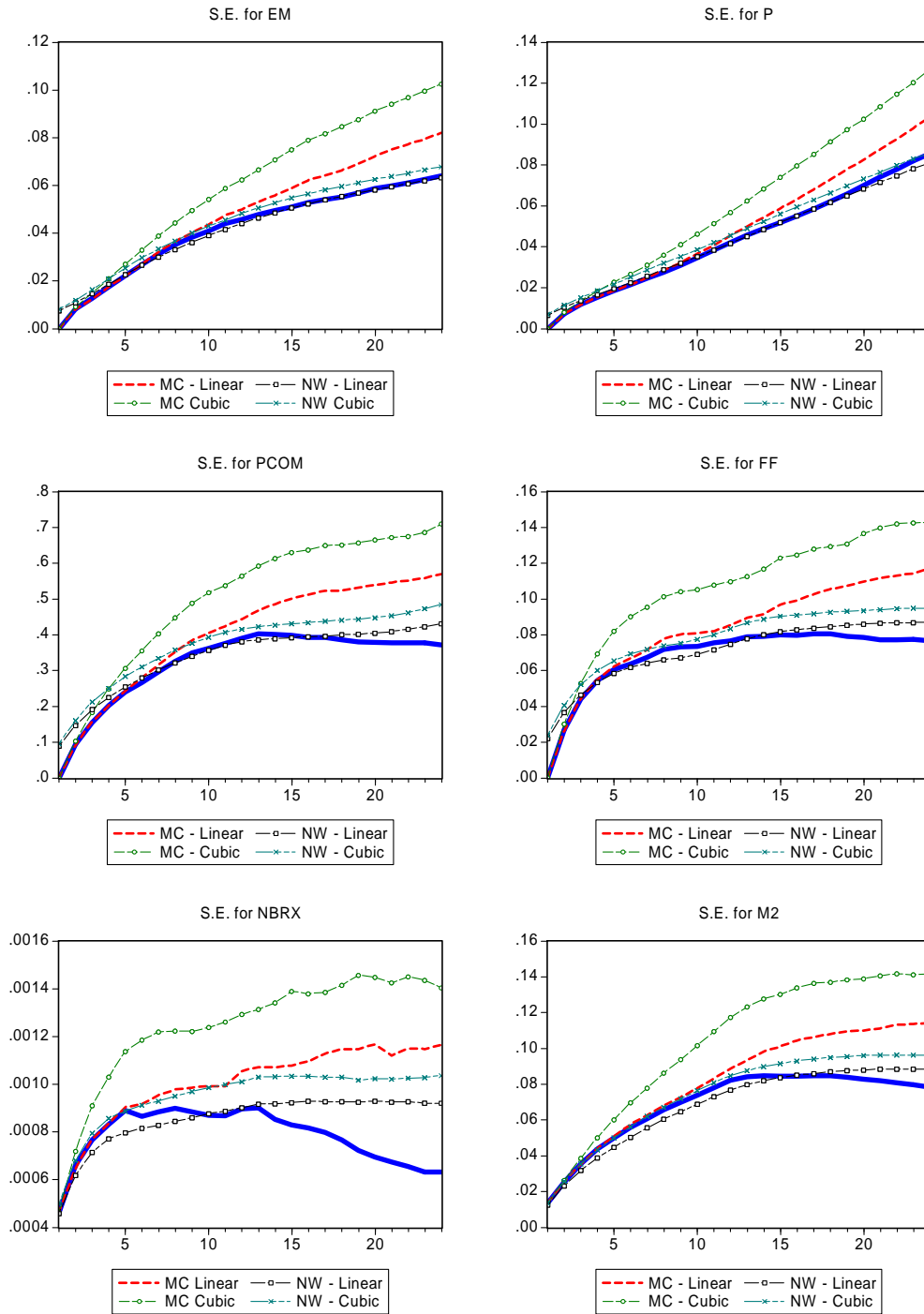
Evans and Marshall (1998) VAR(12) Monte Carlo Experiment. The thick, blue line is the true impulse response based on a VAR(12). The thick-dashed, red lines are Monte Carlo, 2-standard error bands. Three additional impulse responses are compared: (1) the response calculated with a VAR(2), depicted by the black, dashed line; (2) the response calculated with a VAR(6), depicted by the green, dashed line, and (3) the response calculated with a local-linear projection with one lag except for P, where 12 lags are allowed. This is depicted by the purple line with circles. 500 replications.

Figure 3 – Impulse Responses to a Shock in FF. Lag Length: 12



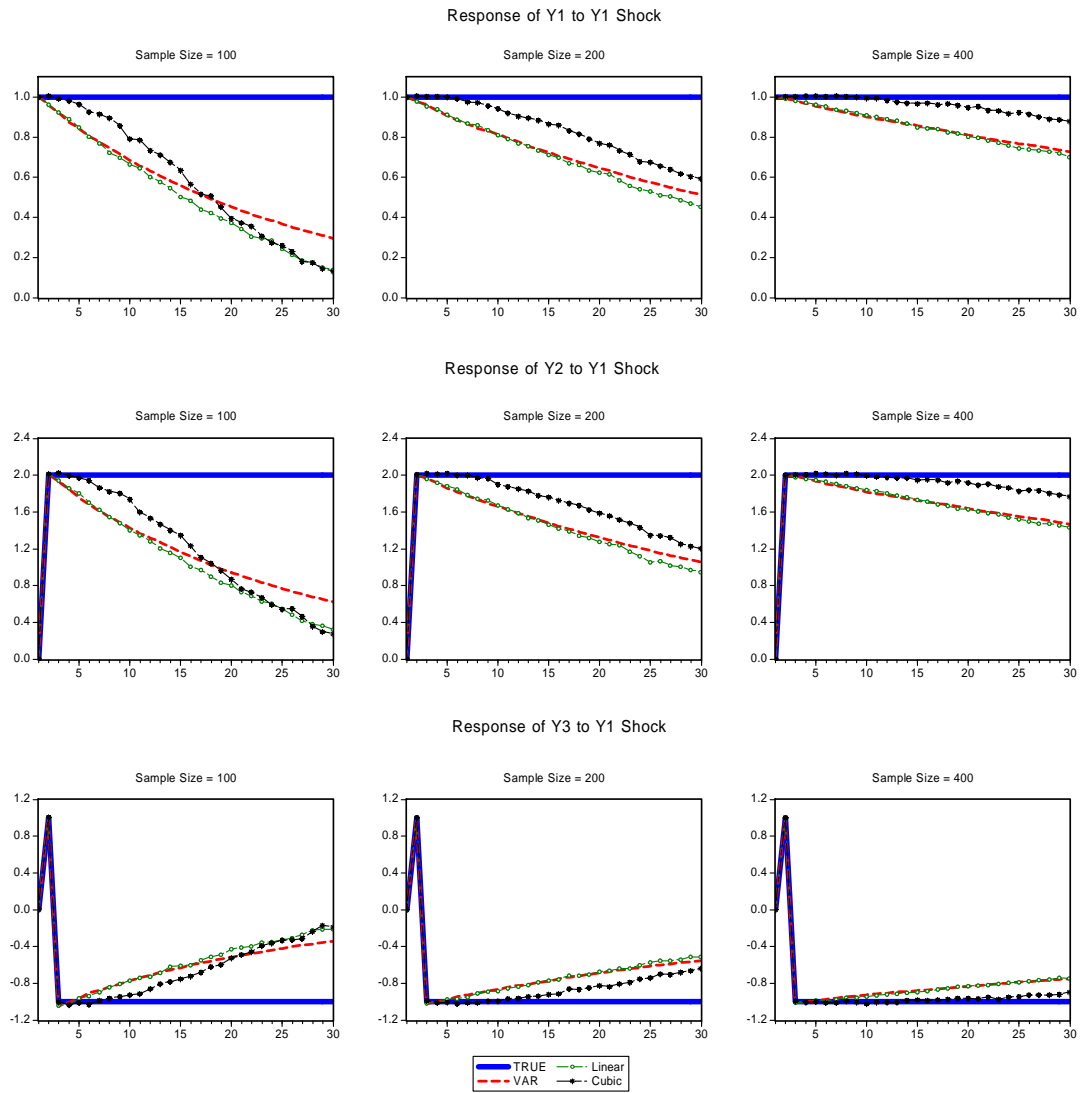
Evans and Marshall (1998) VAR(12) Monte Carlo Experiment. The thick, blue line is the true impulse response based on a VAR(12). The thick-dashed, red lines are Monte Carlo, 2-standard error bands. Two additional impulse responses are compared: (1) the response calculated with a local-linear projection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

Figure 4 – Standard Errors for the Impulse Responses to a Shock in FF. Lag Length: 12



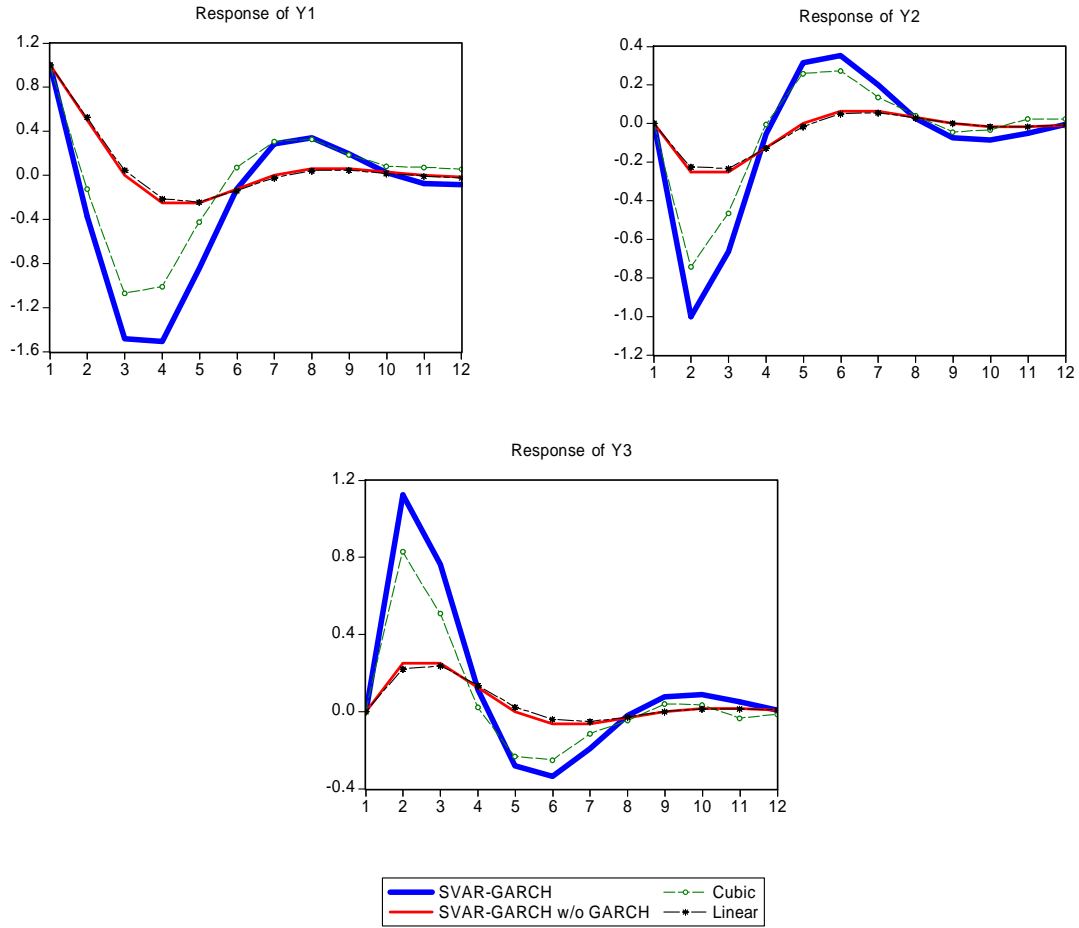
Evans and Marshall (1998) VAR(12) Monte Carlo Experiment. The thick, blue line is the Monte Carlo standard error (MCSE) for the VAR(12). The red, thick dashed line is the MCSE for the local-linear projection. The green, dashed line with circles is the MCSE for the local-cubic projection. The black line with squares is the Newey-West S.E. calculated from the local-linear projection. The green line with stars is the Newey-West S.E. calculated from the local-cubic projection. 500 replications.

Figure 5 – Phillips (1998) Monte Carlo Experiments



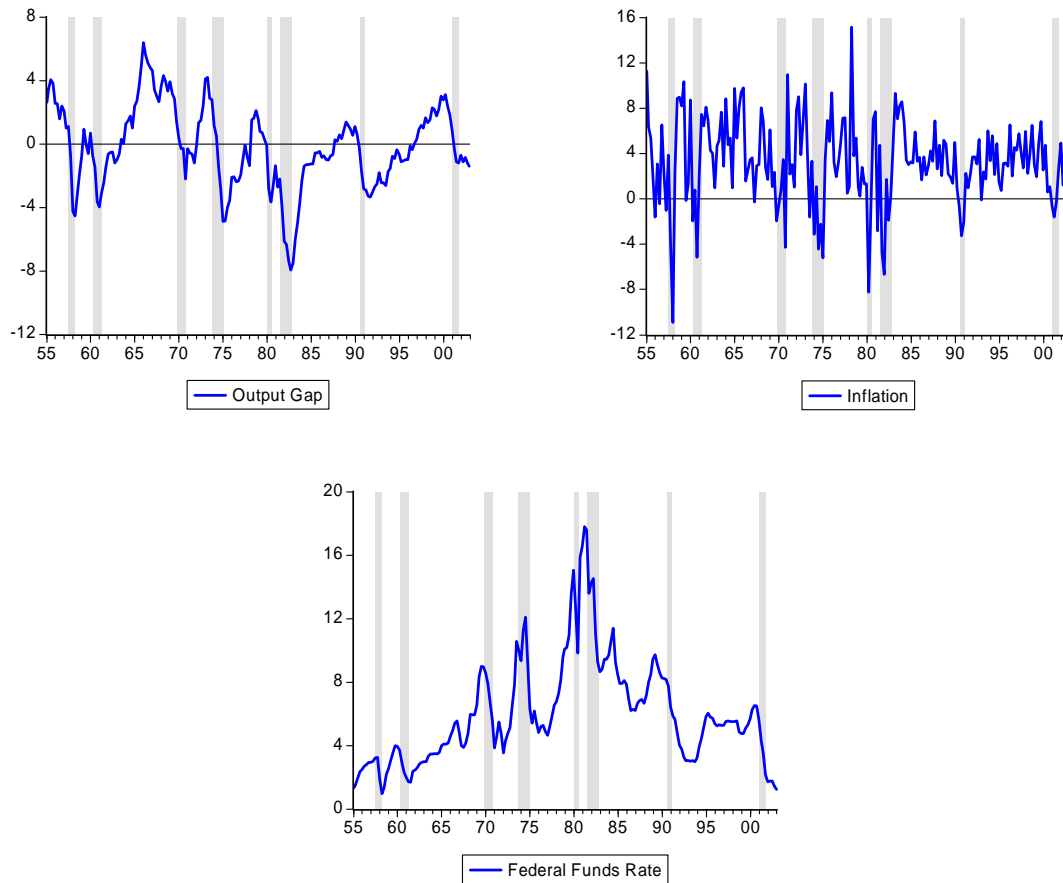
The thick solid blue line displays the true impulse response. The red, dashed line is the impulse response based on a VAR(1), the green line with circles is the local-linear projection with one lag, and the black line with stars is the local-cubic projection.

Figure 6 – Impulse Responses to a Shock in Y1 from a SVAR-GARCH



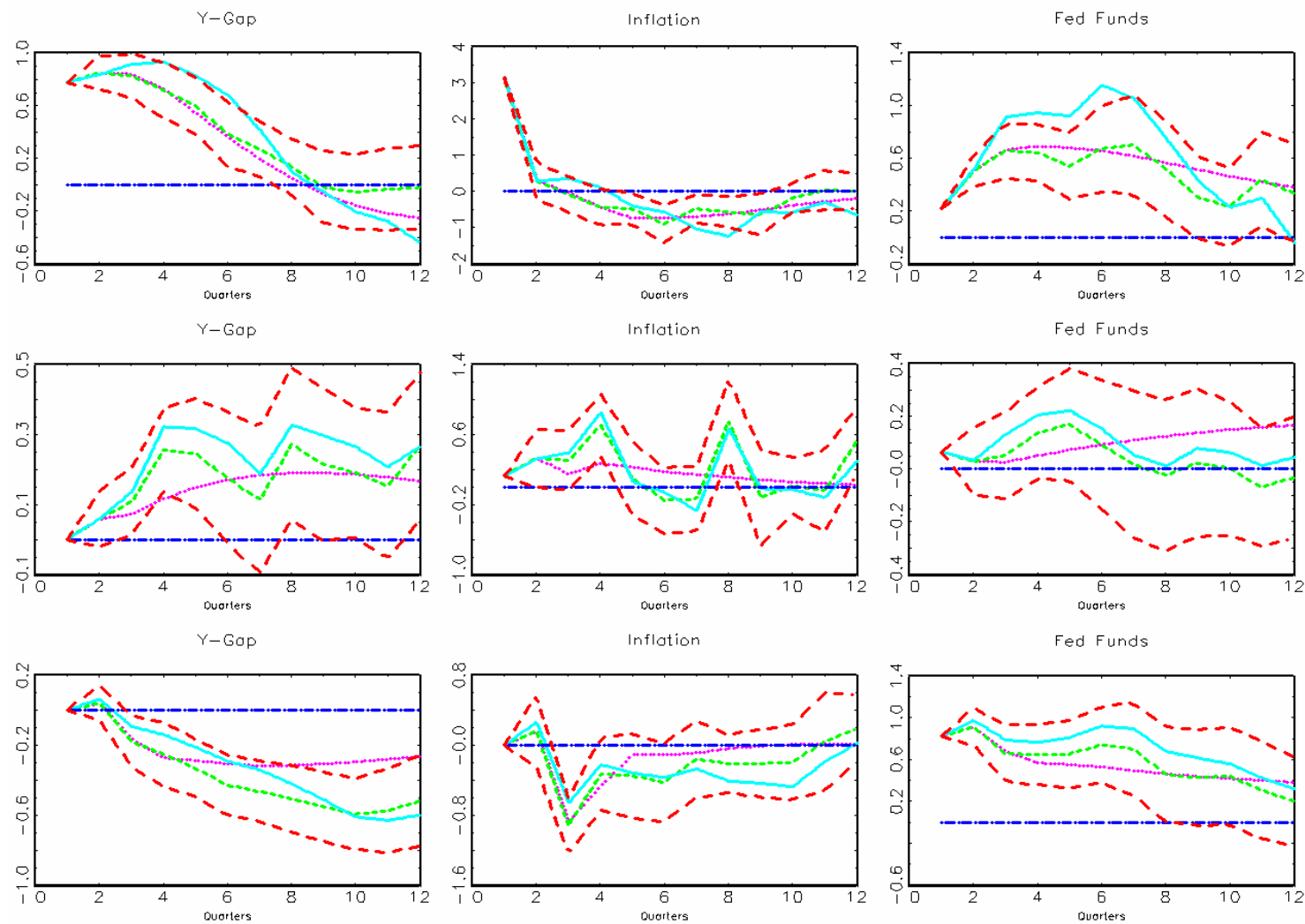
The blue, thick-solid line describes the true impulse response in the VAR-GARCH model. The red, solid line is the impulse response when the variance effects are set to zero (i.e. $B = O_3$). The black, dashed line with stars is the local-linear projection to the impulse response. The green, dashed line with squares is the local-cubic projection to the impulse response. Standard error bands are omitted for clarity but they are fairly narrow (in fact, for periods 2-4 approximately, they exclude the cubic approximation from the truth).

Figure 7 – Time Series Plots of the Output Gap, Inflation, and the Federal Funds Rate



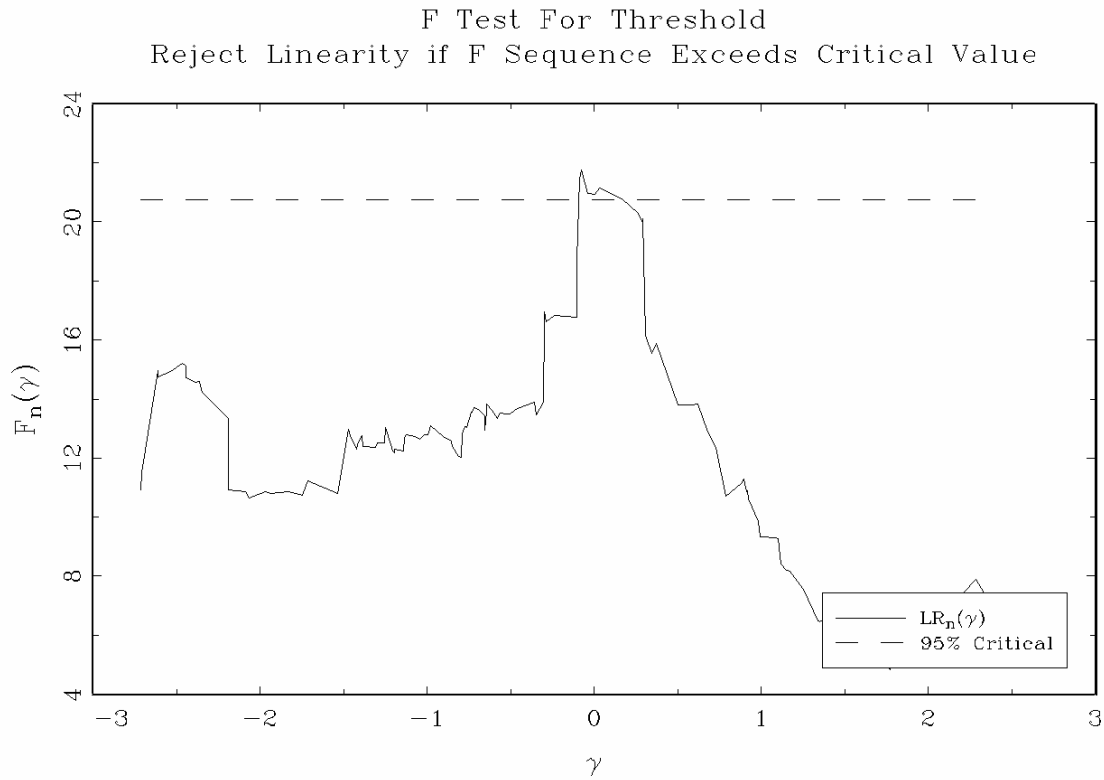
Notes: All variables in annual percentage rates. Shaded areas indicate NBER-dated recessions. Output gap is defined as the percentage difference between real GDP and potential GDP (Congressional Budget Office); Inflation is defined as the percentage change in the GDP, chain-weighted price index at annual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

Figure 8 – Impulse Responses Calculated from: a VAR, a Local-Linear and a Local-Cubic Projections



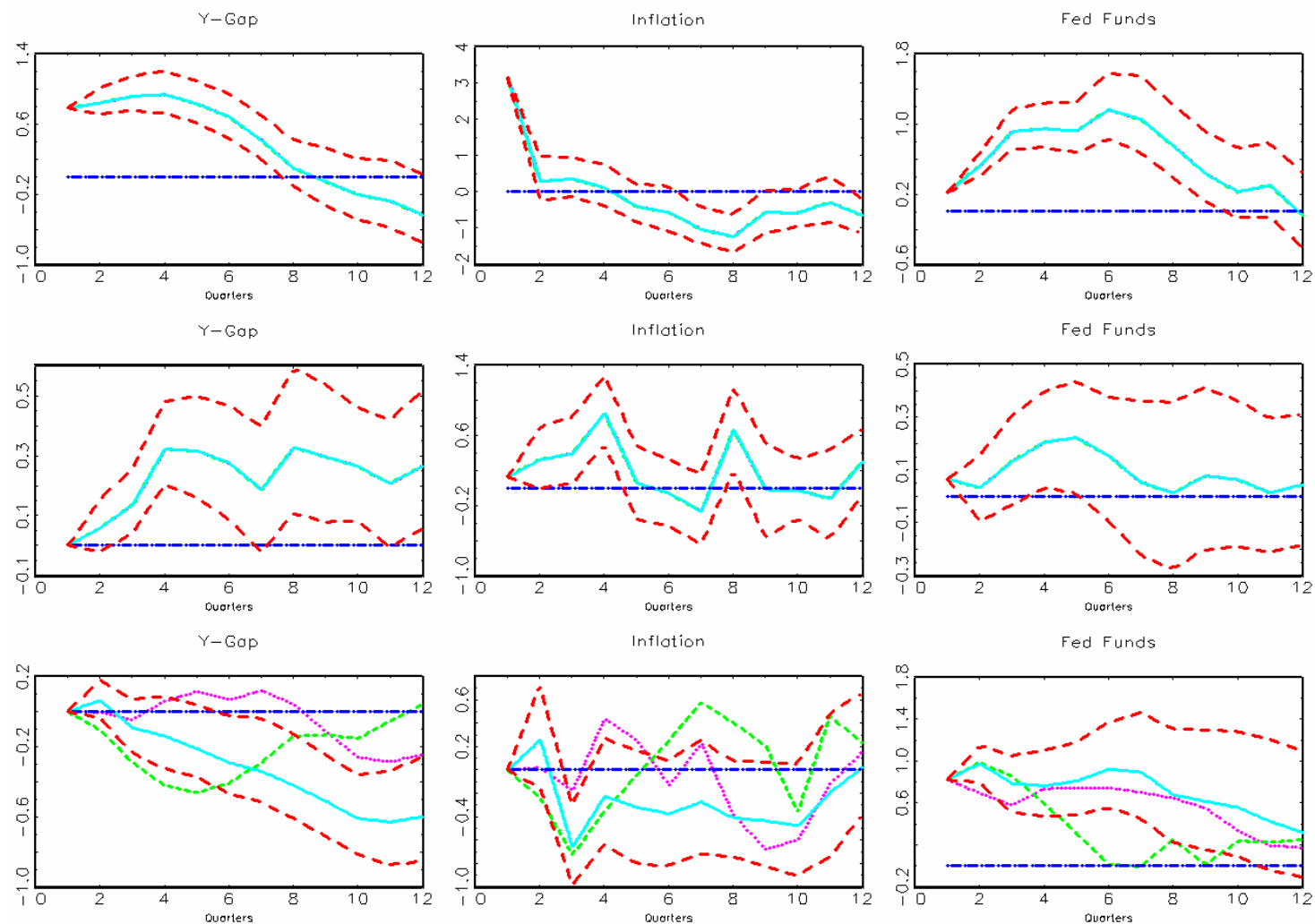
Notes: pink-dotted line is the VAR(3) response, green-dashed line is the IRF based on local linear projection, the red-dashed lines are the corresponding Newey-West corrected 2 S.E. bands for the linear projection. The blue-solid line is the IRF based on cubic projection. The dark blue-dashed line is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

Figure 9 - Sequential Test for a Threshold in y_{t-I} 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 10 – Impulse Responses from Local-Cubic Projections with Threshold Effects for i_t Shocks.



Notes: the blue-solid line is the IRF from a local cubic projection and the red-dashed lines are the corresponding 2 S.E. bands. The pink dotted line is the IRF for a local cubic projection when the output gap is negative while the green-dashed line is the IRF for a local cubic projection when the output gap is positive. All responses in percentages.