IMPULSE RESPONSE ANALYSIS IN CONDITIONAL QUANTILE MODELS AND AN APPLICATION TO MONETARY POLICY

Tae-Hwan Kim[a], Dong Jin Lee[b] and Paul Mizen[c]

[a]School of Economics, Yonsei University, Korea

[b]Economics and Finance Department, Sangmyung University, Seoul, Korea

[c]School of Economics, University of Nottingham, Nottingham, UK

October 2019

Abstract: This paper presents a new method to analyze the effect of shocks on time series using quantile impulse response function (QIRF). While conventional impulse response analysis is restricted to evaluation using the conditional mean function, here, we propose an alternative impulse response analysis that traces the effect of economic shocks on the conditional quantile function. By changing the quantile index over the unit interval, it is possible to measure the effect of shocks on the entire conditional distribution of a given variable in our framework. Therefore we can observe the complete distributional consequences of policy interventions, especially at the upper and lower tails of the distribution as well as at the mean. Using the new approach, it becomes possible to evaluate two distinct features, namely, (i) the degree of uncertainty of a shock by measuring how the dispersion of the conditional distribution is changed after a shock, and (ii) the asymmetric effect of a shock by comparing the responses to an impulse at the lower tails with those at the upper tails of the conditional distribution. None of these features can be observed in the conventional impulse response analysis exclusively based on the conditional mean function. In addition to proposing the

QIRF, our second contribution is to present a new way to jointly estimate a system of multiple quantile functions. Our proposed system quantile estimator is obtained by extending the result of Jun and Pinkse (2009) to the time series context. We illustrate the QIRF on a VAR model in a manner similar to Romer and Romer (2004) in order to assess the impact of a monetary policy shock on the US economy.

Keywords: Quantile vector autoregression; monetary policy shock; quantile impulse response function; structural vector autoregression

JEL classifications: C32, C51

*Corresponding author. Yonsei University, School of Economics,134 Shinchon-dong, Seodaemun-gu, Seoul 120-749, Korea; Tel.: +82-2-2123-5461; fax: +82-2-2123-8638; e-mail address: taehwan.kim@yonsei.ac.kr. Tae-Hwan Kim is grateful for financial support from the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A5A2A01025435).

1 Introduction

Central banks have used past forecast errors or split normal densities of forecasts in order to convey the uncertainty around inflation projections for more than two decades. These projections, known as "fan charts," convey the projected conditional distributions of inflation into the future. The practice began in the Bank of England in 1996, and was quickly followed by other inflation targeting central banks such as the Swedish Riksbank, 1997, and in the Reserve Bank of New Zealand 2002. More recently, in 2017, the Fed added them to its communication tools. The main reasons for the use of fan charts is to convey the uncertainty and asymmetries around the inflation projection, and to avoid spurious precision associated with the use of a single point forecast. Despite the clear logic of this approach, many central banks use fan charts to convey the distribution of inflation in monetary policy communication, but continue to use VAR models and associated impulse response functions at the conditional mean of inflation in order to form monetary policy decisions. There is some inconsistency here, and one that suggests improvements could be made by using distributional information more systematically. To remedy this inconsistency in practice, we propose a quantile impulse response function that can provide conditional quantile impulse responses to show the projections of variables of interest from different parts of the distribution. It is entirely consistent with the logic of fan charts in central bank communication and brings with it all the advantages of showing conditional distributions.

The quantile regression method, originally pioneered by Koenker and Bassett (1978), has become a useful part of the modern econometric toolbox because of its flexibility in permitting researchers to investigate the relationship between economic variables over the entire conditional distribution of interest and particularly at the tails. Recent years have witnessed the surge in applications of the method to time-series models, either theoretically or empirically. Some representative papers include Koenker and Xiao (2006), Galvao (2009), Xiao (2009), Galvao et al. (2009), Greenwood-

Nimmo et al. (2013), Cho et al. (2015), and White et al. (2015), which have provided new insights that conventional mean-centered regression models would not have revealed, such as, for example, a measure of the degree of tail interdependence in terms of value at risk (VaR). Despite these important contributions, scant attention has been paid to the application of the quantile regression method to conjectural economic analysis, especially in measuring the effect of policy shocks on economic variables of interest, such as inflation or output. This is in spite of the fact that the economic environment in which monetary policy is designed since the Great Financial Crisis (GFC) has experienced sluggish growth and exceptionally low inflation and interest rates.

The exception to this story of neglect is White et al. (2015), which traces the effects of shocks in impulse response functions in quantile regression models, as opposed to the conventional mean-centered regressions to derived a pseudo quantile impulse response function tracing the effect of a shock on the conditional quantile function, but in a fairly restrictive setting. The pseudo quantile impulse response function is set up under conditions where (i) they do not allow any dynamics in the first moment of variables in their quantile models; and (ii) they consider only a special case in which a shock is given to the observable variables rather than to the structural error. This paper presents a new and proper impulse response analysis in quantile models by solving the two problems in White et al.(2015). We will allow dynamics in the first moment of structural variables by employing the structural vector-autoregression (SVAR) model, and introduce a shock to structural errors rather than the observable structural variables.

Recently, Chavleishvili and Manganelli (2016) propose another way of deriving quantile impulse response functions independently. Their setting is different from ours in that they consider only a bivariate system of two variables and one of the two is assumed to evolve exogenously to the system. Such a setting may be suitable for financial markets where the market portfolio can be assumed to be exogenous to individual stock returns. The method to define a structural shock is

also different. They set the structural error for the exogenous variable to zero in such way that the exogenous variable is equal to a specific quantile. Hence, the shock is given de facto to the observable exogenous variable, similarly to White et al. (2015). On the contrary, we will consider a general multivariate system where all the variables are endogenous and a shock is given to the corresponding structural error.

To develop the proposed method, we start with the SVAR in the conditional mean, which is used to identify a structural shock.[1] We permit an intervention into the structural errors to affect the entire conditional distribution, and the effect of an identified structural shock on the conditional quantile function is called the "quantile impulse response function" (QIRF). This offers a method to observe the effect of shocks given to the structural error on the entire conditional distribution of the observable structural variables, and not just the mean. It also has two other advantages over conventional impulse responses that are consistent with the logic behind the use of fan charts for inflation projections. First, the impact of shocks on the inflation distribution (represented visually by a fan chart) conveys the uncertainty surrounding a structural shock. The impact of shocks on the distribution can be measured by the conditional quantile ranges of some key economic variables such as inflation based on the dispersion of the conditional distribution. Second, QIRF can capture asymmetry in the responses under different circumstances, so that behavior of economic agents under high inflation risk does not need to be identical (symmetric) to behavior toward low inflation risk. The asymmetry can be captured by the different responses between upper and lower quantiles shown by an asymmetric QIRF with respect to the direction of a shock, that is, positive or negative shocks have different impacts. Therefore, our methods provide researchers and policy makers with

[1] Once a structural shock is identified in the conditional mean VAR model, we do not need further to impose identifying restrictions on the subsequently defined conditional quantile model and therefore, using a reduced form does not pose any problem in obtaining the quantile response to the identified shock.

a broader perspective on the dynamics of macroeconomic and finance variables following a shock.[2] The new methods are useful to central banks in setting policies under conditions that their key variables are likely to be in the tails of their conditional distributions, rather than at the mean, that is, deep recessions, and ultra low inflation and interest rates. In this sense, the proposed QIRF can allow researchers to investigate the effects of a monetary shock to some key macroeconomic variables at the tails as required without making the assumption (which may not be valid) that the effects are the same as those reported at the conditional mean or symmetric around the mean.

In addition to the main contribution of proposing the QIRF, another contribution of this study to the literature is to present a new way to jointly estimate a system of multiple quantile functions. Jun and Pinkse (2009) has developed a system estimation method for multiple conditional quantile functions, but their method is not directly applicable to serially dependent variables such as ours. Hence, we extend the system quantile estimator of Jun and Pinkse (2009) to the time series context. Specifically, we first suggest a set of consistent estimators for all parameters in the system, based on the weighted quantile moment conditions. Then, an efficient GMM type estimator is proposed where the moment weight follows the idea of Jun and Pinkse (2009). The estimator is specialized to Koenker and Vuoung(2009)'s efficient estimator in univariate cases and is equivalent to Jun and Pinkse (2009) if variables are iid. Considering both the possibility of multiple local optima and the curse of the dimensionality problem, we suggest using the Laplace type quantile estimation (LTE) technique of Chernozhukov and Hong (2003). We provide conditions for the consistency, and derive its asymptotic distribution.

We apply the proposed method to assess the impact of monetary policy shocks on the US

[2]It is possible that a positive shock reduces either the conditional variance or the conditional inter-quantile ranges of the whole conditional distribution of the variable of interest, while a negative shock can have the opposite effect. Similar attempts to capture the asymmetric impulse responses have been introduced using Markov-switching or threshold models to the conventional VAR (Ehrmann et al., 2003; Granger and Yoon, 2002; Hatemi, 2014).

economy using a standard three variable VAR, in employment growth, inflation, and the Romer and Romer (2004) measure of the monetary policy shock. Using our QIRF approach, we demonstrate the effects of contractionary and expansionary monetary policy shocks on the whole conditional distributions for employment growth and inflation. We can illustrate the asymmetric responses of the distribution in each of the tails and measure the change in the dispersion of the distribution after contractionary and expansionary monetary policy shocks. These additional pieces of information provide the policy maker with a fuller understanding of the effects of policy on the conditional distribution of variables of interest.

Recently Adrian et al. (2019) have used multiple quantile models to analyze the asymmetric patterns in the conditional distributions of US growth and inflation rates. They assume an asymmetric t-distribution for the growth and inflation variables, and use the fitted values from multiple quantile regressions to estimate the parameters of the conditional distributions. A similar method can be applied to our quantile impulse response functions in such a way that multiple quantile impulse responses can be used to estimate the changes in the shape of generalized parametric conditional distribution functions (such as skewed t-distribution (Fernandez and Steel, 1998), generalized t-distribution (Theodossiou, 1998), and asymmetric power distribution (Fernandez et al., 1995; Komunjer, 2007) after an economic shock. Our VAR model for QIRF can be also generalized to nonlinear quantile models such as CAViaR of Engle and Manganelli (2004) and MQCAViaR of White et al. (2014, 2015) if we add lags of the quantile functions as exogenous variables.

The rest of this paper is organized as follows. Section 2 introduces a linear conditional quantile model in the SVAR framework and Section 3 proposes the quantile impulse response function. Section 4 provides estimation methods. Section 5 shows an application of the quantile impulse response to US monetary policy. Section 6 provides some concluding remarks.

2 The SVAR Model with Heteroscedastic Quantiles

Let us consider a sequence of random variables denoted by $\{z_t\} = \{(y_t', x_t') : t = 1, 2, ..., T\}$ where y_t is a $n \times 1$ vector given by $y_t = (y_{1t}, \ldots, y_{nt})'$ and x_t is a countably dimensioned $m \times 1$ vector. We will assume that z_t has been demeaned. Note that y_t is the set of variables of primary interest and x_t is of secondary interest used to explain y_t. We consider a structural vector-autoregressive (SVAR) model for y_t as follows:

$$A(L)y_t = \epsilon_t$$
$$A(L) = A_0 + A_1 L + \ldots + A_p L^p \tag{1}$$

where $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{nt})'$ is the vector of mean zero disturbances. We impose the following assumption on the SVAR model in (1).

Assumption 1. (i) All values of w satisfying $|A_0 + A_1 w + \ldots + A_p w^p| = 0$ lie outside the unit circle.

(ii) $A(L)$ satisfies the order condition for identifying the structural equation.

(iii) $\{\epsilon_t\}$ is $\phi-$mixing of size $-r/(r-2)$ with $r > 2$, and $\sup_t E|\epsilon_{i,t}^{2(r+\varepsilon)}| < \infty$ for some $\varepsilon > 0$ and for each $i = 1, \ldots, n$.

Assumptions 1(i) through 1(ii) are standard in the SVAR framework. Turning to Assumption 1(iii), we note that $\phi-$mixing is stronger than $\alpha-$mixing. Nevertheless, we impose such a strong condition because it is required to obtain an efficient weight function for the Laplace type quantile estimator, which will be explained in detail in Section 4. If we wish to obtain only a consistent quantile estimator (not necessarily efficient), then the $\phi-$mixing condition can be relaxed to the $\alpha-$mixing condition. We also note that the mixing condition on ϵ_t does not necessarily imply that y_t is a mixing sequence as discussed in Andrews (1983). Instead, Assumption 1(iii) together

with (i) indicates that $\{y_t\}$ is near-epoch-dependent (NED) which is sufficient to obtain the desired asymptotic properties of the efficient quantile estimator as discussed in Section 4.

If the distribution of ϵ_t does not depend on the lags of y_t, the SVAR model in (1) is the same as the conventionally-used SVAR, where an intervention to one of the structural shocks will affect the future dynamics of y_t only through its conditional mean, which will produce the conventional impulse response function. In such a case, the effect of the intervention on other parts (i.e., quantiles) of the conditional distribution can be straightforwardly inferred from the mean effect because all the impulse response functions at different quantiles will be parallel to the conventional mean impulse response function. However, Assumption 1 does not eliminate the possibility that the structural error term ϵ_t can depend on lagged y_t. Such a possibility implies that an economic shock can affect not only the conditional mean, but also the whole distribution of y_t in a non-trivial manner. Since the effect on the conditional mean function is obviously captured by the conventional impulse response function, the objective of this paper is to develop a new method that can capture the effect on the conditional quantile function of y_t.

The relationship between the two effects mentioned above can be easily seen by decomposing $y_{i,t}$ into two parts, the conditional mean and its deviation from the conditional mean, as follows:

$$y_{i,t} = E(y_{i,t}|\mathcal{F}_{t-s}) + u_{i,t|t-s}, \tag{2}$$

where $u_{i,t|t-s} = y_{it} - E(y_{i,t}|\mathcal{F}_{t-s})$ and $\mathcal{F}_{t-s}$ is the σ-algebra generated by $\{z'_{t-s}, z'_{t-s-1}...\}$. Simply speaking, $\mathcal{F}_{t-s}$ is the collection of information available at time $t-s$. As stated before, the conditional mean part in (2) is dealt with by the conventional impulse response function and the remaining part ($u_{i,t|t-s}$) has been left largely unexplained in the literature.

Given that our main methodology is based on quantile models, we first define $F_{i,t|t-s}(y) = P[y_{i,t} \leq y|\mathcal{F}_{t-s}]$ which is the cumulative distribution function of $y_{i,t}$ conditional on $\mathcal{F}_{t-s}$ with the corresponding conditional density function $f_{i,t|t-s}(y)$. Given a quantile index $\alpha \in (0,1)$, the

$\alpha^{th}-$quantile of the distribution of $y_{i,t}$ conditional on the information set $\mathcal{F}_{t-s}$, denoted $q^{\alpha*}_{i,t,s}$, is defined as

$$q^{\alpha*}_{i,t,s} := \inf_{v \in R} \{v : F_{i,t|t-s}(v) \geq \alpha\},$$

and if $F_{i,t|t-s}$ is strictly increasing,

$$q^{\alpha*}_{i,t,s} = F^{-1}_{i,t|t-s}(\alpha).$$

In other words, the conditional quantile $q^{\alpha*}_{i,t,s}$ is such that the conditional probability that $y_{i,t}$ is smaller than $q^{\alpha*}_{i,t,s}$ is α.[3]

Whenever it is convenient, the $\alpha^{th}-$quantile of the distribution of $y_{i,t}$ conditional on the information set $\mathcal{F}_{t-s}$ is also denoted by $Q^{\alpha}(y_{i,t}|\mathcal{F}_{t-s})$ because of its analogy with the corresponding conditional expectation $E(y_{i,t}|\mathcal{F}_{t-s})$. If we restrict our attention to a linear quantile model such that $q^{\alpha*}_{i,t,s} = z'_{t-s}\beta_{i,\alpha}$, the quantile model can be rewritten in a more familiar formulation as:

$$y_{i,t} = z'_{t-s}\beta_{i,\alpha} + \varepsilon^{\alpha}_{i,t}, \tag{3}$$

where $\varepsilon^{\alpha}_{i,t}$ satisfies the quantile restriction $P[\varepsilon^{\alpha}_{i,t} < 0|\mathcal{F}_{t-s}] = \alpha$.

The source of heteroscedastic quantile effect (i.e., $\beta_{i,\alpha}$ varying with α) can come from either heteroscedastic errors or non-separable errors. For example, if we start with $y_{i,t} = z'_{t-s}\beta_i + \varepsilon^{\alpha}_{i,t}$ where β_i is constant and $\varepsilon^{\alpha}_{i,t} = (z'_{t-s}\gamma)\eta_t$ where η_t is independent and identically distributed, then it can be shown that $\beta_{i,\alpha} = \beta_i + z'_{t-s}\gamma q^{\alpha}_{\eta,t,s}$ where $q^{\alpha}_{\eta,t,s}$ is the $\alpha^{th}-$quantile of η_t conditional on $\mathcal{F}_{t-s}$. The following assumption imposes that the conditional quantile function $q^{\alpha*}_{i,t,s}$ has a form of autoregression.

[3]Rather than focusing on a specific quantile index $\alpha \in (0,1)$, we can consider a set of multiple quantile indexes α_k with $k = 1, 2, ..., m$ in which these m quantile indexes are ordered such that $0 < \alpha_1 < ..., < \alpha_m < 1$. Our theory is sufficiently general enough to accommodate such multiple quantile indexes jointly. However, we present the theory in the text using only a specific quantile index α for clarity.

Assumption 2. (i) $F_{i,t|t-s}(y), s = 1, 2, ..., h,\ i = 1, \ldots, n$ is continuous and positive with the density function $f_{i,t|t-s}(y)$ which is finite and continuous for all $y \in R$.

(ii) For a given finite integer p, there exist real $n \times 1$ vectors $\gamma^{\alpha*}_{i,s,j}$ and $m \times 1$ vectors $\phi^{\alpha*}_{i,s}$ for $s = 1, 2, ..., h, i = 1, 2, ..., n$ and $t = 1, 2, ..., T$ such that we have the following

$$q^{\alpha*}_{i,t,s} = \sum_{j=0}^{p-1} \gamma^{\alpha*\prime}_{i,s,j} y_{t-s-j} + \phi^{\alpha*\prime}_{i,s} x_{t-s}. \tag{4}$$

(iii) If x_t is weakly exogenous, it is $\phi-$mixing with the same size and moment condition as ϵ_t. Otherwise x_t is such that $\frac{\partial^m x_{t-s}}{\partial \gamma^{\alpha}_{i,s,j}{}^m},\ m = 1, 2$, exists and is NED on ϵ_t with $E[\|\frac{\partial^m x_{t-s}}{\partial \gamma^{\alpha}_{i,s,j}{}^m}\|^2] < \infty$ and NED numbers $\eta(s) = O(s^{-\eta})$.

We note that the number of lagged terms in (4) is set to be the same as the number of lagged terms in (1) to simplify the notation. Our theory is general enough to accommodate different numbers of lagged terms in both specifications if desired. Defining $q^{\alpha*}_{t,s} := (q^{\alpha*}_{1,t,s}, q^{\alpha*}_{2,t,s}, ..., q^{\alpha*}_{n,t,s})'$, we note that the expression in (4) can be expressed as a vector form:

$$q^{\alpha*}_{t,s} = \Gamma^{\alpha*}_{s}(L) y_{t-s} + \Psi^{\alpha*}_{s} x_{t-s}, \tag{5}$$

where

$$\Gamma^{\alpha*}_{s}(L) = \Gamma^{\alpha*}_{s,0} + \Gamma^{\alpha*}_{s,1} L + \ldots + \Gamma^{\alpha*}_{s,p-1} L^{p-1}$$

$$\Gamma^{\alpha*}_{s,j} = \begin{bmatrix} \gamma^{\alpha*\prime}_{1,s,j} \\ : \\ \gamma^{\alpha*\prime}_{n,s,j} \end{bmatrix},$$

$$\Psi^{\alpha*}_{s} = \begin{bmatrix} \phi^{\alpha*\prime}_{1,s} \\ : \\ \phi^{\alpha*\prime}_{n,s} \end{bmatrix}.$$

Note that Assumption 2 requires the quantile function to be linear for each prediction horizon $s = 1, 2, ..., h$. The reason why we need to impose this condition is that, unlike the conditional mean equation in which the conditional expectation of $y_{i,t+s}$ given $\mathcal{F}_t$ (i.e., $E(y_{i,t+s}|\mathcal{F}_t)$) can be obtained from $E(y_{i,t+1}|\mathcal{F}_t)$ in a recursive manner, $Q^\alpha(y_{i,t+s}|\mathcal{F}_t)$ cannot be obtained from $Q^\alpha(y_{i,t+1}|\mathcal{F}_t)$ recursively.[4] Thus, different models are required for each prediction horizon s. This assumption may be considered rather restrictive and we discuss how such an assumption can be relaxed in the next section to deal with this criticism. Our assumption for x_t is weak enough to cover a broad set of variables which includes the lags of $q^{\alpha*}_{t,s}$. In that case, (5) is a generalization of CAViaR of Engle and Manganelli (2004) and MQCAViaR of White et al. (2014, 2015), which are known to cover nonlinear structures including the ARCH effect.

Consider, for example, the SVAR process in (1) with heteroscedastic errors where ϵ_t is given by

$$\begin{aligned} \epsilon_t &= \Gamma_t \epsilon^*_t, \\ \Gamma_t &= \oplus_{i=1}^n \gamma_{i,t}, \\ \gamma_{i,t} &= \gamma^i_0 + \gamma^{i\prime}_1 y_{t-1} + \gamma^{i\prime}_2 y_{t-2} + \ldots + \gamma^{i\prime}_p y_{t-p}, \end{aligned} \tag{6}$$

where $\epsilon^*_t \sim IID(0, I_n)$, $\oplus$ denotes the matrix direct sum, γ^i_0 is a positive real number, and γ^i_j is an $n \times 1$ real vector for $j = 1, 2, ..., p$. To show that the example process in (6) satisfies Assumption 2, we let $p = 1$ for the sake of simplicity. Let I_n denote the $n \times n$ identity matrix. Then, one can easily show that

$$y_t = \beta_{t,s} y_{t-s} + u^*_{t,s}, \tag{7}$$

[4] Such a recursion for the expectation function is possible thanks to the linearity of the expectation operator, which does not hold for the quantile operator.

where

$$\begin{aligned}
\beta_{t,s} &= \prod_{i=0}^{s-1}(-A_0^{-1}A_1 + A_0^{-1}\Xi_{t-i}\gamma_1) \text{ for } s = 1, 2, ..., h, \\
\Xi_t &= \oplus_{i=1}^{n}\epsilon_{i,t}^*, \\
u_{t,s}^* &= \sum_{i=0}^{s-1}\beta_{t,i}\Xi_{t-i}\gamma_0, \\
\beta_{t,0} &= I_n, \\
\gamma_0 &= (\gamma_0^1, \ldots, \gamma_0^n)', \\
\gamma_1 &= (\gamma_1^1, \ldots, \gamma_1^n)'.
\end{aligned}$$

Note that $Q^\alpha(y_{i,t}|\mathcal{F}_{t-s})$ is the solution to the following equation:

$$P[y_{i,t} \leq Q^\alpha(y_{i,t}|\mathcal{F}_{t-s})|\mathcal{F}_{t-s}] = \alpha. \tag{8}$$

If $E[\psi_\alpha(u_{t,s,i}^*)] = 0$ where $\psi_\alpha(u_{t,s,i}^*) = \alpha - 1_{[u_{t,s,i}^* \leq 0]}$ and $u_{t,s,i}^*$ is the i^{th} element of $u_{t,s}^*$, there exists γ_α such that

$$Q^\alpha(y_t|\mathcal{F}_{t-1}) = \gamma_\alpha y_{t-1},$$

which implies that Assumption 2 is satisfied for the SVAR process with heteroscedastic errors in (6) for $s = 1$. Alternatively, one can show that Assumption 2 is also satisfied if $\gamma_1^1 = \gamma_1^2$ and ϵ_t^* is normally distributed for $s = 1$. We note that we need some additional conditions for the existence of γ_α for $s > 1$.

The quantile function $Q^\alpha(y_{i,t}|\mathcal{F}_{t-s})$ can have some alternative representation which can be derived using $u_{i,t|t-s}$ in (2). If we denote the conditional quantile of $u_{i,t|t-s}$ by either $q_{u,i,t,s}^{\alpha*}$ or $Q^\alpha(u_{i,t|t-s}|\mathcal{F}_{t-s})$, there is one-to-one correspondence between $Q^\alpha(y_{i,t}|\mathcal{F}_{t-s})$ and $Q^\alpha(u_{i,t|t-s}|\mathcal{F}_{t-s})$, which is given by $Q^\alpha(y_{i,t}|\mathcal{F}_{t-s}) = E(y_{i,t}|\mathcal{F}_{t-s}) + Q^\alpha(u_{i,t|t-s}|\mathcal{F}_{t-s})$ so that the equation in (5) can be replaced by

$$q_{u,t,s}^{\alpha*} = \Gamma_{u,s}^{\alpha*}(L)y_{t-s} + \Psi_s^{\alpha*}x_{t-s}, \tag{9}$$

where $q^{\alpha*}_{u,t,s} = (q^{\alpha*}_{u,1,t,s}, q^{\alpha*}_{u,2,t,s}, ..., q^{\alpha*}_{u,n,t,s})'$, $\Gamma^{\alpha*}_{u,s}(L) = \Gamma^{\alpha*}_{s}(L) - A_s(L)$ and $A_s(L)$ is such that $E(y_t|\mathcal{F}_{t-s}) = A_s(L)y_{t-s}$.

3 Quantile Impulse Response

One of the main strengths of using a VAR is that it allows us to examine the dynamic response of a variable to an identified economic shock using impulse response functions, which are conventionally calculated using a moving average transformation of (1) as:

$$y_t = C(L)\epsilon_t,$$
$$C(L) = C_0 + C_1 L + C_2 L^2 + \dots. \tag{10}$$

If ϵ_t is independent and identically distributed, then the response of $y_{i,t+s}$ to a shock in ϵ_{jt} is simply $\frac{\partial y_{i,t+s}}{\partial \epsilon_{jt}} = C^{ij}_s$ where C^{ij}_s is the (i,j) element of C_s. It is well known that the function $\frac{\partial y_{i,t+s}}{\partial \epsilon_{jt}} = C^{ij}_s$ measures the effect of a shock on the conditional mean function of $y_{i,t+s}$ so that it will be referred to as the canonical mean impulse response function (MIRF). However, the dependency of the distribution of ϵ_{t+i} $(i = 1, \dots, s)$ on y_t imposed by Assumption 2 implies that a shock can change not only the conditional mean, but also the whole conditional distribution of y_{t+s} in a non-trivial manner.

In the example in (6) where the structural shock ϵ_t is unexpected but the size of ϵ_t is related to the past (Γ_t) , the change in ϵ_{t+s} with respect to a unit change in ϵ_t is given by $\frac{\partial \Gamma_{t+s}}{\partial \epsilon_t'}\epsilon^*_{t+s}$. Then the impulse response function is obtained as

$$\frac{\partial y_{t+s}}{\partial \epsilon_t'} = C_0 \frac{\partial \Gamma_{t+s}}{\partial \epsilon_t'}\epsilon^*_{t+s} + C_1 \frac{\partial \Gamma_{t+s-1}}{\partial \epsilon_t'}\epsilon^*_{t+s-1} + ... + C_{s-1}\frac{\partial \Gamma_{t+1}}{\partial \epsilon_t'}\epsilon^*_{t+1} + C_s$$

which is due to the fact that $\Gamma_{t+s}, ..., \Gamma_{t+1}$ are functions of ϵ_t as specified in (6). We note that $\frac{\partial y_{t+s}}{\partial \epsilon_t'}$ depends on unknown future error terms $\epsilon^*_{t+j},\ j = 1, \dots, s$. In other words, the entire future

distribution of y_{t+s} is affected by a shock to ϵ_t. We capture such a response of the entire distribution using the changes in its conditional quantiles.

To capture the non-trivial changes in conditional quantiles, we propose two concepts of quantile impulse response functions denoted by $QIRF_1^\alpha(s)$ and $QIRF_2^\alpha(s)$, respectively. Analogous to the MIRF, the first one $QIRF_1^\alpha(s)$ is defined as

$$QIRF_1^\alpha(s) \ = \frac{\partial q_{t+s,s}^{\alpha*}}{\partial \epsilon_t'},$$

where s is the response horizon; $s = 1, 2, ..., h$.

Using the quantile specification in (5), one can easily show that

$$QIRF_1^\alpha(s) \ = \Gamma_{s,0}^{\alpha*} C_0 + \Psi_s^{\alpha*} \frac{\partial x_t}{\partial \epsilon_t'}. \tag{11}$$

Although $QIRF_1^\alpha(s)$ is intuitively appealing due to its analogy to the MIRF, implementing $QIRF_1^\alpha(s)$ can be computationally demanding. Its computation is similar to the local projection method in that, due to the nonexistence of the Wald representation in the quantile series, one needs a different quantile equation for each response horizon $s = 1, ..., h$ as defined in Assumption 2. Thus its implementation can be computationally intensive for large values of n since quantile estimation must be carried out at each horizon $s = 1, ..., h$ and each variable $i = 1, \ldots, n$. Moreover, $QIRF_1^\alpha(s)$ requires a strong condition such as Assumption 2(i) to hold for each response horizon $s = 1, ..., h$, which is often too restrictive. The second concept of QIRF denoted by $QIRF_2^\alpha(s)$ is designed to weaken such strong assumption and restrictions, which is defined as follows:

$$\begin{aligned} QIRF_2^\alpha(s) &= E\left[\frac{\partial q_{t+s,1}^{\alpha*}}{\partial \epsilon_t'} | \mathcal{F}_t\right], \\ &= \sum_{i=1}^{p} \Gamma_{1,i}^{\alpha*} E\left[\frac{\partial y_{t+s-i-1}}{\partial \epsilon_t'} | \mathcal{F}_t\right] + \Psi_1^{\alpha*} E\left[\frac{\partial x_{t+s-1}}{\partial \epsilon_t'} | \mathcal{F}_t\right], \end{aligned} \tag{12}$$

where $E\left[\frac{\partial y_{t+s-i}}{\partial \epsilon_t'} | \mathcal{F}_t\right]$ is by definition the MIRF.

As is evident in (12), $QIRF_2^\alpha(s)$ is based on the quantile at $t+s$ conditional on the information set at $t+s-1$. Intuitively speaking, at each s, $QRF_2^\alpha(s)$ captures the change in distribution occurring

at $t+s$ whereas $QIRF_1^{\alpha}(s)$ tracks the aggregate change in distribution between $t+1$ and $t+s$ for each s. For this reason, $QIRF_2^{\alpha}(s)$ does not require strong conditions such as Assumption 2(i) required for $QIRF_1^{\alpha}(s)$, and does not need to be carried out for each response horizon $s = 1, ..., h$; that is, a single estimation of $q_{t,1}^{\alpha *}$ is sufficient. The concept of $QIRF_2^{\alpha}(s)$ is analoguous to that of the generalized impulse response function of Pesaran and Shin (1998) in that both methods compute the expectation of the change of a variable after a shock.

As noted in the previous section, the concept of QIRF is able to capture the so-called asymmetric response of a variable to economic shocks. For example, consider the quantile impulse response of $y_{i,t+s}$ $(s = 1, 2, ..., h)$ when an impulse is given to ϵ_{jt}. A positive monetary policy shock can make $y_{i,t+s}$ smaller in dispersive order in the sense of Shaked and Shanthikumar (2006), while a negative shock can make it larger. That is, a positive shock shrinks the distribution of $y_{i,t+s}$ given $\mathcal{F}_t$ or $\mathcal{F}_{t+s-1}$, while a negative shock can increase the spread of the whole conditional distribution possibly in an asymmetric manner. Hence, the QIRF is not necessarily symmetric whereas the conventional MIRF is symmetric even in this example. For the sake of illustration, we display an example graphically in Figure 1. In each sub-figure, the boundaries of different shades represent 0.2, 0.4, 0.6, and 0.8 quantiles from left to right, respectively. Figure 1(b) shows that a positive shock reduces the spread of the distribution after the mean shifting. For example, the distance between 0.2^{th} and 0.8^{th} quantiles changes from 1.7 to 0.8. On the other hand, a negative shock increases the spread of the distribution so that the distance becomes 2.6 as shown in Figure 1(d).

To examine a possible asymmetric pattern in QIRF, we conduct Monte Carlo simulations using a bivariate VAR(1) model with heteroscedastic errors as in (1) and (6) with $y_t = (y_{1,t}, y_{2,t})'$ and $\epsilon_t = (\epsilon_{1,t}, \epsilon_{2,t})'$. The structural identification condition is such that A_0 is a lower triangular matrix specified as $A_0 = \begin{pmatrix} 1 & 0 \\ -0.5 & 1 \end{pmatrix}$. The coefficient matrices A_1 and Γ_1 for the bivariate VAR(1) are

set to $A_1 = \begin{pmatrix} 0.4 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}$, and $\Gamma_1 = \begin{pmatrix} 0.3 & -0.2 \\ 0.2 & 0.3 \end{pmatrix}$, respectively. . Once y_t's are generated through these specifications, we compute $QIRF_2^{\alpha}(s)$ for five quantile indexes ($\alpha = 0.1,\ 0.3,\ 0.5,\ 0.7$, and 0.9). The results are shown in Figure 2. In each figure, there are five lines for those selected quantile indexes and each line traces how the corresponding quantile response changes after a shock. When $QIRF_2^{\alpha}(s)$ for five different values of α is graphed against s, all the five lines should start at the same point (i.e., zero) when $s = 0$ because there is no shock when $s = 0$. However, just for easing the comparison of the five lines, the initial starting points are separated based on the corresponding quantiles of the standard normal distribution. For example, $QIRF_2^{\alpha=0.1}(s)$ starts at the 0.1^{th} quantile of the standard normal distribution. It is also noted that each quantile line converges to its starting level when the effect of the shocks disappears. If the distance between the five lines becomes wider after a shock, it implies that the shock increases the spread of the conditional distribution, and vice versa. As shown in Figure 2, the spread of the conditional distribution of $y_{1,t}$ decreases after a positive shock in $\epsilon_{2,t}$, while a negative shock in $\epsilon_{2,t}$ increases the dispersion of the conditional distribution. The spread of the conditional distribution of $y_{2,t}$ tends to move in the opposite direction.

4 Estimation

To compute the quantile impulse response functions discussed in the previous section, we need to estimate both the mean parameters in (1) and the quantile parameters in (4). Since the conditional mean coefficient matrix $A(L)$ can be estimated using any existing consistent estimation method under Assumption 1, this section focuses on quantile estimation. For a particular quantile index α and a specific horizon s, the set of coefficients to be estimated is $\theta_s^{\alpha*} := (\theta_{1,s}^{\alpha*\prime}, ..., \theta_{n,s}^{\alpha*\prime})'$ where

$\theta_{i,s}^{\alpha*} = (\gamma_{i,s,1}^{\alpha*\prime}, \ldots, \gamma_{i,s,p}^{\alpha*\prime}, \phi_{i,s}^{\alpha*\prime})'$ and $\gamma_{i,s,j}^{\alpha*}$, $\phi_{i,s}^{\alpha*}$ are given in (4).[5] We estimate $\theta_s^{\alpha*}$ using a correctly specified model. Let Θ be the relevant compact parameter space and we assume that there exists a sequence of $n \times 1$ vector functions $\{q_{t,s}^{\alpha}(\theta) : \theta \in \Theta\}$ such that for each s and t, the function $q_{t,s}^{\alpha}(\theta)$ for $\theta \in \Theta$ is specified as follows:

$$q_{t,s}^{\alpha}(\theta) = \sum_{j=0}^{p} \Gamma_{s,j}^{\alpha} y_{t-s-j} + \Psi_s^{\alpha} x_{t-s}, \tag{13}$$

where θ is defined analogously with $\theta_s^{\alpha*}$ but using $\Gamma_{s,j}^{\alpha}$ and Ψ_s^{α} which have the same dimensions as $\Gamma_{s,j}^{\alpha*}$ and $\Psi_s^{\alpha*}$ in (5), respectively.

Next, we provide the correct specification condition; that is, the model in (13) is correctly specified which means that the true parameter vector $\theta_s^{\alpha*}$ belongs to the parameter space Θ.

Assumption 3. The true parameter vector $\theta_s^{\alpha*}$ belongs to the interior of a compact parameter space Θ such that for each s and t, we have the following:

$$q_{t,s}^{\alpha*} = q_{t,s}^{\alpha}(\theta_s^{\alpha*}). \tag{14}$$

For notational convenience, we suppress the dependency on s hereafter unless it is required to clarify the notations. For example, $q_{t,s}^{\alpha}(\theta)$ is denoted as $q_t^{\alpha}(\theta)$. Let $\nabla_\theta q_t^{\alpha}$ be the gradient of $q_t^{\alpha}(\theta)$. If x_t is weakly exogenous, $\nabla_\theta q_t^{\alpha}$ is simply $I_n \otimes w_t$ where $w_t = (y_{t-s}', y_{t-s-1}', \ldots, y_{t-s-(p-1)}', x_{t-s}')'$. Define the $n \times 1$ vector $\rho_t^{\alpha}(\theta)$ where the i^{th} element of $\rho_t^{\alpha}(\theta)$ is $1_{\{y_{i,t} < q_{i,t,s}^{\alpha}(\theta)\}} - \alpha$. Then, it can be shown that the following moment condition is satisfied.

$$E[\nabla_\theta q_t^{\alpha*\prime} \Omega_t \rho_t^{\alpha*}] = 0, \tag{15}$$

where $\nabla_\theta q_t^{\alpha*} = \nabla_\theta q_t^{\alpha}(\theta_s^{\alpha*})$, $\rho_t^{\alpha*} = \rho_t^{\alpha}(\theta_s^{\alpha*})$, and $\Omega_t \in \mathcal{F}_{t-s}$ is a $n \times n$ non-singular positive definite matrix of the weight function. In this paper we consider estimators that make the sample

[5] Computing $QIRF_2^{\alpha}(s)$ requires n equations to be estimated while $QIRF_1^{\alpha}(s)$ needs nh equations. The latter can be computationally intensive.

counterpart of (15) close to zero; that is, estimators satisfying the following condition

$$m_T(\hat{\theta}_T^\alpha) = \frac{1}{T}\sum_{t=1}^{T} \nabla_\theta q_t^{\alpha\prime}(\hat{\theta}_T^\alpha)\Omega_t \rho_t^\alpha(\hat{\theta}_T^\alpha) = o_p(\frac{1}{\sqrt{T}}). \tag{16}$$

Existing quantile estimators can be considered as special cases of the estimator $\hat{\theta}_T^\alpha$ obtained from (16) with different choices of Ω_t. For example, in cross-section models with *iid* variables, the condition in (16) can be viewed as the first order condition of the multivariate quantile regression estimator of Chaudhuri (1996) and the univariate median regression estimator of Zhao (2001) if $\Omega_t = I_n$ and $\Omega_t = f_{i,t|t-s}(q_{i,t}^{\alpha*})$, respectively. Using $\Omega_t = F_t T_t^{-1}$ where $F_t = \oplus_{i=1}^n f_{i,t|t-s}(q_{i,t}^{\alpha*})$ and $T_t = E[\rho_t^{\alpha*}\rho_t^{\alpha*\prime}]$ will result in the efficient seemingly unrelated quantile estimator of Jun and Pinske (2009). In time-series models with non-*iid* variables, using the identity matrix for Ω_t is equivalent to the case of the QMLE of White et al. (2015). We also note that the univariate efficient semiparametric estimator of Komunjer and Vuong (2010) is considered as the univariate version of Jun and Pinske (2009).

The following proposition whose proof is based on the idea of Huber (1967) provides the asymptotic properties of the estimator defined in (16).

Proposition 1. Suppose that (i) Ω_t is known and (ii) an estimator $\hat{\theta}_T^\alpha$ satisfies (16). Under Assumptions 1 through 3,

$$\hat{\theta}_T^\alpha \xrightarrow{p} \theta^{\alpha*},$$

$$\sqrt{T}(\hat{\theta}_T^\alpha - \theta^{\alpha*}) \xrightarrow{d} N(0, Q^{-1}VQ^{-1}),$$

where $Q = E[\nabla_\theta q_t^{\alpha*\prime}\Omega_t F_t \nabla_\theta q_t^{\alpha*}]$, $V = E[\nabla_\theta q_t^{\alpha*\prime}\Omega_t T_t \Omega_t \nabla_\theta q_t^{\alpha*}]$, $F_t = \oplus_{i=1}^n f_{i,t|t-s}(q_{i,t}^{\alpha*})$ and $T_t = E[\rho_t^{\alpha*}\rho_t^{\alpha*\prime}]$.

All the technical proofs are provided in the Mathematical Appendix A. The efficiency of the estimator $\hat{\theta}_T^\alpha$ depends on the choice of Ω_t. If Ω_t is a diagonal matrix such as the identity matrix,

the estimator is basically equivalent to what is obtained by estimating each equation separately by regression quantile. In that case, we lose efficiency, analogously to the SUR set-up in OLS regression, if the elements of $\rho_t^{\alpha*}$ are correlated. As noted in Section 2, our model eventually considers multiple quantiles, although our notation uses a single index α for clear presentation. In such a general multi-quantile case, the vector $\rho_t^{\alpha*}$ contains the check functions of the different quantile levels of the same $y_{i,t}$, which is likely to cause high correlation between the elements of $\rho_t^{\alpha*}$.

Since efficiency loss caused by such correlation can be substantial, one can consider the choice of Jun and Pinkse (2009) by setting $\Omega_t = F_t T_t^{-1}$ where $F_t = \oplus_{i=1}^n f_{i,t|t-s}(q_{i,t}^{\alpha*})$ and $T_t = E[\rho_t^* \rho_t^{*\prime}]$. Such a choice of Ω_t denoted as Ω_t^E can produce an efficient estimator. However, the estimation procedure of Jun and Pinkse (2009) has potentially poor finite sample performance and is not applicable to serially dependent series such as ours. Hence, we suggest using a direct GMM estimation method with kernel density estimators. The estimation procedure can be carried out in two steps. The first step is an initial estimation stage to obtain a preliminary proxy estimator for Ω_t^E. In the first step, the true parameter can be estimated by any consistent estimation method such as single equation-by-equation quantile regression or QMLE depending on the property of x_t. The conditional density $f_{i,t|t-s}(\cdot) = f_{i,t}(\cdot|w_t)$ can be estimated using the traditional methods such as Powel (1984) and White et al. (2015). Or it can be directly estimated using the data set by the standard kernel method as follows :

$$\hat{f}_{i,t}(y_i|w) = \frac{1_{[b_T < \hat{f}(w)]} \hat{f}(y_i, w)}{\hat{f}(w)},$$

where

$$\hat{f}(y_i, w) = \frac{1}{Th_T^{k+1}} \sum_{t=1}^{T} K_1(\frac{y_i - y_{i,t}}{h_T}) K_k(\frac{w - w_t}{h_T}), \quad (17)$$

$$\hat{f}(w) = \frac{1}{Th_T^{k}} \sum_{t=1}^{T} K_k(\frac{w - w_t}{h_T}).$$

Note that $K_i(\epsilon)$ is a kernel with $\epsilon \in R^i$, , h_T is a positive bandwidth, k is the dimension of w_t and b_T is a sequence of positive constants designed to eliminate the aberrant behavior of kernel estimators for the conditional distribution (density) in regions where $\hat{f}(w)$ is small. The proxy estimator of Ω_t^E, denoted by $\hat{\Omega}_t^E$, is computed using the first step estimator $\hat{\rho}_t^\alpha \equiv \rho_t^\alpha(\tilde{\theta}_T^\alpha)$ and $\hat{f}_{i,t}(y_i|w)$ where $\tilde{\theta}_T^\alpha$ is any first-stage consistent estimator for $\theta^{\alpha *}$.

The second step for the GMM estimation method is to estimate $\theta^{\alpha *}$ based on (16) using $\hat{\Omega}_t^E$. Specifically, we obtain the GMM estimator of $\theta^{\alpha *}$ by minimizing the following objective function:

$$L_T = T m_T^E(\theta)' \left[\hat{Q}^E\right]^{-1} m_T^E(\theta), \quad (18)$$

where $m_T^E(\theta) = \frac{1}{T} \sum_{t=1}^{T} \nabla_\theta q_t^{\alpha\prime}(\theta) \hat{\Omega}_t^E \rho_t^\alpha(\theta)$, and $\hat{Q}^E$ is a consistent estimator of $Q^E = Var(\sqrt{T} m_T^E(\theta^*)) = E[\nabla_\theta q_t^{\alpha *\prime} F_t T_t^{-1} F_t \nabla_\theta q_t^{\alpha *}]$.

A typical GMM estimation method often leads to computational difficulties because the check function $\rho_t^\alpha(\theta)$ generally yields too many local non-convex regions. To tackle such a problem, we employ the Laplace-type Estimator (LTE) of Chernozhukov and Hong (2003) which is relatively easy to compute and is shown to circumvent the curse of dimensionality which our VAR set-up might have. This method is basically equivalent to the Markov Chain Monte Carlo (MCMC) approach but uses the quasi-posterior distribution function which is defined as

$$p_T = \frac{e^{L_T} \pi(\theta)}{\int e^{L_T} \pi(\theta) d\theta}, \quad (19)$$

where $\pi(\theta)$ is a prior distribution function. The detailed estimation procedure to obtain the LTE is explained in Mathematical Appendix B.

Let $\hat{\theta}^E$ be the LTE which minimizes (18). Note that Proposition 1 cannot be directly used to obtain the asymptotic distribution of $\hat{\theta}^E$ because the estimated weight function $\hat{\Omega}_t^E$ is used instead of the true efficient weight function Ω_t^E. To obtain the asymptotic property of $\hat{\theta}^E$, we need additional assumptions for the density estimator. The assumption for the kernel estimator (17) is as follows.

Assumption 4. (i) $\sup_{\epsilon \in R^i} |K_i(\epsilon)| \leq C_0 < \infty$, $\int \epsilon K_i(\epsilon) d\epsilon = 0$, $\int \epsilon^2 K_i(\epsilon) d\epsilon < \infty$, $i = 1, \ k$.

(ii) $K_i(\cdot)$ has a Fourier transform $\phi_i(\cdot)$ that is absolutely integrable.

(iii) $K_1(\cdot)$ is continuously differentiable on R with derivative satisfying $\sup_{w \in R} |K_1'(w)| < \infty$.

(iv) $h_T \to 0$, $Th_T^{k+1} \to \infty$, and $(Tp_T h_T^3)^{-1} \to 0$.

While using other density estimation methods, similar assumptions such as Assumptions 5 and 7 of White et al. (2015) are needed. The following proposition provides the asymptotic distribution of the LTE $\hat{\theta}^E$ together with its asymptotic variance-covariance matrix.

Proposition 2. Suppose that (i) $(y_t, \nabla_\theta q_t^\alpha)$ is strictly stationary, and (ii) $b_T T^{\frac{\eta}{2\eta+1}} h_T^{k+\frac{1}{2\eta+1}} \to \infty$, and $b_T/(T^{1/4} h_T) \to \infty$ where $b_T = o(T^{-\frac{1}{4\eta}})$, $\eta = \lim_{m \to \infty} \frac{-\ln \sum_{i=m+1}^{\infty} \|\psi_i\|}{\ln m}$, and $\{\psi_i\}$ is the moving average coefficient of (1). Under Assumptions 1 through 4, the asymptotic distribution of $\hat{\theta}^E$ is given by

$$\sqrt{T}(\hat{\theta}^E - \theta^{\alpha *}) \Longrightarrow N(0, Q^{E^{-1}}),$$

where $Q^E = E[\nabla_\theta q_t^{\alpha *\prime} F_t T_t^{-1} F_t \nabla_\theta q_t^{\alpha *}]$.

We note that Q^E can be easily estimated using its sample counterpart $\hat{Q}^E = \frac{1}{T} \sum_{t=1}^T \nabla_\theta q_t' \hat{F}_t \hat{T}_t^{-1} \hat{F}_t \nabla_\theta q_t^\alpha$, and $\hat{T}_t^{-1} = \frac{1}{T} \sum_{t=1}^T \hat{\rho}_t \hat{\rho}_t'$.

5 Empirical Application

In this section, we apply the proposed method to demonstrate how we can explore the effects of a monetary policy shock in greater detail using the $QIRF$. Although studying such an effect can provide significant insights into the effects of monetary policy, most empirical work, inside and outside of central banks, has focused on the average effect with the assumption that contractionary and expansionary monetary policy shocks have the same effect with the opposite sign. The main purpose of our analysis is to demonstrate how the generalization of a traditional mean-based analysis to include $QIRF$s can usefully trace the effects over the whole distribution using our proposed method - illustrating the asymmetric response to policy shocks in each tail and measuring the change in the dispersion of the distributions after contractionary or expansionary monetary policy shocks.[6]

To keep the demonstration simple, the baseline VAR model is

$$A(L)y_t = \epsilon_t, \tag{20}$$

$$A(L) = A_0 + A_1 L + \ldots + A_p L^p \tag{21}$$

where $y_t = (x_t, p_t, R_t)'$ taking x_t as employment growth, p_t as the inflation rate, and R_t as a narrative based monetary policy variable derived by Romer and Romer (2004). The identifying assumption for the VAR model is that A_0 is lower triangular, which implies that policy shocks respond to aggregate employment growth and inflation, but have no contemporaneous impact on them. That is, any contemporaneous correlations between VAR disturbance to the policy variable and the indicator of aggregate production is assumed to reflect causation from other variables to the policy variable, and not the other way around.

[6] As a related issue, central banks have acknowledged the limitation of generating mean-based forecasting, and there has been an increasing attention toward density-based forecasting which can be produced by multiple regression quantiles.

The data used in estimation are monthly observations and the sample period is from January 1973 to December 2000.[7] We estimate the VAR and provide quantile impulse response functions, $QIRF_2^{\alpha}$ to illustrate our new methods in response to monetary policy shocks.

Although our model is simple in its structure, it is well known that a VAR model such as (20) may still display a "price puzzle" - a rise in the aggregate price level in response to a contractionary monetary policy shock that contradicts mainstream theory.[8] Various approaches are suggested to deal with the price puzzle, mainly focusing on isolating monetary policy shocks from the policy response to forecasts. These approaches include using a new measure of monetary policy (Romer and Romer (2004), Keating et al. (2014)), adding forecasts or a proxy of forecasts (Bernanke et al. (2005), Bhuiyan (2014)). This paper uses the narrative measure of monetary policy shocks from Romer and Romer (2004), that is relatively free of endogenous and anticipatory movements.

Our VAR model includes total non-farm employment growth, consumer price inflation, and the monetary shock measure based on Romer and Romer (2004), which is illustrated in Figure 3. The shocks implied by the narrative policy measure comove very closely with changes in the actual federal funds rate - for example both show negative shocks in periods of recession - but there are inevitably discrepancies in some periods such as 1977-1978 and 1991-1992, which may indicate that the Federal Reserve raises (decreases) the interest rate by a lesser (greater) amount than it normally would use, given its forecast of rapid expansion (recession). This does not prevent us from

[7] This avoids complications with the dotcom crisis, which occurred in 2001, to which the Federal Reserve responded by cutting the Fed funds rate eleven times and the discount rate twelve times in 2001.

[8] A traditional interpretation of the puzzle is that the federal reserve board has better inflation forecasts so that changes in the interest rate partly reflect policy response to inflation pressures. In recent decades, there have been many attempts to tackle the problem by eliminating the expected changes in the policy variable. A conventional way is to add a commodity price as a measure of information variable; see Sims (1992), Christiano et al. (1996). In other efforts, Giordani (2004) propose to use a GDP gap instead of output growth while Bernanke and Mihov (1998) suggest a linear combination of total reserves, non-borrowed reserves, and the federal funds rate as policy shocks.

using this model as a demonstration tool to show the usefulness of the quantile impulse response functions, $QIRF_2^{\alpha}$.

Let us suppose, for the purpose of illustration, that the central bank would like to know the effect of a contractionary monetary policy on employment growth and inflation. When we estimate the model we can see that it has similar responses to the original Romer and Romer (2004) model, based on the plots the mean impulse response functions (MIRFs) for employment growth (left panel), and inflation (right panel) in Figure 4. The effect of a monetary contraction (interest rates rise by 100 basis points) initially leads to an increase in employment growth, but after a few quarters, results in a negative response of employment growth as expected. The effect on inflation is more variable, but the effect is again negative. This is the information that a regular ($MIRF$) impulse response from a simple VAR model of Romer and Romer (2004) type would generate, and based this information, the central bank would infer the effect of tightening the policy on employment growth and inflation.

We now compare these results with the quantile impulse response functions, $QIRF_2^{\alpha}$. To do this, we will demonstrate the additional information that is available by reporting the deviations of the quantile estimates around the mean, that is, $QIRF_2^{\alpha} - MIRF$ in two ways, which we explain in sequence below.

First, we show that the impulse response functions for different quantiles often differs significantly from the $MIRF$, and not in a uniform way. Consider Figure 5 for employment growth under a contractionary (100 basis point) monetary policy shock. The employment growth at $\alpha = 0.1$ and 0.3 increases significantly more than $MIRF$ after initial fluctuations, and then returns to the vicinity of zero. On the contrary, the changes at $\alpha = 0.9$ and 0.7 show the movement in the opposite direction initially, and sharply increases more than $MIRF$, reaching maximum points about 0.5 and 0.2, respectively for $\alpha = 0.9$ and 0.7. Here, we can identify two distinct patterns. One pattern is that $QIRF_2^{\alpha} - MIRF$ increases (taking positive values) for low quantiles, but decreases (taking

negative values) for high quantiles. The other pattern is that $QIRF_2^{\alpha} - MIRF$ takes positive values for both low and high quantiles.

The mean impulse response function $MIRF$ can be interpreted as the change in the conditional expectation function caused by an external shock. Notationally, $MIRF = CE^S - CE^{NS}$ where CE^S is the conditional expectation with shock, while CE^{NS} is the conditional expectation with no shock. Analogously, the sample kind of interpretation can be given to $QIRF_2^{\alpha}$; that is, $QIRF_2^{\alpha} = CQ_{\alpha}^S - CQ_{\alpha}^{NS}$ where CQ_{α}^S is the conditional α^{th}-quantile with shock while CQ_{α}^{NS} is the conditional α^{th}-quantile with no shock. Hence, the first pattern in Figure 5 indicates that, for example with $\alpha = 0.1, 0.9$, $QIRF_2^{\alpha=0.1} > MIRF$ and $MIRF > QIRF_2^{\alpha=0.9}$ (when s is very small) which implies that $CQ_{\alpha=0.1}^S - CQ_{\alpha=0.1}^{NS} > CQ_{\alpha=0.9}^S - CQ_{\alpha=0.9}^{NS}$. Rearranging this inequality, one can obtain that $CQ_{\alpha=0.9}^{NS} - CQ_{\alpha=0.1}^{NS} > CQ_{\alpha=0.9}^S - CQ_{\alpha=0.1}^S$. The first term $CQ_{\alpha=0.9}^{NS} - CQ_{\alpha=0.1}^{NS}$ is the quantile distance (or range) between $\alpha = 0.1$ and $\alpha = 0.9$ of the conditional distribution before the shock, whereas $CQ_{\alpha=0.9}^S - CQ_{\alpha=0.1}^S$ is the corresponding distance of the conditional distribution after the shock.[9] Hence, the dispersion of the growth distribution tends to shrink after the shock. The same kind of calculation can be carried out for the second pattern which indicates that, for slightly higher values of s, the shock tends to make the growth distribution skewed to the right.

Figure 6 shows how the inflation distribution reacts to the same magnitude of contractionary monetary policy shock. It can be easily seen $QIRF_2^{\alpha} - MIRF$ is different from zero for some values of s. Based on the empirical evidence indicated in Figures 5 and 6, it would be misleading in this case for the central bank to assume that the effects of contractionary monetary policy on employment growth or inflation correspond at all points on the distribution to the $MIRF$ because Figures 5 and 6 show they do not. By consulting the $QIRF_2^{\alpha}$ or by comparing the response using the difference $QIRF_2^{\alpha} - MIRF$ the central bank could observe differences from the $MIRF$ while

[9] If $\alpha = 0.25$, then $CQ_{\alpha} - CQ_{1-\alpha}$ becomes the well-known interquartile range.

setting monetary policy.

Second, the differences in the responses at the upper and lower tails of the distribution can be illustrated in a single figure. Figure 7 presents $QIRF_2^{\alpha} - MIRF$ for $\alpha = 0.1, 0.3, 0.7, 0.9$ in one graph. We use the same convention as used in Figure 2. That is, all the four lines should start at the same point (i.e., zero), but the initial starting points are separated based on the quantiles of the normal distribution as in Figure 2. Each starting point is indicated by a dotted line. Hence, for example, if $QIRF_2^{\alpha} - MIRF$ becomes lower than the dotted line, it means the difference becomes negative. We present this graph for four scenarios: in the upper row we have employment growth under contractionary monetary policy (left upper panel) and expansionary policy (right upper panel), and in the lower row, we have the equivalents for inflation.

Under contractionary policy, our previous observations can be easily re-confirmed in the left panel of Figure 7. The impact of an expansionary monetary policy shock is shown in the right panel of Figure 7. It shows that the conditional quantile range of responses in employment growth expands in a way that is fairly similar (in this case) with the contraction after a tightening of monetary policy (left panel). It is not necessarily the case that this kind of symmetric response will occur, and our deviations in $QIRF_2^{\alpha} - MIRF$ can provide evidence of asymmetry between tightening and expansionary monetary policy shocks. The same kind of analysis can be undertaken for inflation.

These illustrations show that the $QIRF_2^{\alpha}$ developed in this paper provides important information on the differences in the impulse responses at points on the distribution away from the conditional mean. They show the range of distributional responses, the extent to which the conditional quantile range is widening or narrowing, and the degree of asymmetry in the effects of monetary policy shocks, which might otherwise be assumed (incorrectly, in this case) to be identical to the $MIRF$. It illustrates very clearly that a central bank can be more consistent in its use of

distributional information for the formulation of monetary policy, as well as for the communication of monetary policy.

6 Conclusion

Central banks have made use of past forecast errors or split normal densities of forecasts in order to convey the uncertainty around inflation projections for more than two decades. However, they still rely on conditional mean impulse response functions from models used to form monetary policy decisions. In this paper, we present a new and proper impulse response analysis in quantile models that ensures that the advantages of distributional information are conferred on models used for policy purposes. Our paper also resolves some restrictions in the pseudo quantile impulse response function proposed by White et al. (2015). Using a structural vector autoregression (SVAR) in the conditional mean set-up, which is used to identify a structural shock, we permit an intervention into the structural shock to affect the entire conditional distribution, from which we derive a "quantile impulse response function ($QIRF$)." This allows us to observe the effect of the shock on the entire conditional distribution of the observable structural variable via any changes to the breadth of the distribution under the shock, which measures a form of uncertainty and any asymmetry in the responses to positive and negative shocks. None of these advantages are available using impulse responses from the conditional mean function. Therefore, our methods provide researchers and policy makers with a broader perspective on the dynamics of macroeconomic variables following a shock. The new methods are applied to US monetary policy using the VAR model proposed by Romer and Romer (2004). The results demonstrate the range of distributional responses, the extent to which the conditional quantile range is widening or narrowing and the degree of asymmetry in the effects of monetary policy shocks for tightening and expanding monetary policy.

References

[1] Adrian, T., N. Byarchenko, and D. Giannone (2019): "Vulnerable growth," American Economic Review, forthcoming.

[2] Andrews, D. (1983): "First order autoregressive processes and strong mixing," Cowles Foundation Discussion Paper, 664.

[3] Andrews, D. (1995): "Nonparametric kernel estimation for semiparametric models," Econometric Theory, 11, 560-596.

[4] Barth, M. J., and V. A. Ramey (2002): "The cost channel of monetary transmission," NBER Macroeconomics Annual, 16, 199-256.

[5] Bernanke, B., J. Boivin, and P. S. Eliasz (2005): "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach," The Quarterly Journal of Economics, 120, 387-422.

[6] Bernanke, B., and I. Mihov (1998): "Measuring monetary policy," Quarterly Journal of Economic, 113, 869-902.

[7] Bhuiyan, R. (2014): "The effects of monetary policy shocks in the USA: A forecast-augmented VAR approach," Australian Economic Papers, 53, 139-152.

[8] Chavleishvili, S., and S. Manganelli (2016): "Quantile impulse response functions," working paper.

[9] Chaudhuri, P. (1996): "On a geometric notion of quantiles for multivariate data," Journal of American Statistical Association, 91, 862–872.

[10] Chernozhukov, V., and H. Hong (2003): "An MCMC approach to classical estimation," Journal of Econometrics, 115, 293-346.

[11] Cho, J., T.-H. Kim, and Y. Shin (2015): "Quantile cointegration in the autoregressive distributed-lag modeling framework," Journal of Econometrics, 188, 281-300.

[12] Christiano, L. J., M. Eichenbaum, and C. Evans (1996): "The effects of monetary policy shocks: evidence from the flow of funds," The Review of Economics and Statistics, 78, 16-34.

[13] Davidson, J. (1994): Stochastic Limit Theory: Advanced Texts in Econometrics. Oxford, London, 1st edition.

[14] Ehrmann, M., M. Ellison, and N. Valla (2003): "Regime-dependent impulse response functions in a Markov-switching vector autoregression model," Economics Letters, 78, 295-299.

[15] Engle, R.F., and S. Manganelli (2004): "CAViaR: conditional autoregressive value at risk by regression quantiles," Journal of Business & Economic Statistics, 22, 367-381.

[16] Fernandez, C., and M.F.J. Steel (1998): "On Bayesian modeling of fat tails and skewness," Journal of American Statistical Association, 93, 359-371.

[17] Fernandez, C., J. Osiewalski, and M.F.J. Steel (1995): "Modeling and inference with v-spherical distributions," Journal of the American Statistical Association, 90, 1331-1340.

[18] Galvao, A. (2009): "Unit root quantile autoregression testing using covariates," Journal of Econometrics, 152, 165–178.

[19] Galvao, A., G. Montes-Rojas, and S. Park (2009): "Quantile autoregressive distributed lag model with an application to house price returns," Oxford Bulletin of Econometrics and Statistics, 75, 307-321.

[20] Galvao, A., G. Montes-Rojas, and J. Olmo (2011): "Threshold quantile autoregressive models," Journal of Time Series Analysis, 32, 253-267.

[21] Giacomini, R., and I. Komunjer (2005): "Evaluation and combination of conditional quantile forecasts,"Journal of Business and Economic Statistics, 229, 416-431.

[22] Giordani, P. (2004): "An alternative explanation of the price puzzle," Journal of Monetary Economics, 51, 1271-1296.

[23] Granger, C., and P. Newbold (1986): Forecasting Economic Time Series. Academic Press, Orlando, second edition.

[24] Granger, C, and G. Yoon (2002): "Hidden cointegration," UCSD working paper.

[25] Greenwood-Nimmo, M, Y. Shin, T. K., and T. van Treeck (2013): "Fundamental asymmetries in US monetary policymaking: evidence from a nonlinear autoregressive distributed lag quantile regression model," working paper.

[26] Hanson, M. (2004): "The price puzzle reconsidered," Journal of Monetary Economic, 51, 1385-1513.

[27] Hatemi, A. (2014): "Asymmetric generalized impulse response with an application in finance," Economic Modelling, 36, 18-22.

[28] Huber, P. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 221–233.

[29] Jun, S. J., and J. Pinkse (2009): "Efficient semiparametric seemingly unrelated quantile regression estimation," Econometric Theory, 25, 1392–1414.

[30] Keating, J.W., Kelly, L.J., and V.J,. Valcarcel (2014): "Solving the price puzzle with an alternative indicator of monetary policy quantiles," Economic Letters, 124, 188–194.

[31] Koenker, R., and G. Bassett (1978): "Regression quantiles," Econometrica, 46, 33–50.

[32] Koenker, R., and Z. Xiao (2006): "Quantile autoregression," Journal of the American Statistical Association, 101, 980-1006.

[33] Komunjer, I. (2007): "Asymmetric power distribution: theory and applications to risk management," Journal of Applied Econometrics, 22, 891-921.

[34] Komunjer, I., and Q. Vuong (2010): "Efficient estimation in dynamic conditional quantile models," Journal of Econometrics, 157, 272-285.

[35] Koul, H., and A. Saleh (1995): "Autoregression quantiles and related rank score processes," Annals of Statistics, 23, 670-689.

[36] Li, G., Y. Li, and C. Tsai (2015): "Quantile correlations and quantile autoregressive modeling," Journal of the American Statistical Association, 110, 246-261.

[37] Mario Fornia, L. G. (2010): "The dynamic effects of monetary policy: a structural factor model approach," Journal of Monetary Economics, 57, 203–216.

[38] Pesaran, H., and Y. Shin (1998): "Generalized impulse response analysis in linear multivariate models," Economics Letters, 58, 17-29.

[39] Powell, J. L. (1984): "Least absolute deviations estimation for the censored regression model," Journal of Econometrics, 25, 303–325.

[40] Romer, C. D., and D. H. Romer (2004): "A new measure of monetary shocks: derivation and implications," American Economic Review, 94, 1055–1085.

[41] Shaked, M., and J. G. Shanthikumar (2006): Stochastic Orders. Springer, New York.

[42] Sims, C. (1992): "Interpreting the macroeconomic times series facts: The effects of monetary policy," European Economic Review, 36, 975–1000.

[43] Theodossiou, P. (1998): "Financial data and the skewed generalized t-distribution," Management Science, 44, 1650-1661.

[44] Taylor, J., and D. Bunn (1999): "A quantile regression approach to generating prediction intervals," Management Science, 45, 225–237.

[45] White, H. (2005): Asymptotic Theory for Econometricians, Academic Press, San Diego.

[46] White, H., T.-H. Kim, and S. Manganelli (2014): "Measuring codependence between financial markets using multivariate multi-quantile CAViaR," working paper.

[47] White, H., T.-H. Kim, and S. Manganelli (2015): "VAR for VaR: measuring tail dependence using multivariate regression quantiles," Journal of Econometrics, 187, 169–188.

[48] Xiao, Z. (2009): "Quantile cointegrating regression," Journal of Econometrics, 150, 248–260.

[49] Zhao, Q. (2001): "Asymptotically efficient median regression in the presence of heteroscedasticity of unknown form," Econometric Theory, 17, 765–784.

Mathematical Appendix A: Proofs

To prove Propositions 1 & 2, we need to prove the following lemma first.

Lemma 1. Suppose the model satisfies the conditions for Proposition 2, then for all $i = 1, \ldots, n$

$$\sup_{y_i, w} |\hat{f}(y_i|w) - f(y_i|w)| \to^p 0.$$

Proof of Lemma 1. Similar to the proof of Theorem 1 of Komunjer and Vuong (2010), the lemma can be proved if

$$\sup_{y_i, w} |D^{\lambda}_{y_i} \hat{f}(y_i|w) - D^{\lambda}_{y_i} \bar{f}(y_i|w)| = O_p\left(T^{-\frac{\eta}{2\eta+1}} h_T^{-k-\lambda-\frac{1}{2\eta+1}}\right) + O_p(h_T) \tag{22}$$

where $D^{\lambda}_{y_i} \hat{f}(\cdot)$ and $D^{\lambda}_{y_i} \bar{f}(\cdot)$ are λ^{th} derivative with respect to y_i, $\bar{f}(y_i|w) = f(y|w)\bar{g}(w)$, $\bar{g}(w) = \frac{1}{T}\sum_{t=1}^{T} g_t(w)$ and $g_t(\cdot)$ is the marginal density of w_t. (22) is a modification of Lemma 4 of Komunjer and Vuong (2010) so that the order is adjusted to a NED process case. We will prove $\lambda = 1$ case only. $\lambda = 0, 2$ cases are straightforward from this as is Lemma 4 of Komunjer and Vuong (2010). Using (17), the left hand side can be rewritten as

$$\begin{aligned}
&\sup_{(y,w)} \frac{1}{Th_T} \sum_{t=1}^{T} \left| K_1(\frac{y - y_t}{h_T}) K_k(\frac{w - w_t}{h_T}) - f_t(y, w) \right| = \\
&\quad \sup_{(y,w)} \frac{1}{Th_T} \sum_{t=1}^{T} \left| K_1(\frac{y - y_t}{h_T}) K_k(\frac{w - w_t}{h_T}) - E[K_1(\frac{y - y_t}{h_T}) K_k(\frac{w - w_t}{h_T})] \right| \\
&\quad + \left| E[K_1(\frac{y - y_t}{h_T}) K_k(\frac{w - w_t}{h_T})] - E[f_t(y|w) K_k(\frac{w - w_t}{h_T})] \right| + \left| E[f_t(y|w) K_k(\frac{w - w_t}{h_T}) - f_t(y, w) \right|
\end{aligned}$$

The proofs of the second and the third term are equivalent to those of Lemma 4 of Komunjer and Vuong (2010) which are $O_p(h_T)$ and $O_p(h_T^k)$, respectively. Thus, we have only to show that the first term is $O_p\left(T^{-\frac{\eta}{2\eta+1}} h_T^{-k-\lambda-\frac{1}{2\eta+1}}\right)$. Similar to (A.10) of Andrews (1995), this can be proved if

$$E[\sup_{(h \le h_T, y, w)} \frac{1}{Th_T} \sum_{t=1}^{T} \left| K_1(\frac{y - y_t}{h_T}) K_k(\frac{w - w_t}{h}) - E[K_1(\frac{y - y_t}{h_T}) K_k(\frac{w - w_t}{h_T})] \right|] = O_p\left(T^{-\frac{\eta}{2\eta+1}} h_T^{-k-\lambda-\frac{1}{2\eta+1}}\right) \tag{23}$$

Using the Fourier inversion theorem such that $K_k(\frac{w-w_t}{h_T}) = \int \exp(-iv'(w-w_t)/h_T\phi_i(v)dv$ and Assumption 4 (ii)

$$\sup_{(y,w)}\left|\frac{1}{Th_T^{k+1}}\sum_{t=1}^{T}K_1(\frac{y-y_t}{h_T})K_k(\frac{w-w_t}{h_T}) - \frac{1}{Th_T^{k+1}}\sum_{t=1}^{T}E[K_1(\frac{y-y_t}{h_T})K_k(\frac{w-w_t}{h_T})]\right| \tag{24}$$

$$\leq \int \sup_{(y,w)}\left|\frac{1}{Th_T^{k}}\sum_{t=1}^{T}\left[K_1(\frac{y-y_t}{h_T})\int \exp(-iv'(w-w_t) - E\{K_1(\frac{y-y_t}{h_T})\int \exp(-iv'(w-w_t)\}\right]\phi_i(h_T^k v)\right| dv$$

$$\leq \int \sup_{y}\left|\frac{1}{Th_T^{k}}\sum_{t=1}^{T}\left[K_1(\frac{y-y_t}{h_T})\int \exp(iv'w_t) - E\{K_1(\frac{y-y_t}{h_T})\int \exp(iv'w_t)\}\right]\phi_i(h_T^k v)\right| dv$$

$$= \int \sup_{y}\left|\frac{1}{Th_T^{k}}\sum_{t=1}^{T}\left[\left(K_1(\frac{y-y_t}{h_T})\cos(v'w_t) - E[K_1(\frac{y-y_t}{h_T})\cos(v'w_t)]\right)\right]\right|$$

$$+\left|\frac{1}{Th_T^{k}}\sum_{t=1}^{T}\left[i(K_1(\frac{y-y_t}{h_T})\sin(v'w_t) - E[K_1(\frac{y-y_t}{h_T})\sin(v'w_t)])\right]\right|\phi_i(h_T^k v)dv$$

(25)

Then, similar to (A.13) of Andrews (1995), equation (24) can be proved if there exist bounded constants C_0^*, C_1^*, and C_2^* such that

$$E[\frac{1}{T}\sum_{t=1}^{T}\left|\left[\left(K_1(\frac{y-y_t}{h_T})\cos(v'w_t) - E[K_1(\frac{y-y_t}{h_T})\cos(v'w_t)]\right)\right]\right| < T^{-\frac{\eta}{2\eta+1}}h_T^{-\frac{1}{2\eta+1}}\left[C_0^* + \|vh_T\|C_1^* + C_2^*\right].$$

Let y_t^m and w_t^m be $E[y_t|F_{t-m}^t]$ and $E[w_t|F_{t-m}^t]$, respectively, where $F_{\epsilon,t-m}^t$ is the $\sigma-$field generated by $(\epsilon_{t-m},\ldots,\epsilon_t,x_t')$. Note that

$$K_1(\frac{y-y_t}{h})\cos(v'w_t) - E[K_1(\frac{y-y_t}{h})\cos(v'w_t)] = \left(K_1(\frac{y-y_t}{h}) - K_1(\frac{y-y_t^m}{h})\cos(v'w_t)\right)$$

$$+\left(K_1(\frac{y-y_t^m}{h})\{\cos(v'w_t) - \cos(v'w_t^m)\}\right) + \left(K_1(\frac{y-y_t^m}{h})\cos(v'w_t^m) - E[K_1(\frac{y-y_t^m}{h})\cos(v'w_t^m)]\right)$$

$$+\left(E[K_1(\frac{y-y_t^m}{h})\{\cos(v'w_t^m) - \cos(v'w_t)\}]\right) + \left(E[\{K_1(\frac{y-y_t^m}{h}) - K_1(\frac{y-y_t}{h})\}\cos(v'w_t)]\right)$$

$$= a_t + b_t + c_t + d_t + e_t$$

(26)

Assumption 1 (i) implies that there exists a sequence of absolutely summable $\{\psi_i\}$ such that $y_t = \mu_y + \sum_{i=1}^{\infty} \psi_i \epsilon_{t-i}$. Then,

$$\sup_t E[\|y_t - y_t^m\|] = \sup_t E[\| \sum_{t=m+1}^{\infty} \psi_i \epsilon_{t-i}\|] \leq \sum_{t=m+1}^{\infty} \|\psi_i\| \|Var(\epsilon_{t-i})\|] \to O_p(m^{-\eta}) \tag{27}$$

where η is as defined in the proposition. w_t has the same property by Assumption 2(iii). Then, $\frac{1}{T}\sum E|a_t| \leq C_0 M_{0T}$ and $\frac{1}{T}\sum E|e_t| \leq C_0 M_{0T}$ for a bounded constant C_0 and $M_{0T} = O_p(h_T^{-1} m^{-\eta})$ because $\cos(\cdot) \leq 1$, $\nabla_y K_1(\cdot)$ is bounded by construction, and $K_1(\frac{y-y_t}{h_T}) - K_1(\frac{y-y_t^m}{h_T}) = \nabla_y K_1(\frac{y_t - \bar{y}}{h_T})(\frac{y_t - y_t^m}{h_T}) = C_0 M_{0T}$. $\frac{1}{T}\sum E|b_t| \leq \|v\| C_1 M_{1T}$ and $\frac{1}{T}\sum E|d_t| \leq \|v\| C_1 M_{1T}$ for a bounded constant C_1 and $M_{1T} = O_p(m^{-\eta})$ because $K_1(\cdot)$ is bounded and $\cos(v'w_t) - \cos(v'w_t^m) = sin(v'\bar{w})v'(w_t - w_t^m) \leq \|v\| M_{1T}$. To show the convergence of $\frac{1}{Th}\sum E\,|c_t|$, note that by Assumptions 1(iii) and 2(iii), $\{y_t^m, w_t^m\}$ is $\phi-$mixing with mixing coefficient $\phi(s-m)$ as defined in Assumption 1(iii). Then by Corollary 14.5 of Davidson (1994), $Cov(K_1(\frac{y-y_t^m}{h_T}) cos(v'w_t^m), \widetilde{K}_{1t}^m cos(v'w_t^m)) < 4C_2^r \phi(|t-u|^{(r-2)/r})$ for a bounded constant C_2. Hence, $\backslash\backslash\ Var(\frac{1}{T}\sum_{t=1}^{T} \widetilde{K}_{1t}^m cos(v'w_t^m)) \leq 8C_2 \frac{1}{T}\sum \phi(|t-m|)^{(r-2)/r} \leq C_3 \frac{m}{T}$ for a constant C_3 that depends on C_2 and $\frac{1}{T}\sum \phi(|t-m|^{(r-2)/r}$, which indicates that $\frac{1}{T}\sum E|c_t| \leq C_2(\frac{8C_3 m}{T})^{1/2}$ for bounded constants C_2 and C_3. Consequently, by choosing m as the integer part of $T^{1/(2\eta+1)} h_T^{-2/(2\eta+1)}$,

$$\begin{aligned}\frac{1}{T} E\left| \sum_{t=1}^{T} K_1(\frac{y-y_t}{h}) \cos(v'w_t) - E[K_1(\frac{y-y_t}{h}) \cos(v'w_t)] \right| &\leq C_0 M_{0T} + \|v\| C_1 M_{1T} + C_2 \left(\frac{8C_3 m}{T}\right)^{1/2} \\ &= T^{-\frac{\eta}{2\eta+1}} h_T^{-\frac{1}{2\eta+1}} [C_0^* + \|v h_T\| C_1^* + C_2^*] \end{aligned} \tag{28}$$

for bounded constants $C_0^*,\ C_1^*,\ C_2^*$. This completes the proof. ■

Proof of Proposition 1. We first establish the consistency result. Let $m(\theta) = \frac{1}{T}\sum_{t=1}^{T} E[\nabla_\theta q_t^{\alpha\prime}(\theta) \Omega_t \rho_t^\alpha(\theta)]$. Using triangle inequality

$$\|m(\hat{\theta}_T^\alpha)\| \leq \|m_T(\hat{\theta}_T^\alpha)\| + \sup_{\theta \in \Theta} \|m(\theta) - m_T(\theta)\| \tag{29}$$

The first term is $o_p(1)$ by (16). Note that since $\{y_t^m, w_t^m\}$ is mixing, $\nabla_\theta q_t^\alpha(\theta)'\Omega_t \rho_t^\alpha(\theta)$ is also mixing with the same mixing coefficients as $\{y_t^m, w_t^m\}$. Thus, we can apply the law of large numbers for mixing sequence [Theorem 3.47 of White (2000)] so that $\|m_T(\theta) - m(\theta)\| = o_p(1)$ for all $\theta \in \Theta$. Then, the second term is $o_p(1)$ by Glivenko-cantelli Theorem, which completes the proof by Assumption 3.

Next, we prove the asymptotic normality of the proposed estimator. Since $\hat{\theta}_T^\alpha$ satisfies the asymptotic first order condition by (16), we can apply the proof of Theorem 2 of White et al. (2015). Equations (15) through (19) if replacing $\nabla_\theta q_t^{\alpha\prime}$ by $\nabla_\theta q_t^{\alpha\prime}\Omega_t$ still satisfies assumption of White et al. (2015), denoted by WA1 through WA6. WA1 can be replaced by Assumptions 1 and 2 (iii) because WA1 is required to apply CLT for $\sqrt{T} m_T(\hat{\theta}_T^\alpha)$ and the mixing property of $\nabla_\theta q_t^{\alpha\prime}\Omega_t$ from the assumptions allow to apply appropriate central limit theorems for mixing processes. WA 2, 3, and 4 are equivalent to Assumptions 2(i), (ii), and 3. WA5 (i) and (ii) are satisfied by Assumptions 1(iii) and 2(iii). WA 5(iii) also follows from the same assumptions because Ω_t is finite. Thus, we skip the detailed proof. ■

Proof of Proposition 2. We will prove the case while using the standard kernel estimator in eq.(19). Using the other density estimator will be similar. The proposition can be proved by showing that Assumptions for Theorem 2 of Chernozhukov and Hong (2003) (denoted CA1 to CA4) hold in our set-up. Assumption 3(ii) indicates CA1. Chernozhukov and Hong (2003) state that a quadratic function with the prior in our set-up satisfies CA2. To prove CA3, note that $\frac{1}{T}\sum_{t=1}^T \nabla_\theta q_t^\alpha \hat{\Omega}^E \rho_t^\alpha(\theta) = \frac{1}{T}\sum_{t=1}^T \nabla_\theta q_t^{\alpha\prime}(\hat{\Omega}^E - \Omega^E)\rho_t^\alpha(\theta) + \frac{1}{T}\sum_{t=1}^T \nabla_\theta q_t^{\alpha\prime}\Omega^E \rho_t^\alpha(\theta)$ of which the first term is $o_p(1)$ by Lemma 1, $E\|\nabla_\theta q_t^\alpha\|^2 < \infty$, and the bounded $\rho_t^\alpha(\theta)$.Also, $(y_t, \nabla_\theta q_t^\alpha)$ are NED of size η and it can be easily shown that $\nabla_\theta q_t^{\alpha\prime}\Omega^E \rho_t^\alpha(\theta)$ satisfies Lipschitz condition. Thus, by Theorem 17.12 of Davidson (1994) $\nabla_\theta q_t^{\alpha\prime}\Omega^E \rho_t^\alpha(\theta)$ is also NED of size η, and, together with

Lemma 1, $E\|\nabla_\theta q_t^\alpha\|^2 < \infty$, and $\min\left[\frac{4\eta+2}{3\eta+1}, \frac{8r-9}{7r-9}\right] > 1$, we can apply Theorem 20.19 of Davidson (1994) to obtain that $\frac{1}{T}L_T(\theta) - \frac{1}{T}L_T^0(\theta) \to 0$ in probability uniformly over Θ, where $L_T^0(\theta) = Tm_T^0(\theta)'^E m_T^0(\theta)$ and $m_T^0(\theta) = \frac{1}{T}\sum_{t=1}^T E[\nabla_\theta q_t^{\alpha\prime}\Omega^E \rho_t^\alpha(\theta)]$. Also $L_T^0(\theta)$ is positive except $m_T^0(\theta) = 0$ and by Assumption 3(ii), $m_T^0(\theta) = 0$ if and only if $\theta = \theta_0$. Thus, CA3 is satisfied by Lemma 1 of Chernozhukov and Hong (2003). To prove CA4, we verify that our set-up satisfies Conditions (i) through (iii) of Lemma 2 of Chernozhukov and Hong (2003). L_T and L_T^0 are twice continuously differentiable, which satisfies (i). Since $(y_t, \nabla_\theta q_t^\alpha)$ is NED of size η on $\{\epsilon_t\}$ or on $\{\epsilon_t, z_t\}$, for any vector ι with $\iota'\iota = 1$, the sequence $\iota' L_s^{E-1} m_t^E(\theta^*)$ with $L_s^E L_s^{E\prime} = Q_s^E$ satisfies 24.6(a), (b), and 24.7(c'), (d') of Davidson (1994) for a bounded constant $c_{nt} = c < \infty$. Then, by Corollary 24.7 and Theorem 25.6 of Davidson (1994), $\frac{1}{\sqrt{T}} m_t^E(\theta^*) \Rightarrow N(0, Q_s^E)$ which verifies (ii). To check (iii), let us define

$$r_T(\theta_1, \theta_2) = \frac{\|\hat{m}^E(\theta_1) - \hat{m}^E(\theta_2) - \nabla_T(\theta_1 - \theta_2)\|}{\|\theta_1 - \theta_2\|} \tag{30}$$

where $\nabla_T = \frac{1}{T}\sum \nabla_t$, $\nabla_t = [(q_t^\alpha(\theta_1) - q_t^\alpha(\theta_2))\delta_t(\theta_2)]/\|\theta_1 - \theta_2\|$ and $\delta_t(\cdot)$ is the diagonal matrix of which the diagonal elements are dirac delta function. Note that for any $\theta_1, \theta_2 \in \Theta$

$$\begin{aligned}
&\frac{\sqrt{T}\|\left(\hat{m}^E(\theta_1) - \hat{m}^E(\theta_2)\right) - \left(E[m^E(\theta_1)] - E[m^E(\theta_2)]\right)\|}{1 + \sqrt{T}\|\theta_1 - \theta_2\|} \\
&\leq \frac{\sqrt{T}\left[\|r_t - E[r_t]\|\|\theta_1 - \theta_2\| + \nabla_t - E[\nabla_T]\|\|\theta_1 - \theta_2|\right]}{1 + \sqrt{T}\|\theta_1 - \theta_2\|} \\
&\leq \|\nabla_t - E[\nabla_T]\| + O_p(r_t) = O_p(r_t)
\end{aligned} \tag{31}$$

Thus, (iii) can be satisfied if there exists $\epsilon > 0$ and $\eta > 0$ such that $P[\|r_t\| > \epsilon] < \eta$. Let $\epsilon_t(\theta_i) = y_t - q_t^\alpha(\theta_i)$, $e_t = q_t^\alpha(\theta_1) - q_t^\alpha(\theta_2)$, and r_t as

$$
\begin{aligned}
r_t &= \|\nabla_\theta q_t^{\alpha\prime}\hat{\Omega}_t^E \rho_t^\alpha(\theta_1) - \nabla_\theta q_t^{\alpha\prime}\hat{\Omega}_t^E \rho_t^\alpha(\theta_2) - \nabla_\theta q_t^{\alpha\prime}\hat{\Omega}_t^E \nabla_t(\theta_1-\theta_2)\|/\|\theta_1-\theta_2\| \\
&\leq \|\hat{\Omega}_t^E\|\|1[\epsilon_t^2+e_t]-1[\epsilon_t^2]-e_t\delta_t\|/\|\theta_1-\theta_2\| \\
&\leq \|\hat{\Omega}_t^E\|\|e_t/\|\theta_1-\theta_2\|\|\|1[\epsilon_t^0+e_t]-1[\epsilon_t^0]-e_t\delta_t\|/\|e_t\| \qquad (32)
\end{aligned}
$$

Since $r_T(\theta_1,\theta_2) = \frac{1}{T}\sum r_t$ by definition, we will show that $P[|r_t|>\epsilon] < \eta$ for all t. By Assumptions 1 and 4 $\nabla_\theta q_t^{\alpha\prime}\hat{\Omega}_t = O_p(1)$. Since $q_t^\alpha(\theta)$ is differentiable, the mean value theorem and Assumption 2 implies that $\|e_t/\|\theta_1-\theta_2\|\|\|$ is also bounded in probability. Thus, we have only to show that $Pr[\|1[\epsilon_t^2+e_t]-1[\epsilon_t^2]-e_t\delta_t\|/\|e_t\|>\epsilon] < \eta$. For given $\epsilon>0$, $\eta>0$, there exists $e>0$ such that $\|e_t\|<e$ implies $Pr[\|1[\epsilon_t^0+e_t]-1[\epsilon_t^0]-e_t\delta_t\|/\|e_t\|>\epsilon]<\eta$. Since $q_t^\alpha(\theta)$ is continuous on Θ, there exist some $\iota>0$ such that $\|\theta_1-\theta_2\|<\iota$ implies $\|e_t\|<e$ which proves the inequality. ■

Mathematical Appendix B: LTE Procedure

The Laplace type estimator (LTE) is obtained by the following 4-step procedure.

Step 1. Obtain an initial consistent estimator of θ_s using conventional equation-by-equation methods. Compute $\tilde{f}_t(\cdot)$ and $\tilde{\rho}_{st}$ as described above to calculate $\tilde{T}_s = \frac{1}{T}\sum_{t=1}^T \tilde{\rho}_{st}\tilde{\rho}_{st}'$, $\tilde{V}_{st} = \tilde{F}_{st}\tilde{T}_s^{-1}$, and $\hat{H}_s$.

Step 2. Let θ_{sl} be the l^{th} element of θ_s. For each $l=1,\ldots,nkr$, generate ξ_l from $N(|\xi_l-\theta_{sl}^{(j)}|,\phi)$ where the starting value $\theta_s^{(0)}$ is the estimator in Step 1).

Step 3. Update $\theta_{sl}^{(j+1)}$ from $\theta_{sl}^{(j)}$ for $j=1,2,\ldots$ using

$$
\theta^{(j+1)} = \begin{pmatrix} \xi & \text{with probability } p(\theta^{(j)},\xi) \\ \theta^{(j)} & \text{with probability } 1-p(\theta^{(j)},\xi) \end{pmatrix}, \qquad (33)
$$

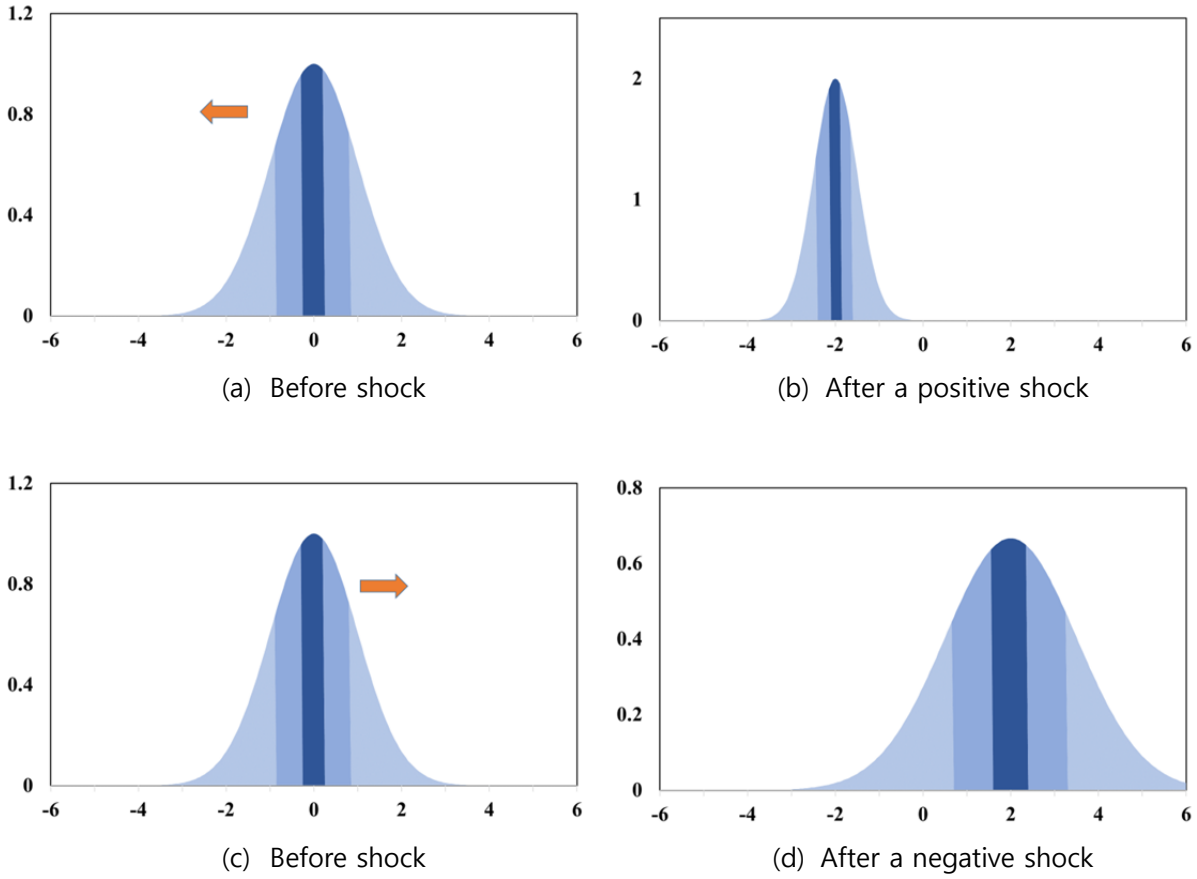
where

$$p(x,y) = inf\left(\frac{e^{L_T(y)}\pi(y)q(x|y)}{e^{L_T(x)}\pi(x)q(y|x)}, 1\right).$$

Step 4. Iterate Step 2 to Step 3 B times. The final estimator is the sample average given by

$$\hat{\theta} = \frac{1}{B}\sum_{j=1}^{B}\theta^{(j)}.$$

Note that ϕ is updated every 100 times so that the rejection rate at Step 3 is approximately 50%.



(a) Before shock (b) After a positive shock

(c) Before shock (d) After a negative shock

Figure 1. **Shifts in distribution of y_{t+s} with respect to a shock**

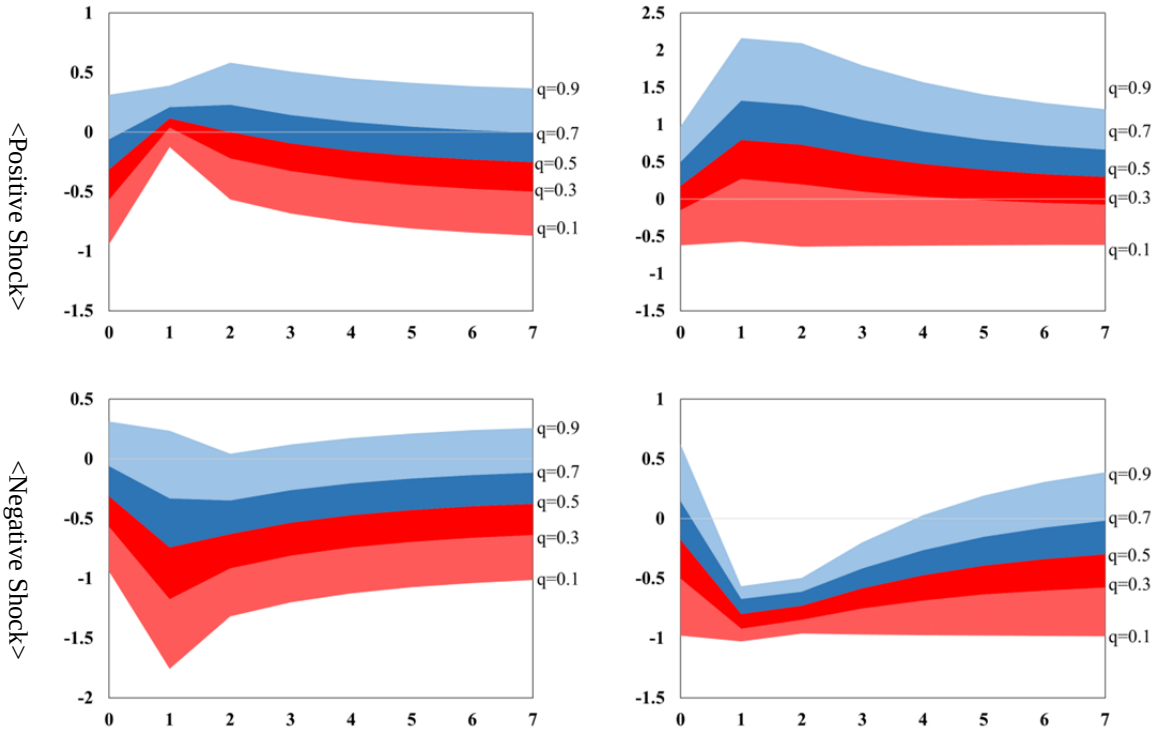
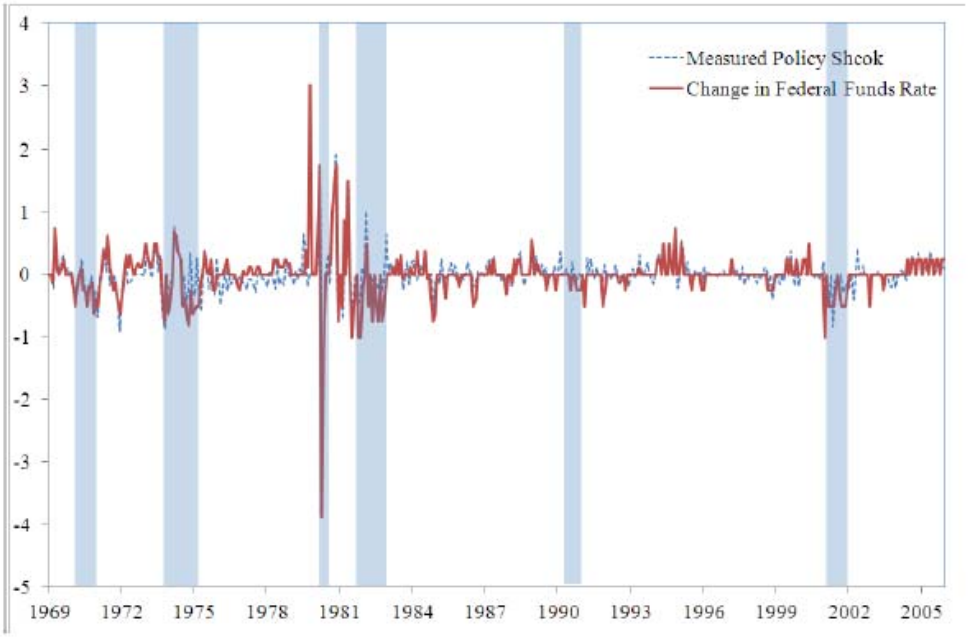
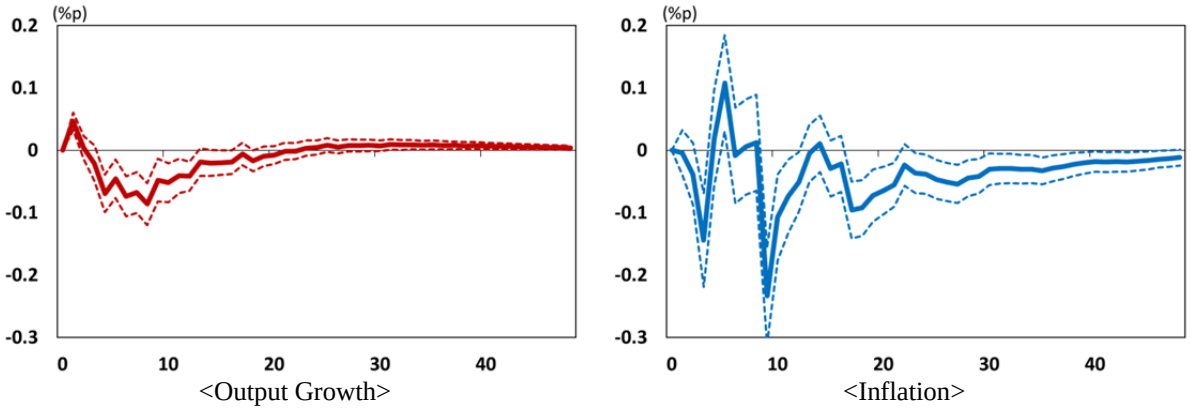


Figure 2. **Simulated quantile impulse response**



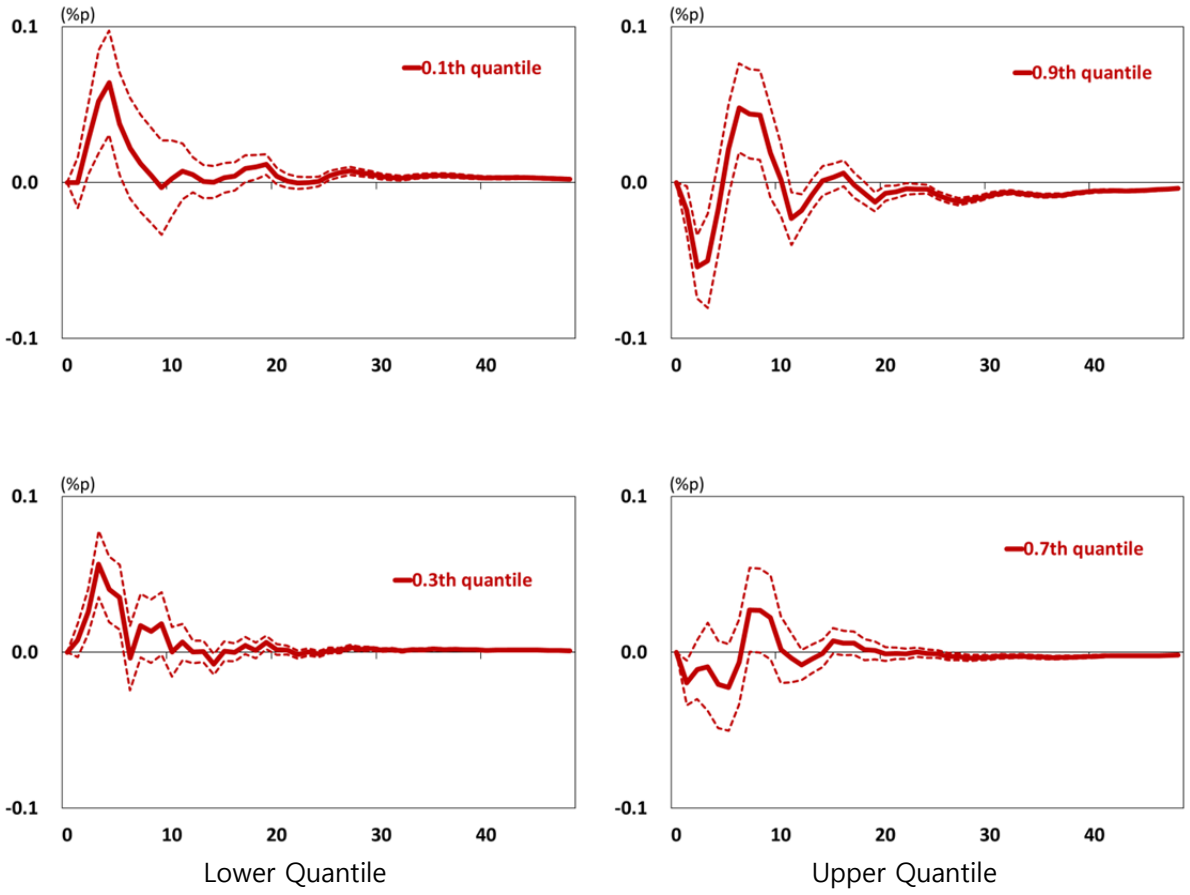
Note: Each shaded region begin at a National Bureau of Economic Research (NBER) business cycle peak, and end at a trough.

Figure 3. **Measured monetary policy shock series**



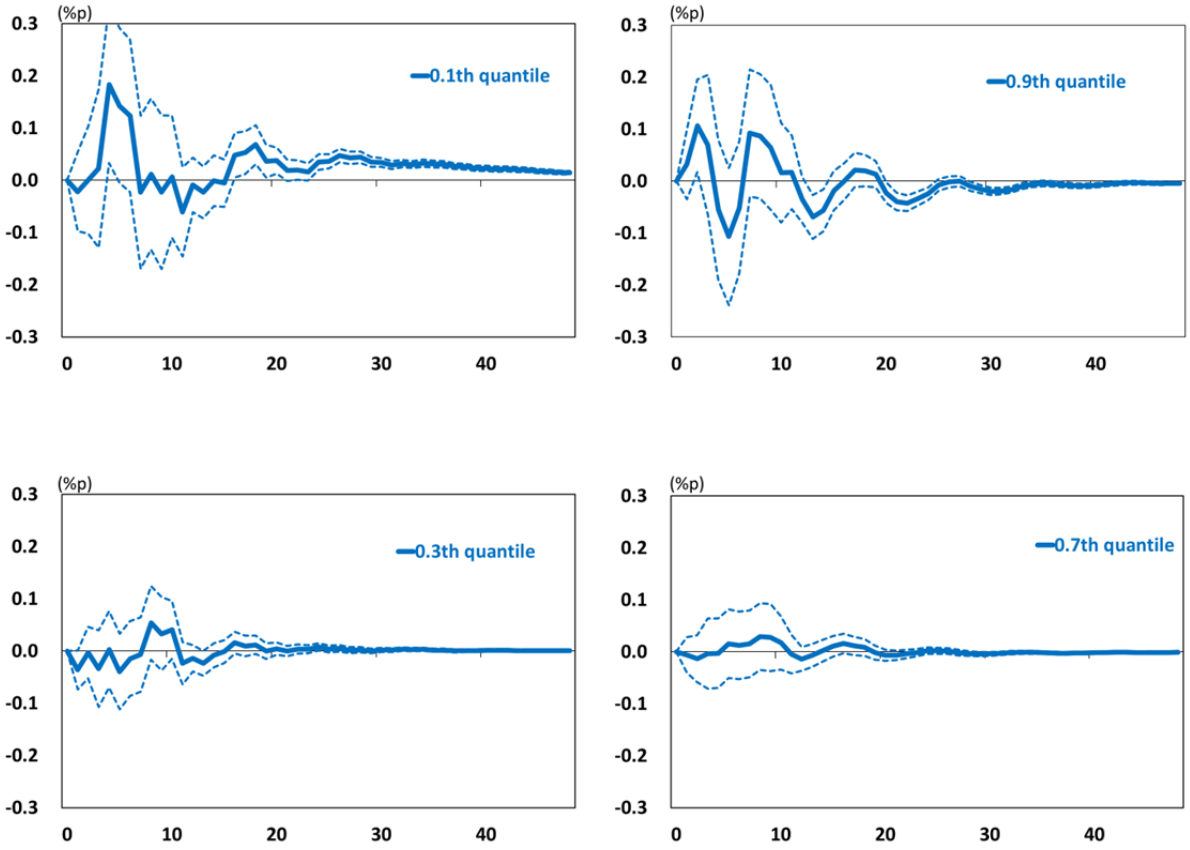
Note: (total non-farm employees, consumer price index, R&R measure of monetary policy hock). Data are from Jan. 1973 through Dec. 2000. Dotted lines represent 67% confidence interval.

Figure 4. **Mean impulse response to 100bp contractionary monetary policy shock**



Note: Difference of $QIRF_2^{\alpha}$ from MIRF. 3 month average. (total non-farm employees, consumer price index, R&R measure of monetary policy hock). Data are from Jan. 1973 through Dec. 2000. Dotted lines represent 67% confidence interval.

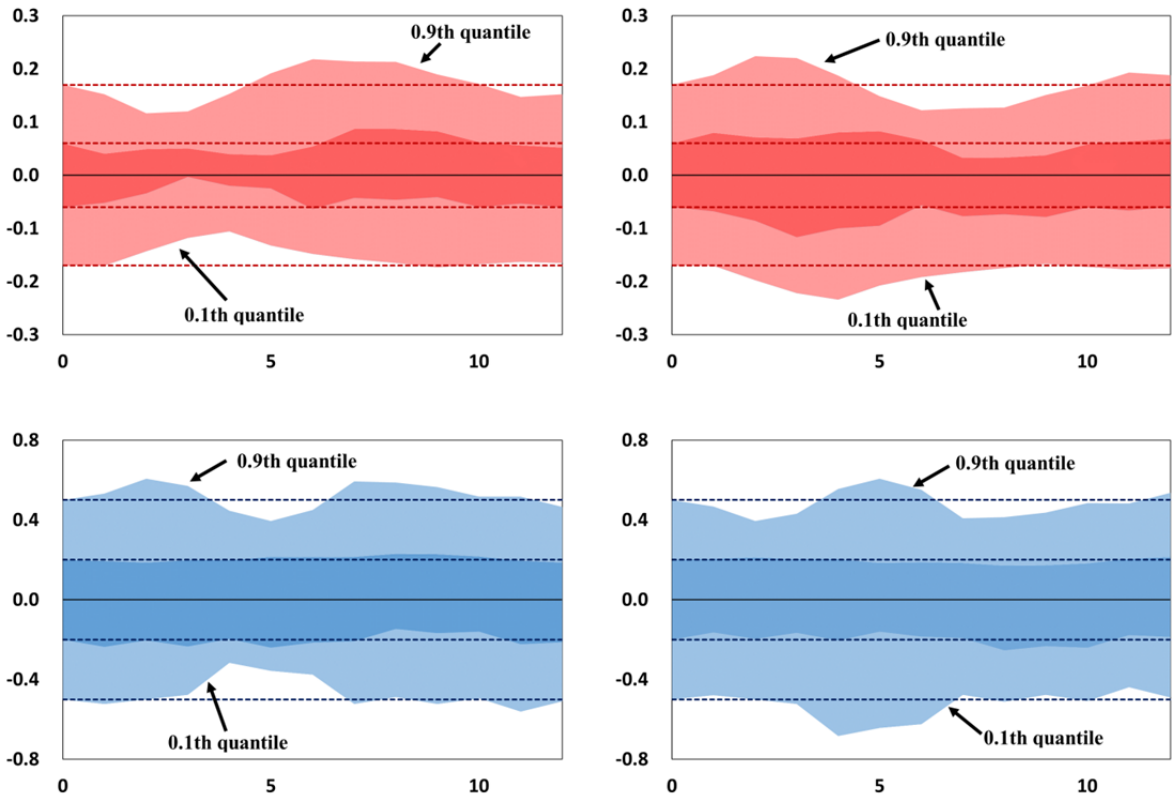
Figure 5. **Quantile impulse response of growth to a contractionary monetary policy shock** ($QIRF_2^{\alpha}$ from MIRF)



<Lower Quantiles> <Upper Quantiles>

Note: Difference of $QIRF_2^{\alpha}$ from MIRF. 3 month average. (total non-farm employees, consumer price index, R&R measure of monetary policy hock). Data are from Jan. 1973 through Dec. 2000. Dotted lines represent 67% confidence interval.

Figure 6. **Quantile impulse response of inflation to a contractionary monetary policy shock** ($QIRF_2^{\alpha}$ from MIRF)



< contractionary shock> <expansionary shock >

Note: Borders of shades represent the quantile impulse responses at 0.1, 0.3, 0.7, and 0.9, respectively. The initial point of each quantile impulse response function has been separated (for clear comparison) based on the corresponding quantiles of the standard normal distribution.

Figure 7. **Quantile impulse responses**