Quantile Local Projections:
Identification, Smooth Estimation, and Inference

Josef Ruzicka[*†]

December 29, 2024

Abstract

Standard impulse response functions measure the average effect of a shock on a response variable. However, different parts of the distribution of the response variable may react to the shock differently. A popular method to capture this heterogeneity are quantile regression local projections. We identify them by short-run restrictions or external instruments, and we establish their asymptotics. To overcome their excessive volatility, we introduce two novel smoothing estimators and propose information criteria for optimal smoothing. In the first empirical application, we show that financial conditions affect the entire distribution of GDP growth and not just its lower part. Thus, financial conditions matter not only for recessions, but also during normal times and even in recovery periods. The second application demonstrates that conventional monetary policy is more effective at curbing inflation than at generating it.

JEL classification: C14, C32, C53, C54.

Keywords: Impulse response, structural identification, quantile regression, regularization.

---

[*]Nazarbayev University, Email: josef.ruzicka@nu.edu.kz

[†]I thank Juanjo Dolado, Juan Carlos Escanciano, Stephan Fahr, Ivan Fernandez-Val, Jesús Gonzalo, Roger Koenker, Nazarii Salish, Victor Troster, Carlos Velasco, participants at EWMES 2020, RGS 2021, IAAE 2021, EEA-ESEM 2021, CFE 2022, AMES 2023, EC2 2023, and RCEA 2024. This work received the best paper award at the 2021 RGS Conference and the award for the best paper presented by a PhD student at the 2021 IAAE Conference.

1 Introduction

Quantile regression (QR) of Koenker and Bassett (1978) is more informative and more robust to outliers than the least squares regression. In macroeconomics, these advantages are crucial: the whole distribution of possible outcomes matters to risk-averse policymakers and outliers abound. As a result, QR has become a popular tool for macroeconomic data.

A key endeavor in macroeconometrics concerns dynamic causal effects of shocks—impulse response analysis. A simple and widely used method are local projections (LP) of Jordà (2005). In contrast to Structural Vector Autoregression (SVAR) of Sims (1980), LP represent impulse response functions by coefficients of horizon-specific regressions. This simplifies estimation and inference, and it enables easy incorporation of nonlinearities. Still, standard LP, which are based on least squares, measure only the average effects of shocks and they are very sensitive to outliers. On the other hand, LP estimated by QR—quantile local projections (QLP)—capture heterogeneous effects and are more robust.

However, empirical researchers who use QLP face three key challenges. First, it is unclear how to identify the dynamic causal effects—the standard SVAR identification theory does not apply. Second, the impulse responses are choppy, which is at odds with macroeconomic theory and hampers interpretation. Third, it is unclear how to construct closed-form confidence intervals.

We overcome these challenges in a coherent QLP framework, which encompasses identification, smooth estimation, and inference. First, we develop a general, nonlinear, structural identification theory for dynamic quantile causal effects of macroeconomic shocks, and we study the special cases of QLP identification by short-run restrictions and external instruments. Second, we introduce two novel estimators: Smooth Quantile Local Projections (SQLP) and Smooth Quantile Local Projections with Instruments (SQLPI). Third, we establish the asymptotic properties of these estimators, which enable us to construct closed-form confidence intervals for QLP estimators.

The first empirical application illustrates SQLP in the setup of Adrian et al. (2019). Leveraging our identification results, we add lags as controls, which alters the conclusions—we show that financial conditions affect the entire distribution of GDP growth and not just its lower part. In the second empirical application, SQLPI uncover heterogeneous effects of monetary policy.

Our estimators rely on the following principle: The quantiles of the response variable should be smooth functions of the forecast horizon. We control smoothness by roughness penalties similar to Koenker et al. (1994). However, our estimators converge at the parametric rate, unlike their estimator (Portnoy, 1997). If the penalties equal zero, the SQLP reduces to local projections estimated by QR, and the SQLPI amounts to local projections estimated by instrumental variable quantile regression (Chernozhukov and Hansen, 2008). In the ℓ^2 setting, Barnichon and Brownlees (2019) introduce smooth local projections, which merge two smoothing approaches: transformation of variables by a two-sided moving average and ridge regression shrinkage towards a polynomial. Our smoothing is the asymmetric ℓ^1 analogue of their second approach.

QLP have become popular both in academic research and in policy work. The earliest reference seems to be Distante et al. (2013), who document heterogeneous effects of technology shocks. A prominent QLP application is by Adrian et al. (2019), who use a two-step procedure. In the first step, they estimate local projections by QR. In the second step, they match three selected quantiles with those of skewed t-distribution. Our approach is different: While they impose a distributional assumption, we smooth across horizons nonparametrically. More importantly, they do not address identification and they do not provide confidence intervals of their QR estimates. A related methodology, particularly suitable for financial risk management, was proposed by Han et al. (2022), who use local projections in the MVMQ-CAViaR of White et al. (2015). However, they cover neither structural identification nor smoothing, and they find a different asymptotic variance.

The paper is structured as follows. Section 2 proposes a novel, general framework for causal identification in dynamic quantile models. Sections 3 and 4 introduce the SQLP and the SQLPI estimators, respectively. Section 5 provides two empirical applications. The main proofs are in the Appendix. The Online Appendix contains technical proofs, bootstrap inference, simulation study, forecasting application, computational aspects, and additional figures for empirical applications.

2 Causality in Dynamic Quantile Models

We propose a framework for identification of quantile treatment effects of macroeconomic shocks. The quantile treatment effect is the effect of a treatment on quantiles of an outcome variable.[2] Here, the treatment is an exogenous intervention—an impulse. Standard SVAR identification theory (see Ramey (2016)) concerns only average treatment effects, whose identification differs from quantile treatment effects. SVAR identification is neither necessary nor sufficient for identification of quantile treatment effects. Even in a univariate setting, identification of a moving average (MA) does not imply identification of quantile treatment effects. Shocks in our framework are not structural vector MA, but exogenous impulses, defined later.

Our framework allows for heterogeneous responses over quantiles not just between different time periods, but even within the same time period, unlike Chavleishvili and Manganelli (2024). For example, we allow the interest rate shock to affect the excess bond premium within the same month, but differently at different quantiles (as in our empirical application in Figure 8.)

The framework comprises a flexible stochastic process in Assumption 1, its structural representation in Lemma 1 (crucial for causal interpretation and to avoid the simultaneity bias), a counterfactual process in Assumption 2, and an impulse response function in Definition 1.

For $\tau \in (0,1)$, random variable Y, and random vector X, denote the τth quantile of Y as $Q_\tau(Y)$, and the τth conditional quantile of Y given X as $Q_\tau(Y|X)$. Throughout the paper, equations involving random variables hold with probability one.

Assumption 1 (Stationary Higher Order Markov Process)**.** *Let* $\{Y_t\}_{t\in\mathbb{Z}} = \{(y_{1t}, y_{2t}, \dots, y_{Kt})'\}_{t\in\mathbb{Z}}$ *be a* K*-variate strictly stationary stochastic process with natural filtration* $\{\mathcal{F}_t\}_{t\in\mathbb{Z}}$*. For* $P \geq 1$*, let* $\underline{Y}_{t-1} = (Y'_{t-1}, \dots, Y'_{t-P})'$*. Assume that* $\forall y \in \mathbb{R}^K$*:* $P(Y_t \leq y|\mathcal{F}_{t-1}) = P(Y_t \leq y|\underline{Y}_{t-1})$*.*

[2] For a comprehensive discussion of quantile treatment effects, see (Koenker, 2005, pp. 26-32).

Lemma 1 (Structural Quantile Form of a Stochastic Process)**.** *Under Assumption 1 there exist functions $S_1, S_2, \ldots, S_K : \mathbb{R}^{K(P+1)} \mapsto \mathbb{R}$, non-decreasing in their first argument, and independent standard uniform random variables $\{u_{1t}, u_{2t}, \ldots, u_{Kt}\}_{t\in\mathbb{Z}}$ such that:*

(i) For all $\tau \in (0,1)$,

$$
\begin{aligned}
Q_\tau\Big(y_{1t} - S_1(\tau|y_{2,t}, y_{3,t}, \ldots, y_{K,t}, \underline{Y}_{t-1})\Big|\{u_{1t}, \ldots, u_{Kt}\} \setminus u_{1t}, \mathcal{F}_{t-1}\Big) &= 0,\\
Q_\tau\Big(y_{2t} - S_2(\tau|y_{1,t}, y_{3,t}, \ldots, y_{K,t}, \underline{Y}_{t-1})\Big|\{u_{1t}, \ldots, u_{Kt}\} \setminus u_{2t}, \mathcal{F}_{t-1}\Big) &= 0,\\
&\vdots\\
Q_\tau\Big(y_{Kt} - S_K(\tau|y_{1,t}, y_{2,t}, \ldots, y_{K-1,t}, \underline{Y}_{t-1})\Big|\{u_{1t}, \ldots, u_{Kt}\} \setminus u_{Kt}, \mathcal{F}_{t-1}\Big) &= 0.
\end{aligned}
\tag{1}
$$

(ii) Y_t is the a.s. unique solution of (1).

The functions $S_1, S_2, \ldots, S_K$ are structural quantile functions in the sense of Chernozhukov and Hansen (2008), which allow endogeneity in quantile regressions. If $K > 1$, the functions $S_1, S_2, \ldots, S_K$ are not unique—that would require more assumptions, such as recursive short-run restrictions, used in linear conditional quantile models by Montes-Rojas (2022) and Chavleishvili and Manganelli (2024), but such restrictions are often unrealistic (Stock and Watson, 2001). However, Lemma 1 applies even if recursive short-run restrictions are violated, as shown in Example 1. Example 2 then illustrates the connection to univariate quantile autoregression.

Example 1 (Bivariate Non-Recursive Gaussian SVAR)**.** *Take $\{(y_{1t}, y_{2t})'\}_{t\in\mathbb{Z}}$ given by*

$$
\begin{aligned}
y_{1t} &= 2.0 + 0.2y_{2t} + 0.8y_{1,t-1} - 0.2y_{2,t-1} + e_{1t},\\
y_{2t} &= 0.5 + 0.6y_{1t} + 0.4y_{1,t-1} + 0.6y_{2,t-1} + e_{2t},
\end{aligned}
\tag{2}
$$

where $\{e_{1t}, e_{2t}\}_{t\in\mathbb{Z}}$ are independent $N(0,1)$. To represent $\{[y_{1t}, y_{2t}]\}_{t\in\mathbb{Z}}$ in the structural quantile form, denote Φ the cdf of $N(0,1)$, and let $u_{1t} = \Phi(e_{1t}), u_{2t} = \Phi(e_{2t})$. Then $\{u_{1t}, u_{2t}\}_{t\in\mathbb{Z}}$ are *independent. The probability integral transform yields $u_{1t}, u_{2t} \sim U[0,1]$. Let S_1, S_2 be given by*

$$S_1(\tau|y_{2t}, y_{1,t-1}, y_{2,t-1}) = 2.0 + 0.2y_{2t} + 0.8y_{1,t-1} - 0.2y_{2,t-1} + \Phi^{-1}(\tau),$$
$$S_2(\tau|y_{1t}, y_{1,t-1}, y_{2,t-1}) = 0.5 + 0.6y_{1t} + 0.4y_{1,t-1} + 0.6y_{2,t-1} + \Phi^{-1}(\tau).$$

Finally, we check the first equation in (1). (Checking the second equation in (1) is analogous.)

$$Q_\tau\big(y_{1t} - S_1(\tau|y_{2t}, y_{1,t-1}, y_{2,t-1})\big|u_{2t}, \mathcal{F}_{t-1}\big) = Q_\tau\big(e_{1t} - \Phi^{-1}(\tau)\big|u_{2t}, \mathcal{F}_{t-1}\big) =$$
$$Q_\tau\big(\Phi^{-1}(u_{1t}) - \Phi^{-1}(\tau)\big|u_{2t}, \mathcal{F}_{t-1}\big) = Q_\tau\big(\Phi^{-1}(u_{1t}) - \Phi^{-1}(\tau)\big) = \Phi^{-1}(\tau) - \Phi^{-1}(\tau) = 0.$$

Example 2 (Univariate Quantile Autoregression)**.** *Let $\{u_{1t}\}_{t\in\mathbb{Z}}$ be independent $U[0,1]$ and let $y_{1t} = 1.5u_{1t}y_{1,t-1} - \log(1-u_{1t})$. The process $\{y_{1t}\}_{t\in\mathbb{Z}}$ is a first-order quantile autoregression (Koenker and Xiao, 2006). It can be shown that $\{y_{1t}\}_{t\in\mathbb{Z}}$ is also AR(1) with heteroskedastic disturbances. The structural quantile form of $\{y_{1t}\}_{t\in\mathbb{Z}}$ is $S_1(\tau|y_{1,t-1}) = 1.5\tau y_{1,t-1} - \log(1-\tau)$.*

Lemma 1 implies Corollary 1, which specifies the structural data generating process.

Corollary 1 (Nonseparable Model)**.** *Under Assumption 1, using notation of Lemma 1, it holds*

$$\begin{aligned} y_{1t} &= S_1(u_{1t}|y_{2,t}, y_{3,t}, \ldots, y_{K,t}, \underline{Y}_{t-1}), \\ y_{2t} &= S_2(u_{2t}|y_{1,t}, y_{3,t}, \ldots, y_{K,t}, \underline{Y}_{t-1}), \\ &\vdots \\ y_{Kt} &= S_K(u_{Kt}|y_{1,t}, y_{2,t}, \ldots, y_{K-1,t}, \underline{Y}_{t-1}). \end{aligned} \tag{3}$$

An impulse response function has a causal interpretation only if the impulse is exogenous. It is popular to take impulses from a structural vector MA, as Lee et al. (2021). However, if one MA error shifts, the distributions of other MA errors may react, e.g. due to heteroskedasticity, putting the exogeneity of such impulses into question. We proceed differently, via Assumption 2.

Assumption 2 (Counterfactual Stochastic Process)**.** *Let* $\{Y_t\}_{t\in\mathbb{Z}}, \{u_{1t}, u_{2t}, \ldots, u_{Kt}\}_{t\in\mathbb{Z}}$*,* $S_1, S_2, \ldots, S_K$ *be as in Lemma 1. Take* $i \in \{1, 2, \ldots, K\}$*,* $s \in \mathbb{R}$ *such that* $\{Y_t\}_{t\in\mathbb{Z}}$ *has the identical support as* $\{\tilde{Y}_t\}_{t\in\mathbb{Z}} = \{(\tilde{y}_{1t}, \tilde{y}_{2t}, \ldots, \tilde{y}_{Kt})'\}_{t\in\mathbb{Z}}$ *that is given by*

$$
\begin{aligned}
\tilde{y}_{1,t} &= \mathbb{1}(t<0)y_{1t} && + \mathbb{1}(t \geq 0)S_1(u_{1,t}|\tilde{y}_{2,t}, \tilde{y}_{3,t}, \ldots, \tilde{y}_{K,t}, \underline{\tilde{Y}}_{t-1}),\\
&\vdots\\
\tilde{y}_{i-1,t} &= \mathbb{1}(t<0)y_{i-1,t} && + \mathbb{1}(t \geq 0)S_{i-1}(u_{i-1,t}|\tilde{y}_{1,t}, \ldots, \tilde{y}_{i-2,t}, \tilde{y}_{i,t}, \ldots, \tilde{y}_{K,t}, \underline{\tilde{Y}}_{t-1}),\\
\tilde{y}_{i,t} &= \mathbb{1}(t\leq 0)y_{it} + \mathbb{1}(t=0)s && + \mathbb{1}(t > 0)S_i(u_{i,t}|\tilde{y}_{1,t}, \ldots, \tilde{y}_{i-2,t}, \tilde{y}_{i,t}, \ldots, \tilde{y}_{K,t}, \underline{\tilde{Y}}_{t-1}),\\
\tilde{y}_{i+1,t} &= \mathbb{1}(t<0)y_{i+1,t} && + \mathbb{1}(t \geq 0)S_{i+1}(u_{i+1,t}|\tilde{y}_{1,t}, \ldots, \tilde{y}_{i,t}, \tilde{y}_{i+2,t}, \ldots, \tilde{y}_{K,t}, \underline{\tilde{Y}}_{t-1}),\\
&\vdots\\
\tilde{y}_{K,t} &= \mathbb{1}(t<0)y_{Kt} && + \mathbb{1}(t \geq 0)S_K(u_{K,t}|\tilde{y}_{1,t}, \tilde{y}_{2,t}, \ldots, \tilde{y}_{K-1,t}, \underline{\tilde{Y}}_{t-1}),
\end{aligned}
$$

where $\underline{\tilde{Y}}_{t-1} = (\tilde{Y}'_{t-1}, \ldots, \tilde{Y}'_{t-P})'$*. We call* $\{\tilde{Y}_t\}_{t\in\mathbb{Z}}$ *the counterfactual process.*

For $t < 0$, the counterfactual process satisfies $\tilde{Y}_t = Y_t$. At $t = 0$, the impulse hits and $\tilde{y}_{i0} = y_{i0}+s$. The other contemporaneous values are determined endogenously, and so are $\tilde{Y}_1, \tilde{Y}_2, \ldots$. By strict stationarity, the intervention period $t = 0$ is without loss of generality. The identical support of $\{Y_t\}$ and $\{\tilde{Y}_t\}$ in Assumption 2 and property (ii) in Lemma 1 ensure $\{\tilde{Y}_t\}_{t\in\mathbb{Z}}$ is well-defined. Next, we define a structural impulse response function to measure the quantile treatment effect of the ith impulse (or shock) of size s on the jth variable after h periods.

Definition 1 (Impulse Response Function)**.** *Let* $\{Y_t\}_{t\in\mathbb{Z}}, \{u_{1t}, u_{2t}, \ldots, u_{Kt}\}_{t\in\mathbb{Z}}$*,* $\{\tilde{Y}_t\}_{t\in\mathbb{Z}}$*,* i *and* s *be as in Assumption 2. Let* $j \in \{1, 2, \ldots, K\}$ *and* $\tau \in (0, 1)$*. Let* $X_{it} = \underline{Y}_{t-1}$ *or* $X_{it} = (y_{1t}, \ldots, y_{it}, \underline{Y}'_{t-1})'$*. Define the response of the* j*th variable to the* i*th impulse (or shock)* s *at quantile* τ *and horizon* $h \in \mathbb{N}_0$ *as*

$$
IR^{h,\tau,s}_{j,i}(Y_0, X_{i0}) = Q_\tau(Q_\tau(\tilde{y}_{jh}|u_{i0}, X_{i0})|Y_0, \underline{Y}_{-1}) - Q_\tau(Q_\tau(y_{jh}|u_{i0}, X_{i0})|Y_0, \underline{Y}_{-1}). \tag{4}
$$

The vector X_{i0} should be set according to the selected identification scheme (Propositions 2 and 5). In general, IR is a function of the observables Y_t, X_{it}, but the inner quantiles in (4) condition on the unobservable u_{i0}, reflecting the counterfactual experiment of independent sampling of u_{i0}. However, under a linearity assumption, IR becomes constant in Y_t, X_{it} (Propositions 2 and 5). To measure the accumulated effect over horizons 0 through h, we define the *Cumulative Impulse Response Function* by replacing (4) with

$$\mathrm{CIR}^{h,\tau,s}_{j,i}(Y_0, X_{i0}) = Q_\tau\left(Q_\tau\left(\sum_{l=0}^{h} \tilde{y}_{jl}\middle| u_{i0}, X_{i0}\right)\middle| Y_0, \underline{Y}_{-1}\right) - Q_\tau\left(Q_\tau\left(\sum_{l=0}^{h} y_{jl}\middle| u_{i0}, X_{i0}\right)\middle| Y_0, \underline{Y}_{-1}\right).$$

3 Smooth Quantile Local Projections

3.1 Identification by Recursive Short-run Restrictions

The most simple SVAR identification, introduced by Sims (1980), is often called recursive short-run restrictions. It amounts to a Wold causal chain (Wold, 1954) on multivariate time series. In the least squares setup, the recursive short-run restrictions are equivalent to a Cholesky decomposition of the covariance matrix of reduced-form innovations. However, this Cholesky decomposition is irrelevant for identification of quantile treatment effects. In fact, the quantile treatment effects in Definition 1 are well-defined even if the covariance does not exist.

In least squares local projections, Barnichon and Brownlees (2019) point out that recursive short-run restrictions are a special case of identification through controls. We show that this feature applies to the quantile setting, too. We translate the recursive short-run restrictions into a quantile setup via Assumption 3, allowing the kth structural quantile function to depend on the lags of all variables and on $y_{1t}, \dots, y_{k-1,t}$, but not on $y_{k+1,t}, \dots, y_{K,t}$.

Assumption 3 (Recursive Short-run Restrictions)**.** *Let $\{Y_t\}, S_1, S_2, \dots, S_K$ be as in Lemma 1. Assume that for all $k \in \{1, 2, \dots, K-1\}$ the function S_k is constant in $y_{k+1,t}, \dots, y_{K,t}$.*

Proposition 2 characterizes the quantile treatment effect of the ith impulse of size s on the jth variable after h periods by local projections with suitable controls.

Proposition 2 (Identification by Recursive Short-run Restrictions)**.** *If Assumption 3 holds, then for all* $i, j \in \{1, \dots, K\}$, $h \in \mathbb{N}_0$ *there is a function* $S_{ji}^h : \mathbb{R}^{1+i+KP} \mapsto \mathbb{R}$ *and an* h*-dependent process* $\{u_{jit}^h\}_{t\in\mathbb{Z}}$ *such that* $y_{j,t+h} = S_{ji}^h(u_{jit}^h | y_{1t}, \dots, y_{it}, \underline{Y}_{t-1})$*, where* $u_{jit}^h \sim U[0,1]$ *is independent of* $\{y_{1t}, y_{2t}, \dots, y_{it}, Y_{t-1}, Y_{t-2}, \dots\}$ *for all* $t \in \mathbb{Z}$*; also, for all* $\tau \in (0,1)$ *it holds*

$$S_{ji}^h(\tau | y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}) = Q_\tau(y_{j,t+h} | y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}). \tag{5}$$

If also Assumption 2 is satisfied, then

$$\begin{aligned} &IR_{j,i}^{h,\tau,s}(Y_t, (y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}')') = \\ &\quad S_{ji}^h(\tau | y_{1t}, \dots, y_{i-1,t}, y_{it} + s, \underline{Y}_{t-1}) - S_{ji}^h(\tau | y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}). \end{aligned} \tag{6}$$

If, in addition, for some $h \in \mathbb{N}_0$ *and* $\tau \in (0,1)$ *the function* $S_{ji}^h(\tau|\cdot)$ *is linear, that is, if*

$$S_{ji}^h(\tau | y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}) = \beta_{j00}^h(\tau) + \sum_{k=1}^{i} \beta_{jk0}^h(\tau) y_{kt} + \sum_{p=1}^{P} \sum_{k=1}^{K} \beta_{jkp}^h(\tau) y_{k,t-p}, \tag{7}$$

then $IR_{j,i}^{h,\tau,s}(Y_t, (y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}')') = \beta_{ji0}^h(\tau) s$*.*

Equation (6) means that the causal effect of the ith shock on the jth variable after h periods at quantile τ coincides with the effect of the variable y_{it} on $y_{j,t+h}$ after controlling for $y_{1t}, \dots, y_{i-1,t}$ and for the lags of all variables. This identification result is nonparametric and does not require linear conditional quantiles. However, if the conditional quantiles are indeed linear, meaning that (7) holds, then the effect of a unit shock is $\beta_{ji0}^h(\tau)$, which can be estimated by QR on equation (7).

However, the linear form of (7) is useful even if the function (5) is nonlinear. In that case, owing to the QR approximation properties under misspecification (Angrist et al., 2006), (7) becomes the best linear approximation of (5) under asymmetric loss, meaning that the pseudo-parameter $\beta_{ji0}^h(\tau)$ approximates the nonlinear causal effect (6). This local approximation (horizon-specific and quantile-specific) is good as long as (5) is approximately linear. If we suspect substantial nonlinearity in (5), we can add nonlinear terms into (7) following Koenker and Xiao (2006), or, if the sample is large, estimate (5) nonparametrically. In either case, the causal effect is given by (6).

Proposition 2 also shows that the horizon-specific errors $\{u_{jit}^h\}_{t\in\mathbb{Z}}$ are h-dependent. This parallels least-squares local projections, whose horizon-specific errors are MA(h).

3.2 Estimation

Our aim now is to estimate impulse response function from Definition 1 under Assumption 3 and linear conditional quantiles (7); we also discuss what happens if linearity fails.

Suppose we observe a sample $\{Y_t\}_{t=1}^T$. In the special case when $i \geq j$ and $h = 0$, the coefficients in (7) need not be estimated as they are already known: $\beta_{jj0}^{(0)}(\tau) = 1$, while the remaining coefficients are 0. Denote the first horizon that can be meaningfully estimated as $h_0 = \mathbb{1}\{i \geq j\}$. At quantile $\tau \in (0,1)$, the standard quantile loss function is $\rho_\tau(x) = (\tau - \mathbb{1}\{x < 0\})x$. Denote the vector of explanatory variables as $x_t = (1, y_{1,t}, \ldots, y_{i,t}, \underline{Y}_{t-1}')'$. For $1 \leq p \leq P$ denote $\beta_{ph} = (\beta_{j1p}^h, \beta_{j2p}^h, \ldots, \beta_{jKp}^h)'$, and further $\beta_{0h} = (\beta_{j10}^h, \beta_{j20}^h, \ldots, \beta_{ji0}^h)'$, $\theta_h = (\beta_{j00}^h, \beta_{0h}', \beta_{1h}', \beta_{2h}', \ldots, \beta_{Ph}')'$.

Definition 2 (QLP Estimator)**.** *For any ith impulse variable, jth response variable, quantile $\tau \in (0,1)$, and horizon $h \in \mathbb{N}_0$, the quantile local projections estimator $\tilde{\theta}_h(\tau)$ is a solution of*

$$\min_{b_h \in \mathbb{R}^{1+i+KP}} \sum_{t=P+1}^{T-h} \rho_\tau\left(y_{j,t+h} - b_h' x_t\right). \tag{8}$$

The estimators $\tilde{\theta}_{h_0}(\tau), \tilde{\theta}_{h_0+1}(\tau), \ldots, \tilde{\theta}_H(\tau)$ in (8) solve unrelated QR problems, so their values typically vary abruptly between neighboring horizons, which we address by incorporating roughness penalties. Denote the parameter vector $\theta(\tau) = (\theta_{h_0}(\tau)', \theta_{h_0+1}(\tau)', \ldots, \theta_H(\tau)')'$ and its dimension $\bar{K} = (1+i+KP)(H+1-h_0)$. For any $b \in \mathbb{R}^{(1+i+KP)(H+1)}$ write $b = (b_0', b_1', \ldots, b_H')'$, where $b_h \in \mathbb{R}^{1+i+KP}$, $h \in \{0, 1, \ldots, H\}$. Denote the kth element of b_h as $b_{k-1,h}$. Let Δ be the difference operator applied over horizons h in the following sense: $\Delta^1 b_{k,h} = b_{k,h} - b_{k,h-1}$, $\Delta^2 b_{k,h} = b_{k,h} - 2b_{k,h-1} + b_{k,h-2}$, $\Delta^3 b_{k,h} = b_{k,h} - 3b_{k,h-1} + 3b_{k,h-2} - b_{k,h-3}$.

Next, we modify (8) to smooth the estimators over forecast horizons and also to control their behavior at the final horizons. We associate each vector $(b_{k,0}, b_{k,1}, \ldots, b_{k,H})$ with a *roughness penalty* $\lambda_k \in \mathbb{R}_0^+$, which shrinks the vector towards a polynomial of degree $D-1$, where $D \in \{1,2,3\}$. At the final horizons we introduce additional shrinkage, controlled by $\mu \in \mathbb{R}_0^+$, and we call it the *long-run penalty* (it is unrelated to identification by long-run restrictions).

Definition 3 (SQLP Estimator)**.** *Take any ith impulse variable, jth response variable, quantile $\tau \in (0,1)$, difference order $D \in \{1,2,3\}$, non-negative roughness penalties $\{\lambda_k\}_{k=0}^{i+KP}$, non-negative long-run penalty μ, maximal horizon $H \in \mathbb{N}, H > D$, and $B \subset \mathbb{R}^{\bar{K}}$, $\theta(\tau) \in B$. The smooth quantile local projections estimator $\hat{\theta}(\tau) = (\hat{\theta}_{h_0}(\tau)', \hat{\theta}_{h_0+1}(\tau)', \ldots, \hat{\theta}_H(\tau)')'$ solves*

$$\min_{\substack{(b'_{h_0}, b'_{h_0+1}, \ldots, b'_H)' \in B \\ \text{s.t. if } i \geq j, \text{ then } b_{j,0}=1 \text{ and } \forall k \neq j:\ b_{k,0}=0, \\ \text{where } b_h = (b_{0,h}, \ldots, b_{i+KP,h})',\ h \in \{0,1,\ldots,H\}}} \Bigg\{ \sum_{h=0}^{H} \sum_{t=P+1}^{T-h} \rho_\tau \left(y_{j,t+h} - b'_h x_t \right) + \sum_{k=0}^{i+KP} \lambda_k \sum_{h=D}^{H} |\Delta^D b_{k,h}| + \mu \left(\lambda_0 |\Delta^2 b_{0,H}| + \sum_{k=1}^{i+KP} \lambda_k |\Delta b_{k,H}| \right) \Bigg\}. \tag{9}$$

If $\lambda_k = 0$ for all k, then SQLP in (9) coincide with the (unsmoothed) QLP in (8). When $\lambda_k > 0$, the roughness of the coefficient $b_{k,h}$ across horizons h is penalized. The higher the values of roughness penalties, the more smoothing occurs. Recall from (7) that $b_{i,0}, b_{i,1}, \ldots, b_{i,H}$ represent the impulse response function—so $\lambda_i > 0$ amounts to impulse response smoothing.

If the long-run penalty μ is positive, then the roughness penalties are complemented by additional shrinkage at the final horizons—the slope coefficients are shrunk towards a constant function, and the intercept coefficient towards a linear function. The rationale is the limit behavior as $h \to \infty$. Specifically, if $\{Y_t\}$ is strictly stationary and ergodic, the slope coefficients in (7) converge to 0 (and thus to a constant), and the intercept converges to a constant (and thus to a linear function). In applications, we set $\mu = 100$, which practically guarantees shrinkage at the final horizons, unless λ_k is very small. This choice makes sense only for H large enough.

Figure 1 illustrates smoothing of impulse response functions. The QLP estimates (black solid line) are volatile, unlike the SQLP estimates. Without the long-run penalty (left subfigure), the curves eventually become downward sloping. The long-run penalty (right subfigure) straightens

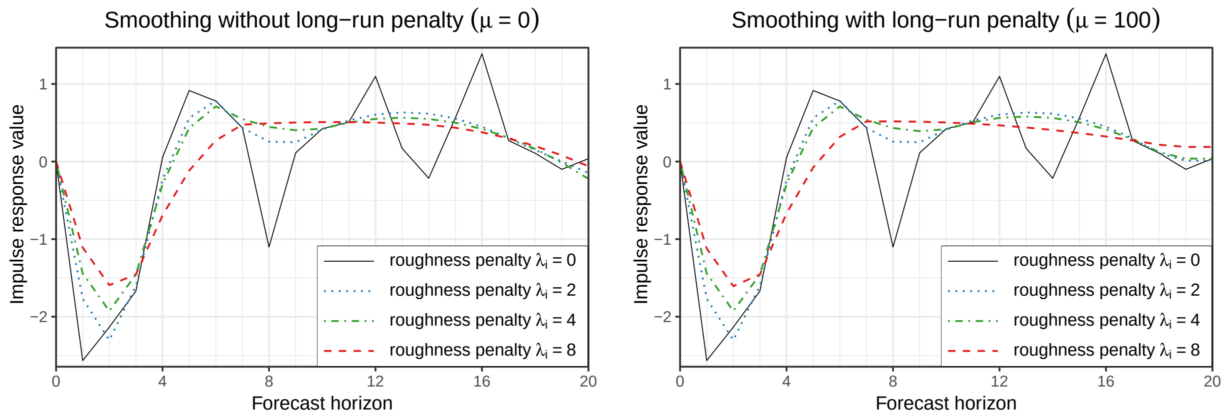

Figure 1: The lines depict the estimated impulse responses at the 75th percentile, for different penalty values. Data and specification are from the empirical application 5.1.

the curves at the end, reconciling the shapes with standard macroeconomic models.

The constraint in (9) for the case $i \geq j, h = 0$ reflects the special case when the coefficients at $h = 0$ are known ex ante. Keeping these coefficients in the estimation problem (even though their values are known) ensures the penalties act on the entire impulse response function, thereby preventing abrupt shifts between horizons 0 and 1.

The linear program (9) can be expressed as QR with a sparse design matrix (see the Online Appendix). We solve it via the sparse Frisch-Newton algorithm of Koenker and Ng (2005), which is fast even for a large number of observations, horizons, and variables.

3.3 Asymptotic Properties

We establish the asymptotics of the SQLP estimator from Definition 3, which will enable us to construct closed-form and weighted bootstrap confidence intervals. These novel results cover also the special case of the unsmoothed QLP estimator from Definition 2. Denote $\bar{\lambda} = (1+\mu)\max_k \lambda_k$

Proposition 3 (SQLP Consistency). *In addition to Assumption 3, assume that $\{Y_t\}$ is ergodic and that S^h_{ji} has the linear form (7) and is increasing in τ at $h = h_0$ a.s. Also assume that $\mathbb{E}|Y_t| < \infty$, $(Y_t', Y_{t-1}', \ldots, Y_{T-P}')'$ is absolutely continuous, $\bar{\lambda} = o(T)$, and B is compact. Take any $\underline{\tau}, \bar{\tau} \in (0,1), \underline{\tau} < \bar{\tau}$. Then* $\sup_{\tau \in [\underline{\tau}, \bar{\tau}]} |\hat{\theta}(\tau) - \theta(\tau)| \xrightarrow{p} 0.$

Let $0 < \underline{\tau} < \bar{\tau} < 1$. We denote by $\Rightarrow$ the weak convergence in the Skorokhod topology on the space of cadlag functions defined on $[\underline{\tau}, \bar{\tau}]$. Denote $e^h_{t,\tau} = y_{j,t+h} - x_t'\theta_h(\tau)$, $\psi_\tau(x) = \tau - \mathbb{1}\{x < 0\}$. Denote the conditional distribution function $F_{t,h}(\cdot) = P(y_{j,t+h} < \cdot | x_t)$, its derivative $f_{t,h}(\cdot)$,

$$\Omega_{h,k}(\tau_1, \tau_2) = \sum_{l=-k}^{h} \mathbb{E}[x_t x_{t+l}' \psi_{\tau_1}(e^h_{t,\tau_1}) \psi_{\tau_2}(e^k_{t+l,\tau_2})], \tag{10}$$

$$\Omega_0(\tau_1, \tau_2) = \begin{pmatrix} \Omega_{0,0}(\tau_1,\tau_2) & \Omega_{0,1}(\tau_1,\tau_2) & \dots & \Omega_{0,H-h_0}(\tau_1,\tau_2) \\ \Omega_{1,0}(\tau_2,\tau_1) & \Omega_{1,1}(\tau_1,\tau_2) & \dots & \Omega_{1,H-h_0}(\tau_1,\tau_2) \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_{H-h_0,0}(\tau_2,\tau_1) & \Omega_{H-h_0,1}(\tau_2,\tau_1) & \dots & \Omega_{H-h_0,H-h_0}(\tau_1,\tau_2) \end{pmatrix},$$

$$f_t(\tau) = \begin{pmatrix} f_{t,h_0}(F^{-1}_{t,h_0}(\tau)) & 0 & \dots & 0 \\ 0 & f_{t,h_0+1}(F^{-1}_{t,h_0+1}(\tau)) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_{t,H}(F^{-1}_{t,H}(\tau)) \end{pmatrix},$$

$$\Omega_1(\tau) = \mathbb{E}[f_t(\tau) \otimes x_t x_t'], \quad \Sigma(\tau_1, \tau_2) = \Omega_1^{-1}(\tau_1)\Omega_0(\tau_1,\tau_2)\Omega_1^{-1}(\tau_2).$$

Assumption 4 (Assumptions for Asymptotic Normality)**.** *In addition to Assumption 3, assume that $\{Y_t\}$ is ergodic, $\mathbb{E}[Y_t^2] < \infty$, $\mathbb{E}[x_t x_t']$ is invertible, $\bar{\lambda} = o(\sqrt{T})$, B is compact, and for all $h \in \{h_0, \dots, H\}$ the function S^h_{ji} has the linear form (7), $f_{t,h}$ is positive, continuous and bounded.*

Proposition 4 (Asymptotic Normality)**.** *Assumption 4 implies $\sqrt{T}(\hat{\theta}(\tau) - \theta(\tau)) \xrightarrow{d} N(0, \Sigma(\tau,\tau))$. If additionally $\mathbb{E}[Y_t^4] < \infty$, then $\sqrt{T}(\hat{\theta}(\tau) - \theta(\tau)) \Rightarrow G(\tau)$ where $G(\tau)$ is a centered $\bar{K}$-dimensional Gaussian process and $Cov[G(\tau_1), G(\tau_2)] = \Sigma(\tau_1, \tau_2)$.*

To construct closed-form confidence intervals for $\theta_{h_0}(\tau), \theta_{h_0+1}(\tau), \dots, \theta_H(\tau)$ using Proposition 4, we need to estimate densities in Ω_1 and long-run variances $\Omega_{h,h}$ for $h \in \{h_0, \dots, H\}$. Following Koenker (2005)[p. 81], we estimate Ω_1 as Powell (1991), with a Gaussian kernel and plug-in bandwidth of Hall and Sheather (1988). For estimating the submatrices $\Omega_{h,h}$, we follow

Galvao and Yoon (2023) and apply the Newey and West (1987) estimator with bandwidth h, as Jordà (2005).

As we saw earlier, if the conditional quantile (5) is not linear, the SQLP estimator approximates the true nonlinear effect. Then, under appropriate weak dependence assumptions, the SQLP estimator remains asymptotically Gaussian, but Ω_0 changes—the series (10) becomes doubly infinite. However, this does not invalidate our estimators of Σ (closed-form or bootstrap), because they do not exploit the finite-sum form of (10).

3.4 Roughness Penalty Selection

The SQLP estimator involves many different roughness penalties—one for the intercept, $i-1$ for the contemporaneous controls, one for the impulse variable, and KP for the lagged controls. To reduce their number, we propose two different ways, depending on the objective of estimation. They both entail just one smoothing parameter, denoted λ. Both ways involve a standardization to account for different scales. For $k \in \{1, 2, \ldots, K\}$, define the standardization coefficient of the kth variable as

$$\nu_k = \frac{1}{T} \sum_{t=1}^{T} \left| y_{kt} - \frac{1}{T} \sum_{s=1}^{T} y_{ks} \right|. \tag{11}$$

The first way is suitable if the aim is to estimate impulse response functions. Then it makes sense to penalize only the coefficient with impulse response interpretation, and set the remaining penalties to zero:

$$\lambda_k = \begin{cases} \lambda \nu_k \text{ where } \lambda \in \mathbb{R}_0^+ & \text{if } k = i, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The second way is appropriate if the goal is to generate forecasts. To make them evolve smoothly, we apply the same penalty on all the coefficients, after standardization:

$$\lambda_{l+Kp+k} = \begin{cases} \lambda \text{ if } l = p = k = 0, \\ \lambda\nu_l \text{ if } p = k = 0 \text{ and } l \in \{1, 2, \ldots, i\}, \\ \lambda\nu_k \text{ if } l = i, p \in \{0, 1, \ldots, P-1\} \text{ and } k \in \{1, 2, \ldots, K\}. \end{cases} \tag{13}$$

The search for appropriate penalty value λ may involve several complementary approaches, such as information criteria, cross-validation, out-of-sample forecast evaluation, and judgment. Information criteria are based on a single estimation, which is convenient computationally. In contrast, cross-validation requires repeated estimations. Out-of-sample predictive accuracy is key in forecasting applications, but goes beyond the scope of our work. Judgment may also play a role—visual inspection of estimated impulse response functions can help to find an appropriate value of λ, or at least rule out unreasonable values.

We propose information criteria similar to Koenker et al. (1994). Denote the effective number of parameters of the SQLP estimator as p_λ. It equals the total number of estimated parameters less those that are locally fitted by polynomials of degree $D-1$, that is,

$$p_\lambda = (D - h_0)(1 + i + KP) + \sum_{k=0}^{i+KP} \sum_{h=D}^{H} \mathbb{1}\{|\Delta^D b_{k,h}| > 0\}.$$

Let us illustrate how p_λ actually works. If $\lambda_k = 0$ for all $k \in \{0, 1, \ldots, i + KP\}$, then typically $|\Delta^D b_{k,h}| > 0$ for all k and h, which implies $p_\lambda = (1 + i + KP)(H + 1 - h_0)$. This is the same number of parameters as in the unsmoothed quantile local projections (8), multiplied by the number of estimation-relevant forecast horizons, $H+1-h_0$. On the other hand, if all the penalties are very large, then $|\Delta^D b_{k,h}| = 0$ for all k and h, so $b_{k,\cdot}$ are polynomials of degree $D - 1$, characterized by D parameters. If $h_0 = 1$, their first value at $h = 0$ is fixed ex ante by the constraint, so $p_\lambda = (1+i+KP)(D-h_0)$. Thus, p_λ satisfies $(1+i+KP)(D-h_0) \leq p_\lambda \leq (1+i+KP)(H+1-h_0)$. The effective number of observations is $N = (T - P - (H + h_0)/2)(H + 1 - h_0)$.

Now we can formulate, for instance, the Bayesian (Schwarz, 1978) information criterion

$$BIC(\lambda) = \log\left(\frac{1}{N}\sum_{h=h_0}^{H}\sum_{t=P+1}^{T-h}\rho_\tau\left(y_{j,t+h} - \hat{\theta}_h' x_t\right)\right) + \frac{p_\lambda \log(N)}{2N}.$$

Information criteria can be complemented by the h-block cross-validation of Racine (1997), used by Barnichon and Brownlees (2019) for smoothing least squares local projections. The information criteria and cross-validation can serve not only to select the roughness penalty, but also the difference order D and the number of lags P.

Note that consistency (Proposition 3) allows a slower convergence rate of $\bar{\lambda}$ than asymptotic normality (Proposition 4). Thus, it makes sense to use smaller λ for confidence intervals than for point estimates (undersmoothing).

4 Smooth Quantile Local Projections with Instruments

4.1 Identification by External Instruments

The recursive short-run restrictions in Section 3 rule out contemporaneous reverse causality, but this is often unrealistic—in most macroeconomic models, the variables mutually affect one another within the same time period. A popular alternative identification relies on an external instrument, that is, an instrumental variable not included among the endogenous or control variables. In essence, the instrument is a proxy for one unobservable shock. Stock and Watson (2018) provide an econometric framework for external instruments within the least squares paradigm; we develop its quantile counterpart. Without loss of generality, take the first impulse, so $i = 1$ in Definition 1. The instrument will be denoted z_t and must satisfy Assumption 5.

Assumption 5 (Contemporaneous and Lead-lag Exogeneity)**.** *Take $\{Y_t\}_{t\in\mathbb{Z}}$ as in Lemma 1. Let $\{z_t\}_{t\in\mathbb{Z}}$ be a sequence of random vectors such that $(Y_t', z_t')'$ is strictly stationary and*

$$z_t \text{ is independent of } \left\{ \bigcup_{s=-\infty}^{\infty} \bigcup_{k=1}^{K} u_{ks} \right\} \setminus u_{1t}. \tag{14}$$

In line with Definition 1, the first impulse ($i = 1$) at time t enters only along u_{1t}. Accordingly, Assumption 5 ensures the instrument z_t may be informative about u_{1t}, but not about any other disturbance. Even though Assumption 5 does not require that z_t and u_{1t} be dependent, such dependence is allowed and desirable.

The following Proposition 5 is the analog of Proposition 2, but with external instruments.

Proposition 5 (Identification by an External Instrument)**.** *If Assumption 5 holds, then for any $j \in \{1, 2, \ldots, K\}$ and all horizons $h \in \mathbb{N}_0$ there is a function $S_j^h : \mathbb{R}^{1+K(P+1)} \mapsto \mathbb{R}$ and standard uniform random variables $\{u_{jt}^h\}_{t\in\mathbb{Z}}$ such that for all $t \in \mathbb{Z}$, $\tau \in (0, 1)$ it holds*

$$Q_\tau\Big(y_{j,t+h} - S_j^h(\tau|Y_t, \underline{Y}_{t-1})\Big|z_t, \underline{Y}_{t-1}\Big) = 0, \quad y_{j,t+h} = S_j^h\Big(u_{jt}^h\Big|Y_t, \underline{Y}_{t-1}\Big), \tag{15}$$

where u_{jt}^h is independent of z_t and $\mathcal{F}_{t-1}$, and $\{u_{jt}^h\}_{t\in\mathbb{Z}}$ is h-dependent; and if also Assumption 2 is satisfied, then

$$IR_{j,1}^{h,\tau,s}(Y_t, \underline{Y}_{t-1}) = S_j^h(\tau|y_{1t} + s, y_{2t}, \ldots, y_{Kt}, \underline{Y}_{t-1}) - S_j^h(\tau|Y_t, \underline{Y}_{t-1}). \tag{16}$$

If, in addition, for some $h \in \mathbb{N}_0$ and $\tau \in (0, 1)$ the function $S_j^h(\tau|\cdot)$ is linear, that is, if

$$S_j^h(\tau|Y_t, \underline{Y}_{t-1}) = \beta_{j00}^h(\tau) + \beta_{j10}^h(\tau) y_{1t} + \sum_{p=1}^{P}\sum_{k=1}^{K} \beta_{jkp}^h(\tau) y_{k,t-p}, \tag{17}$$

then $IR_{j,1}^{h,\tau,s}(Y_t, \underline{Y}_{t-1}) = \beta_{j10}^h(\tau) s$.

Proposition 5 establishes the existence of structural quantile functions S_j^h, which reflect the dynamic causal effect of the first impulse in (16). Under linearity (17), the effect of a unit impulse equals $\beta^h_{j10}(\tau)$, but unlike in Proposition 2, the coefficient pertains to an endogenous variable, y_{1t}. This endogeneity is due to simultaneity—both y_{1t} and the local-projection disturbance u^h_{jt} depend on the disturbances $u_{2t}, u_{3t}, \dots, u_{Kt}$, and thus y_{1t} is dependent on u^h_{jt}.

If u_{kt} and the lags of z_t are dependent, (14) is violated. This problem can be overcome by adding $z_{t-1}, \dots, z_{t-P}$ alongside $\underline{Y}_{t-1}$ in Proposition 5, and in particular, as controls into (17).

Next, we turn to point and interval estimation of the impulse response coefficient $\beta^h_{j10}(\tau)$ under Assumption 5 and linearity (17); we also address ramifications of nonlinearities.

4.2 Point Estimation

In the special case when $j = 1$ and $h = 0$, the values of the coefficients in (17) are known ex ante: $\beta^0_{110}(\tau) = 1$ and all the remaining coefficients equal 0. Denote the first forecast horizon that can be meaningfully estimated as $h_1 = \mathbb{1}\{j = 1\}$.

Denote $\beta_{ph} = (\beta^h_{j1p}, \beta^h_{j2p}, \dots, \beta^h_{jKp})'$, $\beta_h = (\beta^h_{j00}, \beta'_{1h}, \beta'_{2h}, \dots, \beta'_{Ph})'$, $\alpha_h(\tau) = \beta^h_{j10}(\tau)$, $x_t = (1, Y'_{t-1}, \dots, Y'_{t-P})'$, $y_t = y_{jt}$, $d_t = y_{1t}$. Rewrite (17) as

$$S^h_j(\tau|Y_t, \underline{Y}_{t-1}) = \alpha_h(\tau) d_t + \beta_h(\tau)' x_t. \tag{18}$$

Equation (18) can be estimated by IV-QR of Chernozhukov and Hansen (2008), but we adjust their assumptions and asymptotics due to local projections' time dependence—we call this method Quantile Local Projections with Instruments (QLPI). Consider the QR objective function,

$$Q_{h,T}(\tau, \alpha, \beta, \gamma) = \frac{1}{T - P - h} \sum_{t=1+P}^{T-h} \rho_\tau(y_{t+h} - \alpha d_t - \beta' x_t - \gamma' z_t), \tag{19}$$

and the solution corresponding to fixed α and h, denoted

$$(\hat\beta_h(\alpha, \tau), \hat\gamma_h(\alpha, \tau)) = \underset{\beta \in \mathbb{R}^{KP}, \gamma \in \mathbb{R}^{\dim(\gamma)}}{\arg\min} Q_{h,T}(\tau, \alpha, \beta, \gamma). \tag{20}$$

Let $\mathscr{A} \subset \mathbb{R}$ be compact such that $\alpha_h(\tau) \in \mathscr{A}$ for all $h \in \{0, \dots, H\}, \tau \in (0, 1)$. Let $\hat\Gamma_{T,h}(\alpha, \tau)$

be an estimator of the asymptotic variance of $\hat{\gamma}_h(\alpha, \tau)$ (it can be constructed as in Section 3). Assume it is consistent, that is, $\hat{\Gamma}_{T,h}(\alpha, \tau) \xrightarrow{p} \lim_{T\to\infty} \text{Var}[\sqrt{T}\hat{\gamma}_h(\alpha, \tau)]$. Denote the Wald statistic

$$\hat{W}_{T,h}(\alpha, \tau) = (T - P - h)\hat{\gamma}_h(\alpha, \tau)'\hat{\Gamma}_{T,h}^{-1}(\alpha, \tau)\hat{\gamma}_h(\alpha, \tau). \tag{21}$$

Definition 4 (QLPI Point Estimator)**.** *Take any jth response variable and quantile $\tau \in (0, 1)$. For every horizon $h \in \{0, 1, \ldots, H\}$, the QLPI estimator $(\tilde{\alpha}_h(\tau), \tilde{\beta}_h(\tau))$ solves*

$$\begin{aligned} \tilde{\alpha}_h(\tau) &= \underset{\alpha_h \in \mathscr{A}}{\arg\min}\, \hat{W}_{T,h}(\alpha_h, \tau), \\ \tilde{\beta}_h(\tau) &= \hat{\beta}_h(\tilde{\alpha}_h(\tau), \tau). \end{aligned} \tag{22}$$

To reduce the excessive volatility of the estimated impulse response function, we enhance QLPI with penalties and call it Smooth QLPI (SQLPI). Specifically, we smooth the coefficients with impulse response interpretation: $\{\alpha_h\}_{h=0}^{H}$.

Definition 5 (SQLPI Point Estimator)**.** *Take any jth response variable, quantile $\tau \in (0, 1)$, difference order $D \in \{1, 2, 3\}$, roughness penalty $\lambda_1 \geq 0$, long-run penalty $\mu \geq 0$, and maximal horizon $H \in \mathbb{N}, H > D$. The SQLPI estimator $\{(\hat{\alpha}_h(\tau), \hat{\beta}_h(\tau))\}_{h=0}^{H}$ solves*

$$\begin{aligned} \hat{\alpha}(\tau) &= \underset{\substack{\{\alpha_h\}_{h=0}^{H} \in \mathscr{A}^{H+1} \\ \text{s.t. if } j=1, \text{ then } \alpha_0 = 1}}{\arg\min} \sum_{h=0}^{H} \hat{W}_{T,h}(\alpha_h, \tau) + \lambda_1 \sum_{h=D}^{H} |\Delta^D \alpha_h| + \mu\lambda_1 |\Delta \alpha_H|, \\ \hat{\beta}_h(\tau) &= \hat{\beta}_h(\hat{\alpha}_h(\tau), \tau). \end{aligned} \tag{23}$$

The two penalties λ_1 and μ play the same role as in SQLP in Definition 3. In applications, we set $\mu = 10$, which activates shrinkage at final forecast horizons, except for very small λ.

Whereas the QLP and SQLP problems are linear programs, the QLPI problem is IV-QR and can be solved by the inverse QR, described in Chernozhukov and Hansen (2008). Inverse QR discretizes the set $\mathscr{A}$ by a grid of M possible values of the coefficient $\alpha(\tau)$, and then proceeds in two steps. In the first step, for each value a in the grid, (20) is solved by QR and the solution

$(\hat{\beta}_h(a,\tau), \hat{\gamma}_h(a,\tau))$ is stored. In the second step, the minimum (22) is found over all the solutions from the previous step; this is the IV-QR solution.

The inverse QR is computationally demanding if there are many endogenous variables, but fortunately, (22) has just one endogenous variable. In fact, the datasets in empirical macroeconomics typically have just a few hundred observations, making the QR problems (19) ideal for the modified simplex algorithm of Koenker and D'Orey (1987), which is extremely fast in small samples, while giving solutions for all quantiles in $[0,1]$.

The solution of SQLPI also proceeds in two steps. We represent the set $\mathscr{A}$ by an equally spaced grid of M possible values of the coefficient $\alpha(\tau)$, formally $\mathscr{A} = [a_1, a_2, \ldots, a_M]$. Then we solve (20) for every h and store the solutions. It remains to find the minimum (23), using the solutions from the previous step. In contrast to (22), where the minimum is taken over M distinct values, the minimum in (23) is taken over all the possible sequences, of which there are M^{H+1}. Evaluating the expression in (23) for each of these sequences is computationally infeasible in practice. Instead, we reformulate (23) as a shortest path problem and solve it by the algorithm of Dijkstra (1959). The details are in the Online Appendix.

4.3 Interval Estimation

For $\alpha \in \mathscr{A}$ let $y_{t+h,\alpha} = y_{t+h} - \alpha d_t$ and consider the population counterpart of QR (20), whose random coefficient representation is

$$y_{t+h,\alpha} = \beta_h'(\alpha, \tilde{u}_{\alpha,t}^h)x_t + \gamma_h'(\alpha, \tilde{u}_{\alpha,t}^h)z_t, \text{ where } \tilde{u}_{\alpha,t}^h \sim U[0,1],$$

and the deviations from the conditional quantile satisfy for all $\alpha \in \mathscr{A}, \tau \in (0,1)$

$$v_{t,\tau}^h(\alpha) = y_{t+h} - \alpha d_t - \beta_h'(\alpha,\tau)x_t - \gamma_h'(\alpha,\tau)z_t \text{ and } \mathbb{E}[\psi_\tau(v_{t,\tau}^h)(x_t', z_t')'] = 0.$$

Let $f_{t,h,\alpha}(y_{t+h} - \alpha d_t | x_t, z_t)$ be the conditional density of $y_{t+h} - \alpha d_t$ given x_t, z_t. Denote

$$\Omega_{0,h,\alpha_1,\alpha_2}(\tau_1,\tau_2) = \sum_{l=-\infty}^{\infty} \mathbb{E}[(x'_{t+l}, z'_{t+l})'(x'_t, z'_t)\psi_{\tau_1}(v_{t+l,h,\tau_1}(\alpha_1))\psi_{\tau_2}(v_{t,h,\tau_2}(\alpha_2))],$$

$$\Omega_{1,h,\alpha}(\tau) = \mathbb{E}[f_{t,h,\alpha}(F^{-1}_{t,h,\alpha}(\tau))(x'_t, z'_t)'(x'_t, z'_t)],$$

$$\Sigma_{h,\alpha_1,\alpha_2}(\tau_1,\tau_2) = \Omega^{-1}_{1,h,\alpha_1}(\tau_1)\Omega_{0,h,\alpha_1,\alpha_2}(\tau_1,\tau_2)\Omega^{-1}_{1,h,\alpha_2}(\tau_2).$$

We adopt the dual inference of Chernozhukov and Hansen (2008). For $p \in (0,1)$, let c_p be the pth quantile of $\chi^2_{\dim(\gamma)}$ distribution. A confidence region for α at confidence level p is $CR_h(\tau,p) = \{\alpha \in \mathscr{A} : \hat{W}_{T,h}(\alpha,\tau) \leq c_p\}$. It is unrelated to $\hat{\alpha}(\tau)$, and so it is unaffected by smoothing $\hat{\alpha}(\tau)$. Smoothed confidence intervals are defined next.

Definition 6 (SQLPI Interval Estimator)**.** *Take $j, \tau, D, \lambda_1, \mu, H$ and $\{(\hat{\alpha}_h(\tau), \hat{\beta}_h(\tau))\}_{h=0}^H$ as in Definition 5. The SQLPI interval estimator $\{(\underline{\alpha}_h(\tau), \bar{\alpha}_h(\tau))\}_{h=0}^H$ is defined as*

$$\bar{\alpha}(\tau,p) = \max\left\{ \underset{\substack{\{\alpha_h\}_{h=0}^H \in \mathscr{A}^{H+1} \\ s.t.\ \alpha_h \geq \hat{\alpha}_h, \\ if\ j=1,\ then\ \alpha_0=1}}{\arg\min} \sum_{h=0}^H |\hat{W}_{T,h}(\alpha_h,\tau) - c_p| + \lambda_1 \sum_{h=D}^H |\Delta^D \alpha_h| + \mu\lambda_1 |\Delta \alpha_H| \right\},$$

$$\underline{\alpha}(\tau,p) = \min\left\{ \underset{\substack{\{\alpha_h\}_{h=0}^H \in \mathscr{A}^{H+1} \\ s.t.\ \alpha_h \leq \hat{\alpha}_h, \\ if\ j=1,\ then\ \alpha_0=1}}{\arg\min} \sum_{h=0}^H |\hat{W}_{T,h}(\alpha_h,\tau) - c_p| + \lambda_1 \sum_{h=D}^H |\Delta^D \alpha_h| + \mu\lambda_1 |\Delta \alpha_H| \right\}.$$

If $\lambda_1 = 0$, then $(\underline{\alpha}_h(\tau,p), \bar{\alpha}_h(\tau,p)) = (\min(CR_h(\tau,p)), \max(CR_h(\tau,p)))$. For $\lambda_1 > 0$ the roughness penalty plays the same role as when smoothing the point estimates in (23).

4.4 Asymptotic Properties

Assumption 6 (Assumptions for Dual Inference)**.** *Assume that $\{Y_t\}$ is ergodic, $\mathbb{E}[Y_t^4] < \infty$, $\mathbb{E}[(x'_t, z'_t)'(x'_t, z'_t)]$ is finite and invertible, $(1+\mu)\lambda_1 = o(1)$, and for all $h \in \{h_0, \ldots, H\}$ the function S_j^h has the linear form (17). Further, for all τ and h, $\alpha_h(\tau) \in \mathscr{A}$, where $\mathscr{A}$ is compact.*

Finally, for all $h \in \{h_0, \ldots, H\}$, $\alpha \in \mathscr{A}$ and $(\beta, \gamma) = (\beta(\alpha, \tau), \gamma(\alpha, \tau))$, the density $f_{t,h,\alpha}(\cdot)$ is positive, continuous and bounded.

The assumptions above are similar to Chernozhukov and Hansen (2008), but we do not impose independent sampling.

Proposition 6 (Inference with Weak Instruments)**.** *Under Assumptions 5 and 6, for all $h \in \{h_1, h_1+1, \ldots, H\}$: $\sqrt{T}(\hat{\gamma}_h(\alpha_h(\tau), \tau) - \gamma_h(\alpha_h(\tau), \tau)) \Rightarrow G_h(\tau)$ where $G_h(\tau)$ is a centered dim(γ)-dimensional Gaussian process and Var$[G_h(\tau)] = \Gamma_h(\alpha(\tau), \tau)$. Also,*
$P[\alpha_h(\tau) \in CR_h(\tau, p)] \to p$ and $\liminf_{T\to\infty} P[\alpha_h(\tau) \in (\underline{\alpha}_h(\tau, p), \bar{\alpha}_h(\tau, p))] \geq p$ for all $\tau, p \in (0, 1)$.

The main point of Proposition 6 is that in large samples, the confidence interval contains the impulse response value $\alpha_h(\tau)$ with probability p or higher, irrespective of instrument relevance.

Although the SQLPI estimator assumes linear structural quantile function (17), it may be useful even if linearity fails. The QR (19) approximates the true nonlinear conditional quantile function (Angrist et al., 2006), so we can interpret $\alpha_h(\tau)$ as an approximation, albeit suboptimal, of the nonlinear causal effect (16). The estimator $\hat{\Gamma}_h^{-1}(\alpha, \tau)$ under misspecification is still justified for the same reasons as in the SQLP case.

4.5 Roughness Penalty Selection

The SQLPI framework involves only one roughness penalty, which corresponds to the coefficients on the endogenous variable—these are interpreted as the impulse response function. To avoid scaling effects, we set $\lambda = \lambda_1 \nu_1 (a_2 - a_1)$ where ν_1 is the standardization coefficient (11) and $(a_2 - a_1)$ is the step size of the equally spaced grid.

The effective number of parameters p_λ (for a fixed quantile index τ) is

$$p_\lambda = (1 + PK)(H + 1 - h_1) + D - h_1 + \sum_{h=D}^{H} \mathbb{1}\{|\Delta^D \alpha_h| > 0\}.$$

Denote the effective number of observations as $N = (T - P - (H + h_1)/2)(H + 1 - h_1)$.

The SQLPI estimator is a not a maximum likelihood estimator, so standard information criteria are not applicable. That is why we propose the following quasi-information criterion

$$BIC(\lambda) = \frac{1}{N}\hat{W}_{T,h}(\alpha, \tau) + \frac{p_\lambda \log N}{2N}. \tag{24}$$

Its first term substitutes the fidelity, while the second term is the penalty of Schwarz (1978).

5 Empirical Applications

5.1 SQLP: Financial Conditions and GDP

Adrian et al. (2019) study the effect of tighter financial conditions on GDP growth distribution in USA. They find the effect is large and negative in the lower part of GDP growth distribution, but negligible in the upper part. However, Adrian et al. (2019) do not report confidence intervals of their quantile regression estimates and they do not include any lags as controls, which is at odds with our Proposition 2. For our application we use the variables from their dataset—the Chicago Fed's National Financial Conditions Index (NFCI) and real, seasonally adjusted GDP growth rate. Higher NFCI means tighter financial conditions. Denote the values of these variables at time t as $NFCI_t$ and $gGDP_t$, respectively. The quarterly data cover 1973Q1–2015Q4. The main specification of Adrian et al. (2019) is

$$Q_\tau(gGDP_{t+h}|gGDP_t, NFCI_t) = \beta_{100}^{(h)}(\tau) + \beta_{110}^{(h)}(\tau)gGDP_t + \beta_{120}^{(h)}(\tau)NFCI_t, \tag{25}$$

where $\beta_{120}^{(h)}(\tau)$ aims to measure the effect of a 1 unit financial conditions shock on the τth quantile of GDP growth distribution after h quarters.

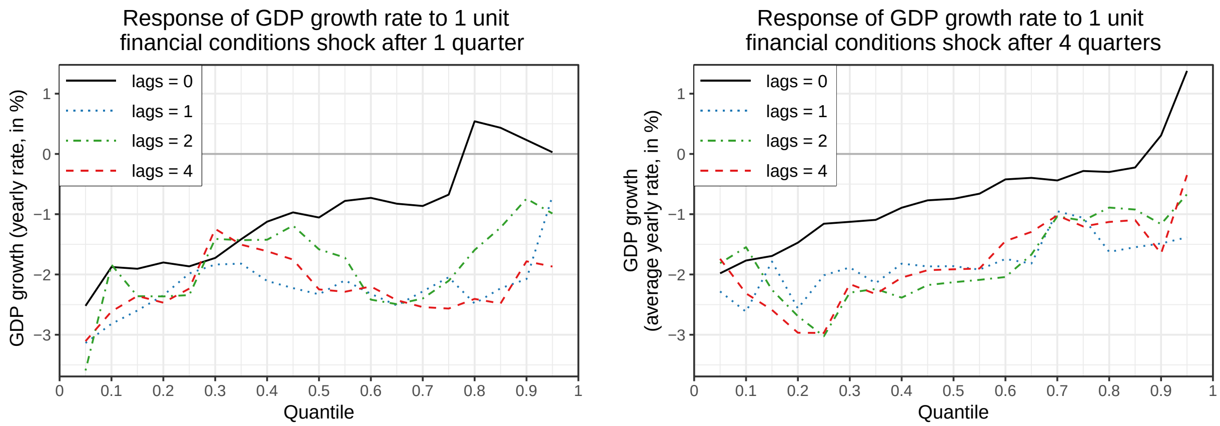

Figure 2: The lines show the effect of financial conditions on GDP growth, depending on lag length. The solid lines (no lags as controls) are the estimates of Adrian et al. (2019) in their main specification.

Contrasting (25) against Proposition 2, we see that (25) embeds the following identification assumption: NFCI has no contemporaneous effect on GDP growth distribution. Even though Adrian et al. (2019) do not mention this restriction, it is essential for causal interpretation.[3] Proposition 2 also highlights the importance of controlling for lags—their omission may lead to inconsistency. To see the lag length sensitivity, we estimate the main specification (25) of Adrian et al. (2019), and then we include 1, 2 or 4 lags. Figure 2 shows that once at least one lag is included, the estimated effect of Adrian et al. (2019) in the upper part of the distribution flips the sign from positive to negative. Also, the asymmetry across quantiles greatly diminishes.

Next, we invoke the SQLP estimator. While Adrian et al. (2019) only consider one-quarter ahead and four-quarter ahead predictions, we look at all horizons up to 20 quarters ahead. Moreover, we consider two ways of capturing the effect: First, the response is plain GDP growth rate, quarter on quarter, at yearly rate. Second, we use the cumulative GDP growth rate as the response. We control for 4 lags and focus on the quantiles 0.1, 0.25, 0.5, 0.75, 0.9. (This is just to simplify exposition; more quantiles are depicted in Figure 6.)

We set a roughness penalty of the form (12) with $D = 3$. To ease comparisons, we take the

[3]In principle, this identification assumption could be justified as follows: NFCI reflects financial markets that quickly react to economic news, while GDP reacts more slowly (with a lag).

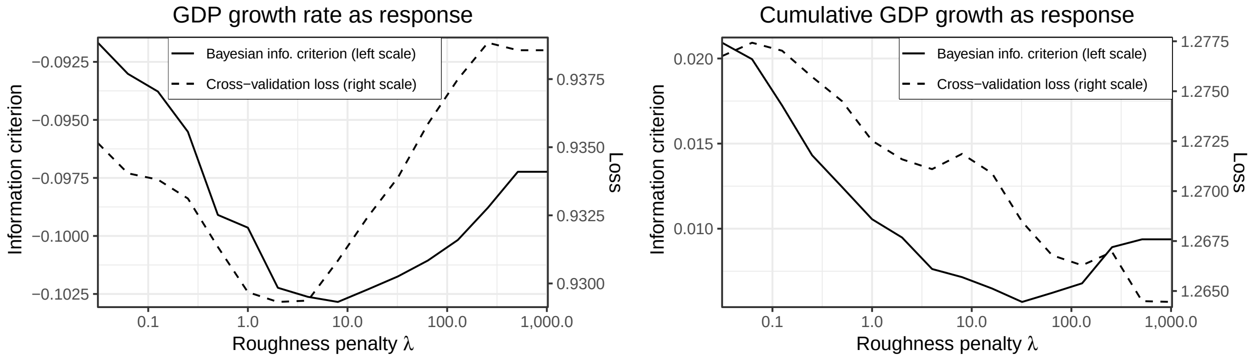

Figure 3: The lines show how the information criterion and cross-validation loss vary with the penalty.

same penalty value for all 5 quantiles. For point estimates we choose the penalty by the Bayesian information criterion (averaged over the 5 quantiles) from $\{2^{-5}, 2^{-4}, \ldots, 2^{10}\}$, but for confidence intervals we use a 4-times smaller penalty (undersmoothing in line with Propositions 3 and 4). Figure 3 illustrates the bias-variance tradeoff.

Figure 4 shows the estimated impulse responses, both for $\lambda = 0$ (unsmoothed) and for $\lambda = 8$ (smoothed). The unsmoothed estimates are very volatile, which might even lead to type I errors. For instance, at the 90th percentile at horizons 4 and 16, the effect suddenly surges and gets statistically significant, which seems implausible. The dependent variable in Figure 4 is the (noncumulative) GDP growth rate. We already mentioned that an alternative is the cumulative GDP growth rate, which we consider now. The Bayesian information criterion suggests $\lambda = 32$ (Figure 3 to the right). We report the impulse responses in Figure 5. While Figure 5 plots the effect across horizons for a few selected quantiles, it's also possible to plot the effect across quantiles for a few selected horizons—we do so in Figure 6 for quantiles $0.10, 0.11, \ldots, 0.89, 0.90$ (keeping λ the same to ease the comparison). All the confidence bands are closed-form (see Proposition 4 and the paragraph thereafter). Bootstrap confidence bands look similar (reported in the Online Appendix).

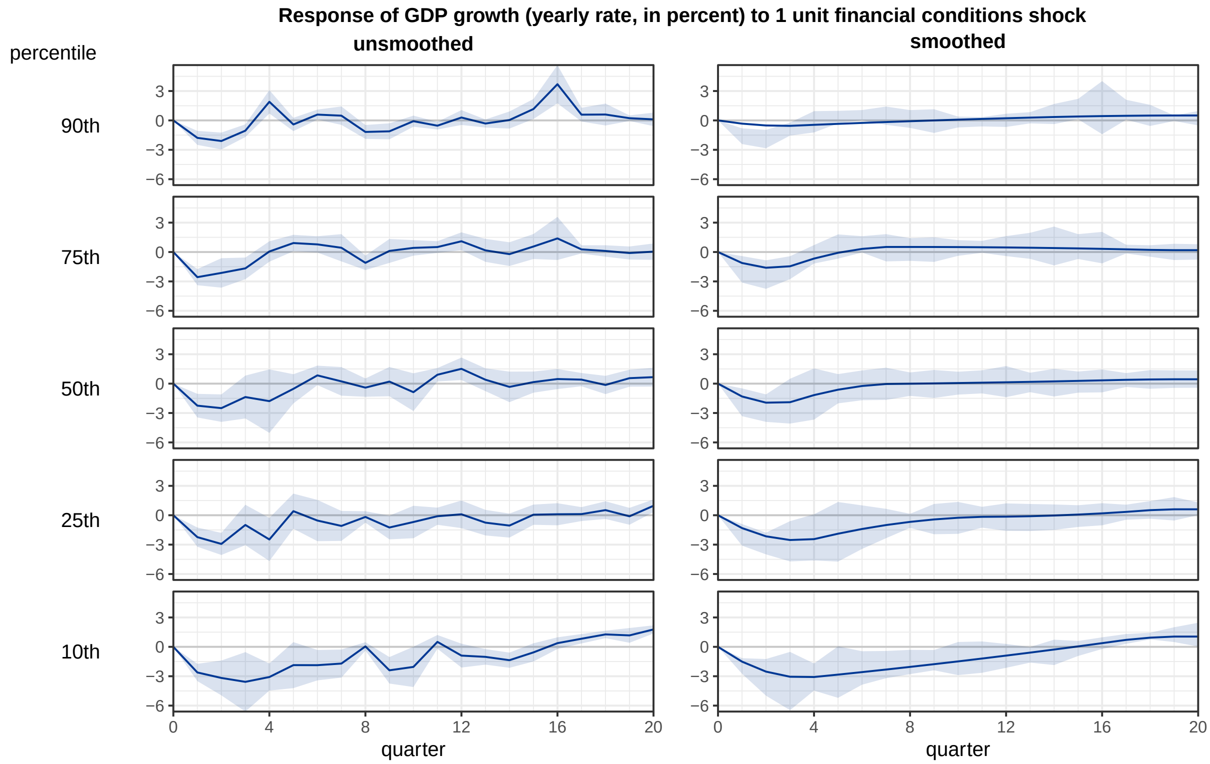

Figure 4: The lines show how financial conditions affect the GDP growth rate (quarter on quarter, at yearly rate). The shaded areas are the 90% confidence intervals.

Although our estimates confirm the finding of Adrian et al. (2019) on the relevance of financial conditions for the low quantiles of GDP growth, the results differ in the upper quantiles, because their specification does not fully capture the causal effect—it does not control for the lags, which is one of the insights of our Proposition 2. While they estimate a small positive effect of financial conditions on GDP growth in upper quantiles, our estimate there is negative and statistically significant. We conclude that financial conditions always matter for the economy: They are relevant not only for recessions, but also during normal times and even in recovery periods.

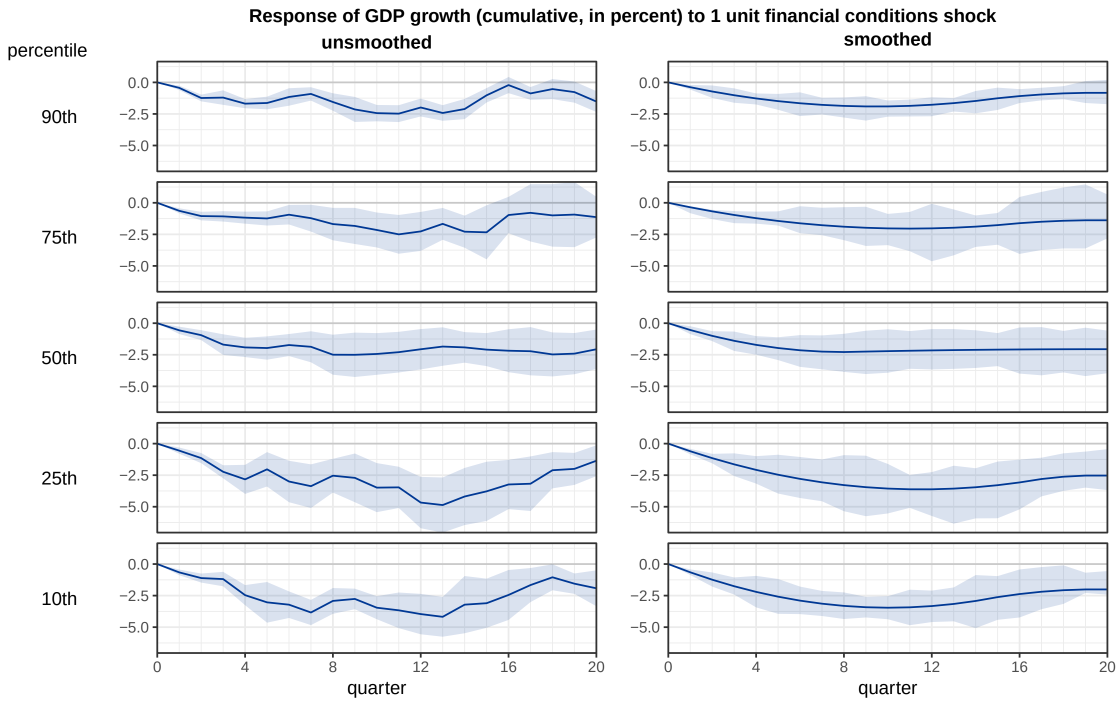

Figure 5: The lines show how financial conditions affect the GDP growth rate (cumulative). The shaded areas are the 90% confidence intervals.

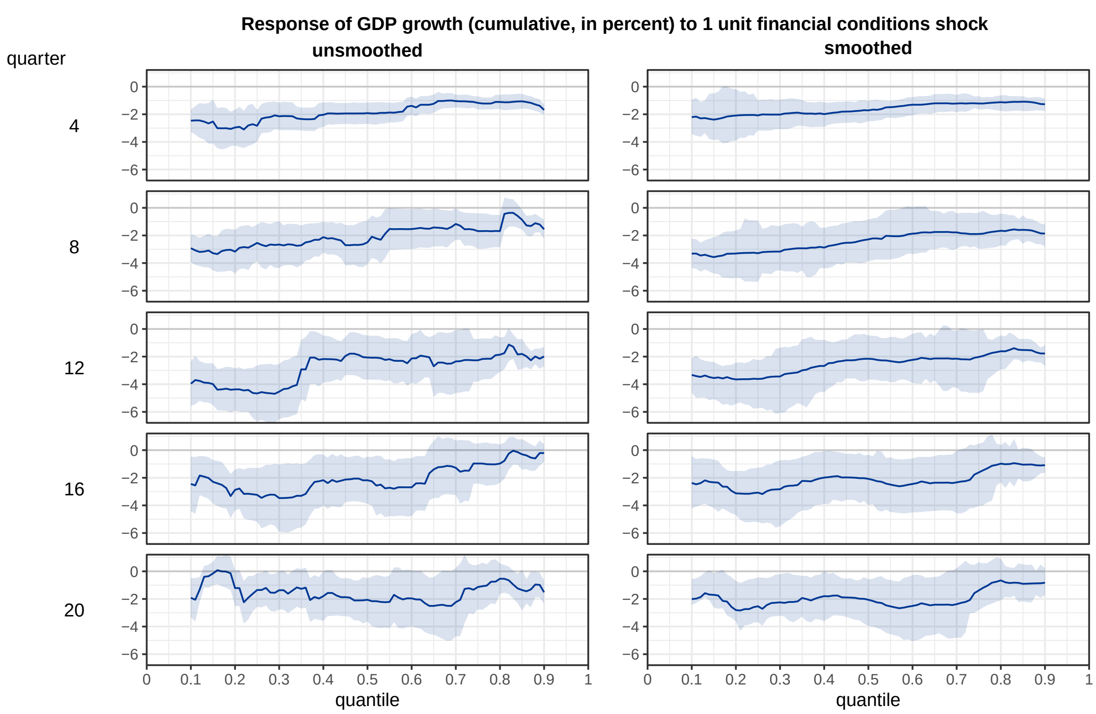

Figure 6: The lines show how financial conditions affect the GDP growth rate (cumulative). The shaded areas are the 90% confidence intervals.

5.2 SQLPI: Dynamic Causal Effects of Monetary Policy

Stock and Watson (2018) identify the effects of interest rates by monetary policy surprises. We use the same variables as them, except for the interest rate—instead of the one-year treasury yield we use the Federal funds rate, which is the main policy rate. Thus, our model includes the following variables: the Federal funds rate, industrial production growth rate, consumer price index (CPI) growth rate, and the excess bond premium from Gilchrist and Zakrajšek (2012). We use the changes in the Federal Funds futures rates around FOMC announcements as instrument. There are 270 observations, covering 1990m1–2012m6.

We start with effect of interest rates on the consumer price index, and we make the following choices. For the coefficient on the impulse variable (the interest rate) we use an equally spaced grid on the interval $[-20, 20]$ with step size $2/3$. We impose a penalty on its second derivative ($D = 2$). We select the optimal penalty from $\{2^{-2}, 2^{-1}, \dots, 2^5\}$ by the Bayesian quasi-information criterion, but for confidence intervals we use 4-times smaller penalty (undersmoothing). Following Stock and Watson (2018), we use 4 lags of all variables and of the instrument as controls.

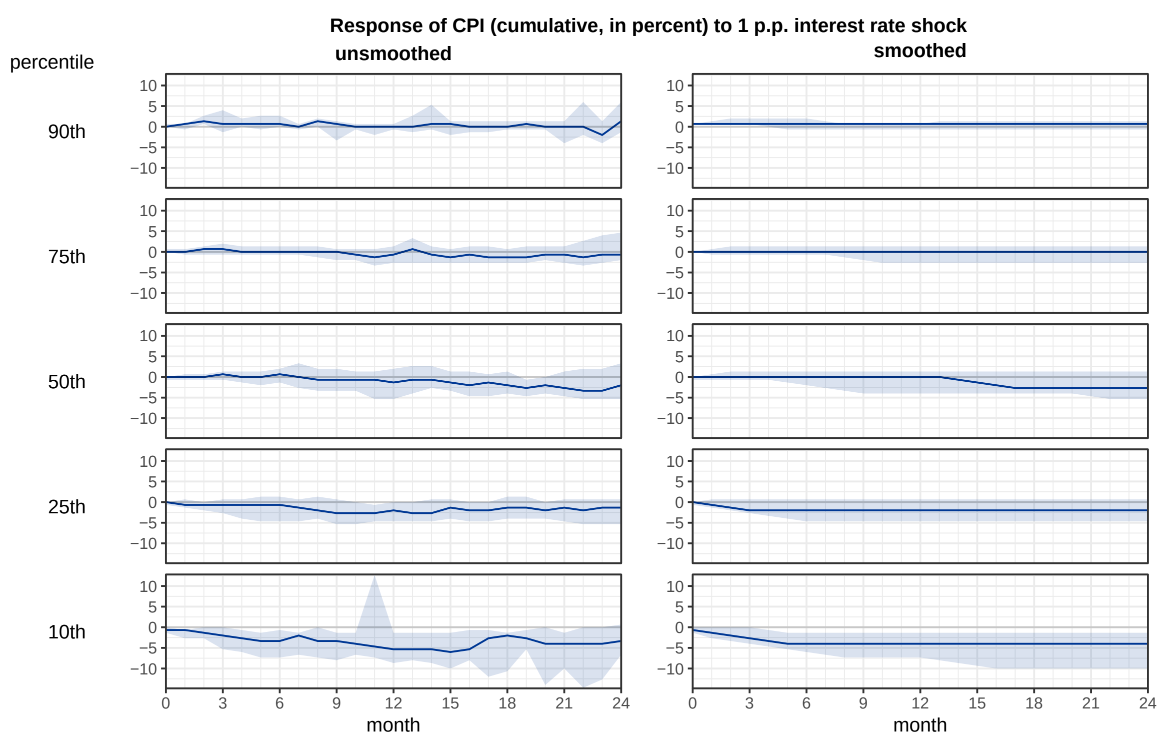

Figure 7: The lines show how interest rates affect consumer prices, along with 90% confidence intervals.

The estimated effects are in Figure 7. One percentage point (p.p.) interest rate increase reduces the center of inflation's distribution after two years by about 3 p.p. (cumulative effect). However, the median effect is mostly not statistically significant, corroborating Stock and Watson (2018), whose estimated mean effect is not statistically significant either.

In addition, the Figure 7 reveals an asymmetry: The interest rate affects inflation distribution predominantly at lower quantiles, where the effect is large and statistically significant, while at upper quantiles the effect is close to zero and not statistically significant. Thus, the interest rate affects the odds of declining inflation, but not the odds of rising inflation. In practical terms, this asymmetry means that conventional monetary policy is more effective at curbing inflation than at generating it.

We repeat the exercise for the other three response variables[4] and show the responses in Figure 8. They all exhibit heterogeneity across quantiles. The interest rate response is pronounced and statistically significant only at upper quantiles. This means that some monetary policy shocks are just one-off interventions without a lasting effect on the interest rates (lower quantiles), while other monetary policy shocks have more lasting effects (upper quantiles). The response of the industrial production is negative and statistically significant only at lower quantiles. Thus, conventional monetary policy is more effective at reducing economic activity than at stimulating it.[5]

Finally, the response of the excess bond premium is positive and statistically significant only at upper quantiles, so contractionary monetary policy induces an uncertain and heterogeneous response of the financial sector, elevating the risks of financial stress and credit tightening. As robustness checks, we use bootstrap, set the penalty to 0, and also switch to a finer grid (25% smaller step size)—we still reach the same conclusions.

---

[4]For interest rate and the excess bond premium, the grid is on $[-10, 10]$ with step size $\frac{1}{3}$, while for industrial production on $[-30, 30]$ with step size 1. We select the penalties as before by the Bayesian information criterion.

[5]This is in line with Klein (2001), who argued: "*I believe that monetary policy has a chronic defect. It is asymmetric—it works better in restraining an economy than in stimulating an economy.*"

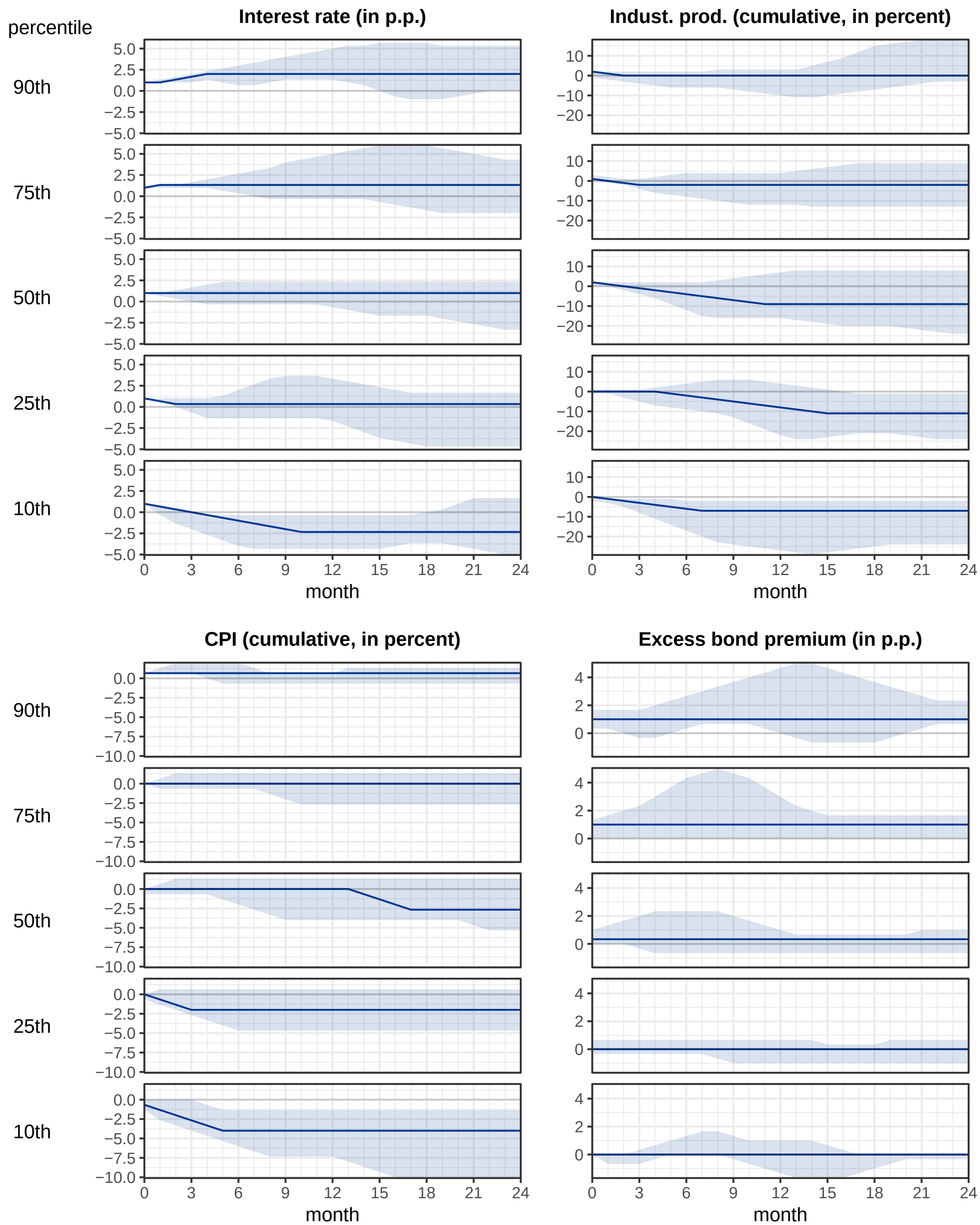

Figure 8: The lines show the effect of one percentage point rise in the Federal funds rate on the Federal funds rate, the industrial production index, the consumer price index and the excess bond premium. The solid lines represent point estimates, the shaded areas are 90% confidence intervals.

6 References

Adrian, T., Boyarchenko, N., Giannone, D., 2019. Vulnerable Growth. American Economic Review 109, 1263–89.

Angrist, J., Chernozhukov, V., Fernández-Val, I., 2006. Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure. Econometrica 74, 539–563.

Barnichon, R., Brownlees, C., 2019. Impulse Response Estimation by Smooth Local Projections. The Review of Economics and Statistics 101, 522–530.

Chavleishvili, S., Manganelli, S., 2024. Forecasting and stress testing with quantile vector autoregression. Journal of Applied Econometrics 39, 66–85.

Chernozhukov, V., Hansen, C., 2008. Instrumental variable quantile regression: A robust inference approach. Journal of Econometrics 142, 379 – 398.

Dijkstra, E. W., 1959. A note on two problems in connexion with graphs. Numer. Math. 1, 269–271.

Distante, R., Petrella, I., Santoro, E., 2013. Asymmetry reversals and the business cycle. Tech. rep., Fondazione Eni Enrico Mattei (FEEM).

Fu, W., Knight, K., 2000. Asymptotics for lasso-type estimators. The Annals of Statistics 28, 1356 – 1378.

Galvao, A. F., Yoon, J., 2023. HAC Covariance Matrix Estimation in Quantile Regression. Journal of the American Statistical Association 0, 1–12.

Gilchrist, S., Zakrajšek, E., 2012. Credit spreads and business cycle fluctuations. American Economic Review 102, 1692–1720.

Hall, P., Sheather, S. J., 1988. On the distribution of a studentized quantile. Journal of the Royal Statistical Society. Series B (Methodological) 50, 381–391.

Han, H., Jung, W., Lee, J. H., 2022. Estimation and Inference of Quantile Impulse Response Functions by Local Projections: With Applications to VaR Dynamics. Journal of Financial Econometrics 2022, 1–29.

Jordà, Ò., 2005. Estimation and inference of impulse responses by local projections. American Economic Review 95, 161–182.

Klein, L., 2001. Keynesianism again. Challenge 44, 6–16.

Knight, K., 1998. Limiting Distributions for L1 Regression Estimators under General Conditions. The Annals of Statistics 26, 755–770.

Koenker, R., 2005. Quantile Regression. Econometric Society Monographs, Cambridge University Press.

Koenker, R., Bassett, G., 1978. Regression Quantiles. Econometrica 46, 33–50.

Koenker, R., Ng, P., 2005. A frisch-newton algorithm for sparse quantile regression. Acta Mathematicae Applicatae Sinica, English Series 21, 225–236.

Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673–680.

Koenker, R., Xiao, Z., 2006. Quantile Autoregression. Journal of the American Statistical Association 101, 980–1006.

Koenker, R. W., D'Orey, V., 1987. Algorithm as 229: Computing regression quantiles. Journal of the Royal Statistical Society. Series C (Applied Statistics) 36, 383–393.

Lee, D. J., Kim, T.-H., Mizen, P., 2021. Impulse response analysis in conditional quantile models with an application to monetary policy. Journal of Economic Dynamics and Control 127, 104102.

Montes-Rojas, G., 2022. Estimating Impulse-Response Functions for Macroeconomic Models using Directional Quantiles. Journal of Time Series Econometrics 14, 199–225.

Newey, W., West, K., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–08.

Newey, W. K., McFadden, D., 1994. Chapter 36 Large sample estimation and hypothesis testing, vol. 4 of *Handbook of Econometrics*. Elsevier.

Portnoy, S., 1997. Local asymptotics for quantile smoothing splines. Ann. Statist. 25, 414–434.

Powell, J., 1991. Estimation of monotonic regression models under quantile restrictions. In: Barnett, W. A., Powell, J., Tauchen, G. E. (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge University Press.

Racine, J., 1997. Feasible cross-validatory model selection for general stationary processes. Journal of Applied Econometrics 12, 169–179.

Ramey, V., 2016. Chapter 2 - Macroeconomic Shocks and Their Propagation. Elsevier, vol. 2 of *Handbook of Macroeconomics*, pp. 71–162.

Rosenblatt, M., 1952. Remarks on a Multivariate Transformation. The Annals of Mathematical Statistics 23, 470–472.

Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464.

Sims, C. A., 1980. Macroeconomics and Reality. Econometrica 48, 1–48.

Stock, J. H., Watson, M. W., 2001. Vector autoregressions. Journal of Economic Perspectives 15, 101–115.

Stock, J. H., Watson, M. W., 2018. Identification and estimation of dynamic causal effects in macroeconomics using external instruments. The Economic Journal 128, 917–948.

White, H., Kim, T.-H., Manganelli, S., 2015. VAR for VaR: Measuring tail dependence using multivariate regression quantiles. Journal of Econometrics 187, 169 – 188.

Wold, H., 1954. Causality and Econometrics. Econometrica 22, 162–177.

Proofs

Proof of Lemma 1. Let $\{Y_t\}_{t\in\mathbb{Z}}$ be as in Assumption 1. If $K > 1$, the representation (1) is not unique; we are going to show that one such representation exists. The proof relies on the transformation of Rosenblatt (1952), but we do not assume absolute continuity. Let $x_{1t} = \underline{Y}_{t-1}$ and for $k \in \{2, \dots, K\}$ let $x_{kt} = (y_{1t}, y_{2t}, \dots, y_{k-1,t}, \underline{Y}'_{t-1})'$. For $k \in \{1, \dots, K\}$, let $F_k(y) = P(y_{kt} \leq y|x_{kt})$, $G_k(\tau) = Q_\tau(y_{kt}|x_{kt})$, $\tilde{u}_{kt} = \inf\{\tau \in (0,1) : y_{kt} \leq G_k(\tau)\}$.

We first consider the case when $F_k(y)$ is continuous in y a.s. Then $G_k(\tau)$ is increasing in τ a.s., so $\tilde{u}_{kt} = G_k^{-1}(y_{kt})$ a.s. Take any $\tau \in (0,1)$. We have $P(\tilde{u}_{kt} \leq \tau|x_{kt}) = P(G_k^{-1}(y_{kt}) \leq \tau|x_{kt}) = P(y_{kt} \leq G_k(\tau)|x_{kt}) = P(y_{kt} \leq Q_\tau(y_{kt}|x_{kt})|x_{kt}) = \tau$. As a result, $\tilde{u}_{kt} \sim U[0,1]$. We will show that $\{\tilde{u}_{kt}\}$ are independent by showing that $\{\tilde{u}_{k_i t_i}\}_{i=1}^n$ are independent for any $k_i \in \{1, 2 \dots, K\}, t_i \in \mathbb{Z}, n \in \mathbb{N}$. Without loss of generality, suppose that $i > j$ implies either $t_i > t_j$ or $t_i = t_j, k_i > k_j$. Take any $y_1, \dots, y_n \in [0,1]$ and let $\mathcal{F}_{0,t} = \sigma(Y_{t-1}, Y_{t-2}, \dots)$, for $k \in \{2, \dots, K\}$ let $\mathcal{F}_{k-1,t} = \sigma(y_{1t}, y_{2t}, \dots, y_{k-1,t}, \mathcal{F}_{0,t})$. Using induction,

$$P\left(\bigcap_{i=1}^n \{\tilde{u}_{k_i t_i} \leq y_i\}\right) = \mathbb{E}\left[\prod_{i=1}^n \mathbb{1}\{\tilde{u}_{k_i t_i} \leq y_i\}\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^n \mathbb{1}\{\tilde{u}_{k_i t_i} \leq y_i\}\middle|\mathcal{F}_{k_n-1,t_n}\right]\right] =$$
$$\mathbb{E}\left[\prod_{i=1}^{n-1} \mathbb{1}\{\tilde{u}_{k_i t_i} \leq y_i\}\mathbb{E}\big[\mathbb{1}\{\tilde{u}_{k_n t_n} \leq y_n\}\big|\mathcal{F}_{k_n-1,t_n}\big]\right] = \mathbb{E}\left[\prod_{i=1}^{n-1} \mathbb{1}\{\tilde{u}_{k_i t_i} \leq y_i\}\right.$$
$$\left.P\big(\tilde{u}_{k_n t_n} \leq y_n\big|\mathcal{F}_{k_n-1,t_n}\big)\right] = y_n \mathbb{E}\left[\prod_{i=1}^{n-1} \mathbb{1}\{\tilde{u}_{k_i t_i} \leq y_i\}\right] = \dots = \prod_{i=1}^n y_i = \prod_{i=1}^n P(\tilde{u}_{k_i t_i} \leq y_i).$$

Now we construct disturbances $\{u_{kt}\}$ in the general case when F_k may be discontinuous. For that, we disperse the point mass of $\tilde{u}_{kt}$ at the discontinuity points of F_k. Let $\mathcal{D}_k = \{x \in R^{k-1+KP} : F_k(\cdot|x_{kt} = x)$ is discontinuous$\}$ and $D_{k,x} = \{y \in \mathbb{R} : F_k(y|x_{kt} = x)$ is discontinuous at $y\}$. For all $x \in \mathcal{D}_k$, $d \in D_{k,x}$, let $F^-_{k,x,d} = \lim_{y\to d^-} F_k(y|x_{kt} = x)$, $F^+_{k,x,d} = F_k(d|x_{kt} = x)$. Take $U_{k,t,x,d} \sim U[0, F^+_{k,x,d} - F^-_{k,x,d}]$ such that the collections of random variables $\{U_{k,t,x,d} : k \in \{1, 2, \dots, K\}, t \in \mathbb{Z}, x \in \mathcal{D}_k, d \in D_{k,x}\}$ and $\{\tilde{u}_{kt} : k \in \{1, 2, \dots, K\}, t \in \mathbb{Z}\}$ are independent. Let $u_{kt} = \tilde{u}_{kt} + U_{k,t,x_{kt},d}\mathbb{1}\{d \in D_{k,x_{kt}}, \tilde{u}_{kt} = F^-_{k,x_{kt},d}\}$. Take any $d \in D_{k,x_{kt}}$. Then $P(u_{kt} \leq \tau|\tau \notin [F^+_{k,x_{kt},d} -$

$F^{-}_{k,x_{kt},d}]) = P(u_{kt} \leq \tau | \tilde{u}_{kt} \neq F^{-}_{k,x_{kt},d}) = P(\tilde{u}_{kt} \leq \tau) = \tau$. On the other hand, $P(u_{kt} \leq \tau | \tau \in [F^{+}_{k,x_{kt},d} - F^{-}_{k,x_{kt},d}]) = P(u_{kt} \leq \tau | \tilde{u}_{kt} = F^{-}_{k,x_{kt},d}) = P(F^{-}_{k,x_{kt},d} + U_{k,t,x_{kt},d} \leq \tau | \tilde{u}_{kt} = F^{-}_{k,x_{kt},d}) = \tau$. Thus, $u_{kt} \sim U[0,1]$. Since $\{\tilde{u}_{kt}\}$ and $\{U_{k,t,x,d}\}$ are independent, $\{u_{kt}\}$ are also independent.

We will show that $y_{kt} = Q_{u_{kt}}(y_{kt}|x_{kt})$. Take any $d \in D_{k,x_{kt}}$. First, $Q_{u_{kt}}(y_{kt}|x_{kt}, \tilde{u}_{kt} \neq F^{-}_{k,x_{kt},d}) = \inf\{y : F_k(y|x_{kt}) \geq \tilde{u}_{kt}\}$. Second, $Q_{u_{kt}}(y_{kt}|x_{kt}, \tilde{u}_{kt} = F^{-}_{k,x_{kt},d}) = \inf\{y : F_k(y|x_{kt}) \geq \tilde{u}_{kt} + U_{k,t,x_{kt},d}\} = \inf\{y : F_k(y|x_{kt}) \geq \tilde{u}_{kt}\}$. Therefore, $Q_{u_{kt}}(y_{kt}|x_{kt}) = \inf\{y : F_k(y|x_{kt}) \geq \tilde{u}_{kt}\} = \inf\{y : F_k(y|x_{kt}) \geq F_k(y_{kt}|x_{kt})\} = y_{kt}$.

To show claim (i), take any $t \in \mathbb{Z}$, $\tau \in (0,1)$. We are going to verify (1). Let

$$
\begin{aligned}
S_1(\tau|y_{2,t}, y_{3,t}, \ldots, y_{K,t}, \underline{Y}_{t-1}) &= Q_\tau(y_{1t}|x_{1t}), \\
S_2(\tau|y_{1,t}, y_{3,t}, \ldots, y_{K,t}, \underline{Y}_{t-1}) &= Q_\tau(y_{2t}|x_{2t}), \\
&\vdots \\
S_K(\tau|y_{1,t}, y_{2,t}, \ldots, y_{K-1,t}, \underline{Y}_{t-1}) &= Q_\tau(y_{Kt}|x_{Kt}).
\end{aligned} \tag{26}
$$

Since Q_τ is non-decreasing in τ, the same is true for $S_1, S_2, \ldots, S_K$. Let $u_t = (u_{1t}, \ldots, u_{Kt})$. From $y_{kt} = Q_{u_{kt}}(y_{kt}|x_{kt})$ it follows that x_{kt} is measurable w.r.t. $u_{1t}, \ldots, u_{k-1,t}, \mathcal{F}_{t-1}$. Hence,

$$
\begin{aligned}
&Q_\tau\Big(y_{kt} - Q_\tau(y_{kt}|x_{kt})\Big| u_t \setminus u_{kt}, \mathcal{F}_{t-1}\Big) = Q_\tau\Big(y_{kt} - Q_\tau(y_{kt}|x_{kt})\Big| x_{kt}, u_t \setminus u_{kt}, \mathcal{F}_{t-1}\Big) = \\
&Q_\tau(y_{kt}|x_{kt}, u_t \setminus u_{kt}, \mathcal{F}_{t-1}) - Q_\tau(y_{kt}|x_{kt}) = Q_\tau(y_{kt}|x_{kt}, \mathcal{F}_{t-1}) - Q_\tau(y_{kt}|x_{kt}) = 0,
\end{aligned}
$$

which together with (26) means that claim (i) is verified. From $y_{kt} = Q_{u_{kt}}(y_{kt}|x_{kt})$ and from (26),

$$
y_{1t} = S_1(u_{1t}|y_{2,t}, y_{3,t}, \ldots, y_{K,t}, \underline{Y}_{t-1}) = Q_{u_{1t}}(y_{1t}|x_{1t}), \tag{27}
$$

$$
y_{2t} = S_2(u_{2t}|y_{1,t}, y_{3,t}, \ldots, y_{K,t}, \underline{Y}_{t-1}) = Q_{u_{2t}}(y_{2t}|x_{2t}), \tag{28}
$$

$$
\vdots
$$

$$
y_{Kt} = S_K(u_{Kt}|y_{1,t}, y_{2,t}, \ldots, y_{K-1,t}, \underline{Y}_{t-1}) = Q_{u_{Kt}}(y_{Kt}|x_{Kt}). \tag{29}
$$

Conditional on any draw of the random vectors $u_t, \underline{Y}_{t-1}$, there is unique y_{1t} solving equation (27).

Given $u_t, \underline{Y}_{t-1}, y_{1t}$, there is unique y_{2t} solving equation (28), etc. Thus, there are unique values $y_{1t}, y_{2t}, \dots, y_{Kt}$ satisfying (27) through (29), and claim (ii) is verified. $\square$

Proof of Corollary 1. From Lemma 1 (ii) we know that the solution $\{Y_t\}$ of (1) is a.s. unique, so we only need to show that (3) implies (1). Suppose (3) holds. By symmetry, it suffices to check the first equation in (1). Take any $\tau \in (0,1)$ and any event $E \subset \sigma(\{u_{1t}, \dots, u_{Kt}\} \setminus u_{1t}, \mathcal{F}_{t-1})$, $P(E) > 0$. We know that u_{1t} is independent of E and $S_1(\tau, \cdot)$ is non-decreasing in τ. Hence, $Q_\tau(S_1(u_{1t}|y_{2,t}, y_{3,t}, \dots, y_{K,t}, \underline{Y}_{t-1}) - S_1(\tau|y_{2,t}, y_{3,t}, \dots, y_{K,t}, \underline{Y}_{t-1}|E) = 0$, which implies the first equation in (1). $\square$

Proof of Proposition 2. If $h = 0, j \leq i$, set $S^0_{ji}(\tau|y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}) = y_{jt}$, while $\{u^0_{jit}\}_{t\in\mathbb{Z}}$ is any sequence of independent $U[0,1]$ random variables, independent of $\{Y_t\}_{t\in\mathbb{Z}}$. In (7) we set $\beta^0_{jj0}(\tau) = 1$, the remaining coefficients equal 0.

Let $u_t = (u_{1t}, \dots, u_{Kt})$, $x_{1t} = \underline{Y}_{t-1}$, for $k \in \{2, \dots, K\}$ let $x_{kt} = (y_{1t}, y_{2t}, \dots, y_{k-1,t}, \underline{Y}'_{t-1})'$. For the other cases $h = 0, j > i$ or $h > 0$, the functions $S_1, \dots, S_K$ in (27)–(29) satisfy Assumption 3, which was actually used in the Proof of Lemma 1. From there, $y_{kt} = Q_{u_{kt}}(y_{kt}|x_{kt})$. By $hK + j - i - 1$ consecutive iterations we obtain a function $G^h_{ji} : \mathbb{R}^{j+KP+hK} \mapsto \mathbb{R}$ such that

$$y_{j,t+h} = G^h_{ji}(u_{j,t+h}, \dots, u_{1,t+h}, u_{t+h-1}, \dots, u_{t+1}, u_{K,t}, \dots, u_{i+1,t}, y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}) \tag{30}$$

Let $S^h_{ji}(\tau|y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}) = Q_\tau(y_{j,t+h}|y_{1t}, \dots, y_{it}, \underline{Y}_{t-1})$, $u^h_{jit} = \inf\{\tau \in (0,1) : y_{j,t+h} \leq S^h_{ji}(\tau|y_{1t}, \dots, y_{it}, \underline{Y}_{t-1})\}$. Then $y_{j,t+h} = Q_{u^h_{jit}}(y_{j,t+h}|y_{1t}, \dots, y_{it}, \underline{Y}_{t-1})$. As in the Proof of Lemma 1, it can be shown that if S^h_{ji} is increasing in τ a.s., then $u^h_{jit} \sim U[0,1]$. (Even if S^h_{ji} is not increasing in τ a.s., a modified $U[0,1]$ version of u^h_{jit} can be constructed the same way as in the Proof of Lemma 1.) Likewise, u^h_{jit} is independent of $y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}$. Because of (30), u^h_{jit} is measurable w.r.t. $u_{j,t+h}, \dots, u_{1,t+h}, u_{t+h-1}, \dots, u_{t+1}, u_{K,t}, \dots, u_{i+1,t}$, and thus $\{u^h_{jit}\}$ is h-dependent.

From the Proof of Lemma 1, recall that u_{it} is independent of $y_{1t}, \dots, y_{i-1,t}, \underline{Y}_{t-1}$. Let $X_{it} = (y_{1t}, \dots, y_{it}, \underline{Y}'_{t-1})'$. Recall that $\{Y_t\}$ differs from the counterfactual version $\{\tilde{Y}_t\}$ due to altering $S_i(\cdot)$ to $S_i(\cdot)+s$. Thus, $(y_{1t}, \dots, y_{i-1,t}) = (\tilde{y}_{1t}, \dots, \tilde{y}_{i-1,t})$, $Y_{t-p} = \tilde{Y}_{t-p}$ for $p \geq 1$, and $\tilde{y}_{it} = y_{it}+s$.

Using Definition 1, $y_{it} = Q_{u_{it}}(y_{it}|x_{it})$, and $y_{j,t+h} = Q_{u^h_{jit}}(y_{j,t+h}|y_{1t}, \ldots, y_{it}, \underline{Y}_{t-1})$, we get

$$
\begin{aligned}
&\mathrm{IR}^{h,\tau,s}_{j,i}(Y_t, X_{it}) = Q_\tau(Q_\tau(\tilde{y}_{j,t+h}|u_{it}, X_{it})|Y_t, \underline{Y}_{t-1}) - Q_\tau(Q_\tau(y_{j,t+h}|u_{it}, X_{it})|Y_t, \underline{Y}_{t-1}) \\
&= Q_\tau(Q_\tau(\tilde{y}_{j,t+h}|X_{it})|Y_t, \underline{Y}_{t-1}) - Q_\tau(Q_\tau(y_{j,t+h}|X_{it})|Y_t, \underline{Y}_{t-1}) = Q_\tau(\tilde{y}_{j,t+h}|X_{it}) - Q_\tau(y_{j,t+h}|X_{it}) \\
&= Q_\tau(y_{j,t+h}|y_{1t}, \ldots, y_{i-1,t}, y_{it} + s, \underline{Y}_{t-1}) - Q_\tau(y_{j,t+h}|y_{1t}, \ldots, y_{i-1,t}, y_{it}, \underline{Y}_{t-1}). \qquad (31)
\end{aligned}
$$

If additionally (7) holds, (31) gives $\mathrm{IR}^{h,\tau,s}_{j,i}(y_{1t}, \ldots, y_{it}, \underline{Y}_{t-1}) = \beta^h_{ji0}(\tau)s$. $\square$

Proof of Proposition 3. Denote the argument of the criterion (9) as $b = (b'_{h_0}, b'_{h_0+1}, \ldots, b'_H)' \in \mathbb{R}^{\bar{K}}$, and represent the criterion (9) as

$$
\begin{aligned}
&\sum_{h=h_0}^{H} \frac{1}{T} \sum_{t=P+1}^{T-h} l_{t,h}(\tau, b_h) + p(b), \text{ where } l_{t,h}(\tau, b_h) = \rho_\tau(y_{j,t+h} - b'_h x_t), \\
&p(b) = \frac{1}{T} \sum_{k=0}^{i+KP} \lambda_k \sum_{h=D}^{H} |\Delta^D b_{k,h}| + \frac{1}{T}\mu\left(\lambda_0|\Delta^2 b_{0,H}| + \sum_{k=1}^{i+KP} \lambda_k |\Delta b_{k,H}|\right).
\end{aligned}
$$

Take a compact set $B_H \subset \mathbb{R}^K$ such that $B \subset \times_{h=h_0}^{H} B_H$. For all t, the function $l_{t,h}$ is continuous on the compact set $[\underline{\tau}, \bar{\tau}] \times B_H$. Also, for all $\tau \in [\underline{\tau}, \bar{\tau}], b_h \in B_H$ it holds $l_{t,h}(\tau, b_h) < |(y_{j,t+h} - b'_h x_t)| \leq |y_{j,t+h}| + |b'_h||x_t|$, so $\mathbb{E}|l_t(\tau, b)| \leq \mathbb{E}|y_{j,t}| + \sup_{b_h \in B_H} |b_h|'\mathbb{E}|x_t| < \infty$. By the Lemma 2.4 of Newey and McFadden (1994), it follows that for all $h \in \{h_0, \ldots, H\}$

$$
\sup_{b_h \in B_H, \tau \in [\underline{\tau}, \bar{\tau}]} \left| \frac{1}{T-h-P} \sum_{t=P+1}^{T-h} l_{t,h}(\tau, b_h) - \mathbb{E}[\rho_\tau(y_{j,t+h} - b'_h x_t)] \right| \xrightarrow{p} 0.
$$

The assumption $\bar{\lambda} = o(T)$ and the Dini's theorem imply $\sup_{b \in B} p(b) \to 0$. The continuous mapping theorem then yields

$$
\sup_{b \in B} \sup_{\tau \in [\underline{\tau}, \bar{\tau}]} \left| \sum_{h=h_0}^{H} \frac{1}{T} \sum_{t=P+1}^{T-h} l_{t,h}(\tau, b_h) + p(b) - \sum_{h=h_0}^{H} \mathbb{E}[\rho_\tau(y_{j,t+h} - b'_h x_t)] \right| \xrightarrow{p} 0. \qquad (32)
$$

Since $(Y'_t, Y'_{t-1}, \ldots, Y'_{T-P})'$ is absolutely continuous, so is x_t. Thus, x_t has a joint density w.r.t. the Lebesgue measure. Hence, for any $a \in \mathbb{R}^K \setminus 0$ the variable $a'x_t$ also has a density w.r.t. the

Lebesgue measure, so $P(a'x_t = 0) = 0$.

The assumption that $S_{ji}^{h_0}(\tau|y_{1t}, \dots, y_{it}, \underline{Y}_{t-1})$ is increasing in τ a.s. implies that its inverse function $P(y_{j,t+h_0} \leq y|y_{1t}, \dots, y_{it}, \underline{Y}_{t-1})$ is increasing in y a.s. Thus, $\forall h \geq h_0 : P(y_{j,t+h} \leq y|y_{1t}, \dots, y_{it}, \underline{Y}_{t-1}) = \mathbb{E}[P(y_{j,t+h} \leq y|y_{1,t+h-h_0}, \dots, y_{i,t+h-h_0}, \underline{Y}_{t+h-h_0-1})]|y_{1t}, \dots, y_{it}, \underline{Y}_{t-1})]$ is also increasing in y a.s. Consequently, every b_h that minimizes $\mathbb{E}[\rho_\tau(y_{j,t+h} - b_h' x_t)]$ satisfies $(\theta_h(\tau) - b_h)' x_t = 0$ a.s., and invoking the result from the previous paragraph, $b_h = \theta_h(\tau)$ a.s.

The result (32) together with $\theta_h(\tau)$ being the a.s. unique minimizer of $\mathbb{E}[\rho_\tau(y_{j,t+h} - b_h' x_t)]$ imply $\sup_{b \in B} \sup_{\tau \in [\underline{\tau}, \bar{\tau}]} |\hat{\theta}(\tau) - \theta(\tau)| \xrightarrow{p} 0$. □

Proof of Proposition 4. Let $v_h = \sqrt{T}(b_h - \theta_h)$ and denote $\theta_{k-1,h}, \hat{\theta}_{k-1,h}, v_{k-1,h}$ the kth entry of $\theta_h, \hat{\theta}_h, v_h$, respectively. These are functions of τ; restrict their domain to $[\underline{\tau}, \bar{\tau}]$ for some $0 < \underline{\tau} < \bar{\tau} < 1$. Further, denote $v = (v_{h_0}', \dots, v_H')'$,

$$
\begin{aligned}
Z_h(v_h) &= \sum_{t=P+1}^{T-h} \rho_\tau\left(y_{j,t+h} - b_h' x_t\right), \\
V(v) &= \sum_{k=0}^{i+KP} \lambda_k \sum_{h=D}^{H} |\Delta^D b_{kh}| + \mu \left(\lambda_0 |\Delta^2 b_{0,H}| + \sum_{k=1}^{i+KP} \lambda_k |\Delta b_{kH}| \right).
\end{aligned}
$$

Subtract $V(0)$ from the criterion in (9) to see that (9) is equivalent to minimizing

$$
\sum_{h=h_0}^{H} Z_h(v_h) + V(v) - V(0). \tag{33}
$$

Using $\bar{\lambda} = o(\sqrt{T})$ and following steps similar to Fu and Knight (2000) we get

$$V(v) - V(0) = \sum_{k=0}^{i+KP} \lambda_k \sum_{h=D}^{H} (|\Delta^D b_{kh}| - |\Delta^D \theta_{kh}|) + \mu \Bigg(\lambda_0 (|\Delta^2 b_{kH}| - |\Delta^2 \theta_{kH}|) +$$
$$\sum_{k=1}^{i+KP} \lambda_k (|\Delta b_{kH}| - |\Delta \theta_{kH}|) \Bigg) = \sum_{k=0}^{i+KP} \frac{\lambda_k}{\sqrt{T}} \sum_{h=D}^{H} \frac{\left|\Delta^D v_{kh}/\sqrt{T} + \Delta^D \theta_{kh}\right| - |\Delta^D \theta_{kh}|}{1/\sqrt{T}} +$$
$$\mu \Bigg(\frac{\lambda_0}{\sqrt{T}} \frac{\left|\frac{1}{\sqrt{T}} \Delta^2 v_{kH} + \Delta^2 \theta_{kH}\right| - |\Delta^2 \theta_{kH}|}{1/\sqrt{T}} + \sum_{k=1}^{i+KP} \frac{\lambda_k}{\sqrt{T}} \frac{\left|\frac{1}{\sqrt{T}} \Delta v_{kH} + \Delta \theta_{kH}\right| - |\Delta \theta_{kH}|}{1/\sqrt{T}} \Bigg) =$$
$$= \sum_{k=0}^{i+KP} \frac{\lambda_k}{\sqrt{T}} \sum_{h=D}^{H} O(1) - \mu \Bigg(\frac{\lambda_0}{\sqrt{T}} O(1) + \sum_{k=1}^{i+KP} \frac{\lambda_k}{\sqrt{T}} O(1) \Bigg) \to 0. \tag{34}$$

For any x_t from the domain of $f_{t,h}$, the assumption $f_{t,h} > 0$ implies $F'_{t,h} > 0$. Therefore, $F_{t,h}^{-1}(\tau) : (0,1) \mapsto \mathbb{R}$ exists and is continuous. However, $F_{t,h}^{-1}(\tau) = Q_\tau(y_{j,t+h}|x_t) = x_t' \theta_h(\tau)$, so $\theta_h : (0,1) \mapsto \mathbb{R}$ is continuous, and thus θ_h is bounded on $[\underline{\tau}, \bar{\tau}]$. This, together with the compact parameter set assumption, implies that v is also bounded on $[\underline{\tau}, \bar{\tau}]$. Since $V(v)$ is continuous, $V(v) - V(0) = o(1)$ holds uniformly in τ. From (34) and Koenker and Xiao (2006)[p. 121], (33) is equivalent to minimizing

$$\sum_{h=h_0}^{H} \sum_{t=1+P}^{T-h} \left(\rho_\tau \left(e_{t,\tau}^h - \frac{v_h' x_t}{\sqrt{T}} \right) - \rho_\tau(e_{t,\tau}^h) \right) + o(1) = \sum_{h=h_0}^{H} Z_h(v_h) + o(1), \tag{35}$$

whose minimum is o(1) and is attained at $v_h = \sqrt{T}(\hat{\theta}_h(\tau) - \theta_h(\tau))$ for all $h \in \{h_0, \ldots, H\}$. By the identity of Knight (1998), (35) equals

$$-\sum_{h=h_0}^{H} \frac{v_h'}{\sqrt{T}} \sum_{t=1+P}^{T-h} x_t \psi_\tau(e_{t,\tau}^h) + \sum_{h=h_0}^{H} \sum_{t=1+P}^{T-h} \int_0^{\frac{v_h' x_t}{\sqrt{T}}} (\mathbb{1}\{e_{t,\tau}^h \le s\} - \mathbb{1}\{e_{t,\tau}^h < 0\}) ds + o(1). \tag{36}$$

Denote $\Psi_{t,\tau} = (\psi_\tau(e_{t,\tau}^{h_0}), \ldots, \psi_\tau(e_{t,\tau}^H))'$, and also denote (35) as $Z(v)$. We express (36) as

$$Z(v) = -\frac{v'}{\sqrt{T}} \sum_{t=1}^{T} \Psi_{t,\tau} \otimes x_t + \sum_{h=h_0}^{H} \sum_{t=1}^{T} \int_0^{\frac{v_h' x_t}{\sqrt{T}}} (\mathbb{1}\{e_{t,\tau}^h \le s\} - \mathbb{1}\{e_{t,\tau}^h < 0\}) ds + o_p(1), \tag{37}$$

where the term $o_p(1)$ holds uniformly in τ.

Regarding the first term in (37), from Lemma 7 in the Online Appendix

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Psi_{t,\tau} \otimes x_t \xrightarrow{d} N(0, \Omega_0(\tau, \tau)). \tag{38}$$

Regarding the second term in (37), from Lemma 8 in the Online Appendix

$$\sum_{h=h_0}^{H} \sum_{t=1}^{T} \int_0^{\frac{v_h' x_t}{\sqrt{T}}} (\mathbb{1}\{e_{t,\tau}^h \leq s\} - \mathbb{1}\{e_{t,\tau}^h < 0\}) ds \xrightarrow{p} \frac{1}{2} v' \Omega_1(\tau) v. \tag{39}$$

Due to (37)-(39), we apply the arguments of Koenker and Xiao (2006) in the proof of their Theorem 3.1, and obtain $\sqrt{T}(\hat{\theta}(\tau) - \theta(\tau)) \xrightarrow{d} N(0, \Sigma(\tau, \tau))$.

Now additionally impose $\mathbb{E}[Y_t^4] < \infty$. From Lemma 7 in the Online Appendix, (38) holds with $\xrightarrow{d}$ replaced by $\Rightarrow$. From Lemma 8 in the Online Appendix, (39) holds with $\xrightarrow{p}$ replaced by $\Rightarrow$. Finally, by the arguments of Koenker and Xiao (2006) in the proof of their Theorem 3.1, we conclude $\sqrt{T}(\hat{\theta}(\tau) - \theta(\tau)) \Rightarrow G(\tau)$. $\square$

Proof of Proposition 5. In the special case $j = 1, h = 0$, set $S_1^0(\tau | y_{1t}, \underline{Y}_{t-1}) = y_{1t}$. $\{u_{1t}^0\}$ is any sequence of independent $U[0,1]$ random variables, independent of $\{Y_t\}$.

Now consider $j > 1$ or $h > 0$. Let $u_{jt}^h = \inf\{\tau \in (0,1) : y_{j,t+h} \leq Q_\tau(y_{j,t+h} | u_{1t}, \underline{Y}_{t-1})\}$. Then $y_{j,t+h} = Q_{u_{jt}^h}(y_{j,t+h} | u_{1t}, \underline{Y}_{t-1})$ and u_{jt}^h is independent of $u_{1t}, \mathcal{F}_{t-1}$. By Corollary 1 there is $G : \mathbb{R}^{(K+1)P} \mapsto \mathbb{R}^K$ such that $G(u_t, \underline{Y}_{t-1}) = Y_t$. Hence, there is $G_j^{(0)}$ such that $y_{j,t} = G_j^{(0)}(u_t | \underline{Y}_{t-1})$. For any $h \geq 0$, substituting yields G_j^h such that $y_{j,t+h} = G_j^h(u_{t+h}, \ldots, u_t | \underline{Y}_{t-1})$. Thus, $\sigma(u_{jt}^h) \subset \sigma(y_{j,t+h}, u_{1t}, \underline{Y}_{t-1}) \subset \sigma(u_{t+H}, \ldots, u_t, \underline{Y}_{t-1})$. Since u_{jt}^h is independent of $\underline{Y}_{t-1}$ it follows that u_{jt}^h is measurable w.r.t. $\{u_t, u_{t+1}, \ldots, u_{t+h}\}$, so u_{jt}^h is h-dependent.

Let $\check{u}_{1t} = \inf_{\tau \in [0,1]}\{y_{1t} \leq S_1(\tau | y_{2,t}, \ldots, y_{K,t}, \underline{Y}_{t-1})\}$. Then $S_1(u_{1t} | \cdot) = S_1(\check{u}_{1t} | \cdot)$, which together with $y_{j,t+h} = Q_{u_{jt}^h}(y_{j,t+h} | u_{1t}, \underline{Y}_{t-1})$ implies $y_{j,t+h} = Q_{u_{jt}^h}(y_{j,t+h} | \check{u}_{1t}, \underline{Y}_{t-1})$. And since $\check{u}_{1t}$ is a function of Y_t and $\underline{Y}_{t-1}$, there exists a function S_j^h such that $y_{j,t+h} = S_j^h(u_{jt}^h | Y_t, \underline{Y}_{t-1})$, so the second equation in (15) holds. Also,

$$Q_\tau\big(y_{j,t+h} - S_j^h(\tau|Y_t, \underline{Y}_{t-1})\big|u_{1t}, \underline{Y}_{t-1}\big) = Q_\tau\big(y_{j,t+h} - Q_\tau(y_{j,t+h}|u_{1t}, \underline{Y}_{t-1})\big|u_{1t}, \underline{Y}_{t-1}\big) = 0. \quad (40)$$

We have $\sigma(z_t) = (\sigma(z_t)\backslash\sigma(u_{1t}))\cup(\sigma(z_t)\cap\sigma(u_{1t}))$. By Assumption 5, $\sigma(z_t)\backslash\sigma(u_{1t})$ is independent of $\{Y_t\}$, and thus (40) implies the first equation in (15). If $Q_\tau(y_{j,t+h}|u_{1t}, \underline{Y}_{t-1})$ is increasing in τ a.s., then $u_{j,t}^h \sim U[0,1]$, as in the proof of Lemma 1. (Even if $Q_\tau(y_{j,t+h}|y_{1t}, \underline{Y}_{t-1})$ is not increasing in τ a.s., modified version of u_{jt}^h can be constructed, as in the proof of Proposition 1.)

Recall that $y_{j,t+h} = S_j^h(u_{jt}^h|Y_t, \underline{Y}_{t-1})$, where u_{jt}^h is independent of $\{u_{1t}, \underline{Y}_{t-1}\}$. Thus, invoking Assumption 2, $\tilde{y}_{j,t+h} = S_j^h(u_{jt}^h|y_{10}+s, y_{20}, \dots, y_{K0}, \underline{Y}_{-1})$. These results together with (4) imply

$$\begin{aligned}
&\mathrm{IR}_{j,1}^{h,\tau,s}(Y_0, \underline{Y}_{-1}) = Q_\tau(Q_\tau(\tilde{y}_{jh}|u_{10}, \underline{Y}_{-1})|Y_0, \underline{Y}_{-1}) - Q_\tau(Q_\tau(y_{jh}|u_{10}, \underline{Y}_{-1})|Y_0, \underline{Y}_{-1}) = \\
&Q_\tau(Q_\tau(S_j^h(u_{jt}^h|y_{10}+s, y_{20}, \dots, y_{K0}, \underline{Y}_{-1})|u_{10}, \underline{Y}_{-1})|Y_0, \underline{Y}_{-1}) - \\
&\quad Q_\tau(Q_\tau(S_j^h(u_{jt}^h|y_{10}, y_{20}, \dots, y_{K0}, \underline{Y}_{-1})|u_{10}, \underline{Y}_{-1})|Y_0, \underline{Y}_{-1}) = \\
&Q_\tau(S_j^h(\tau|y_{10}+s, y_{20}, \dots, y_{K0}, \underline{Y}_{-1})|Y_0, \underline{Y}_{-1}) - Q_\tau(S_j^h(\tau|y_{10}, y_{20}, \dots, y_{K0}, \underline{Y}_{-1})|Y_0, \underline{Y}_{-1}) = \\
&S_j^h(\tau|y_{10}+s, y_{20}, \dots, y_{K0}, \underline{Y}_{-1}) - S_j^h(\tau|Y_0, \underline{Y}_{-1}) \qquad (41)
\end{aligned}$$

If S_j^h is linear, that is, if (17) is satisfied, then (41) equals $\beta_{j10}^h(\tau)$. □

Proof of Proposition 6. Showing $\sqrt{T}(\hat{\gamma}_h(\alpha_h(\tau), \tau) - \gamma_h(\alpha_h(\tau), \tau)) \Rightarrow G(\tau)$ is analogous to the proof of Proposition 4. Take any $h \in \mathbb{N}_0, \tau \in (0,1)$. From (17) we know $\gamma_h(\alpha_h(\tau), \tau) = 0$, so $\hat{W}_{T,h}(\alpha_h(\tau), \tau) \xrightarrow{d} \chi^2_{\dim(\gamma)}$. Since $\mathscr{A}$ is compact and $(1+\mu)\lambda = o(1)$, it follows that $\max\{\{\alpha_h\}_{h=0}^H \in \mathscr{A}^{H+1} : \lambda\sum_{h=D}^H |\Delta^D \alpha_h| + \mu\lambda|\Delta\alpha_H|\} \to 0$. Thus, $\liminf_{T\to\infty} P(\alpha_h(\tau) \in (\underline{\alpha}_h(\tau,p), \bar{\alpha}_h(\tau,p)) \geq \liminf_{T\to\infty} P(\hat{W}_{T,h}(\alpha_h(\tau), \tau) \leq c_p) = P(\chi^2_{\dim(\gamma)} \leq c_p) = p$. □